**UNIVERSIDADE FEDERAL DE JUIZ DE FORA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL**

**Camila Martins Saporetti**

# Integração de dados petrofísicos, petrográficos e de técnicas de inteligência computacional para a caracterização litológica de reservatórios de petróleo

Juiz de Fora
2020

**Camila Martins Saporetti**

# Integração de dados petrofísicos, petrográficos e de técnicas de inteligência computacional para a caracterização litológica de reservatórios de petróleo

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutor em Modelagem Computacional. Área de concentração:

Orientador: Prof. D.Sc. Leonardo Goliatt da Fonseca

Coorientador: Prof. D.Sc. Egberto Pereira

Juiz de Fora

2020

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Saporetti, Camila Martins.
Integração de dados petrofísicos, petrográficos e de técnicas de inteligênciacomputacional para a caracterização litológica de reservatórios de petróleo / Camila Martins Saporetti. -- 2020.
203 p.

Orientador: Leonardo Goliatt da Fonseca
Coorientador: Egberto Pereira
Tese (doutorado) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2020.

1. Inteligência Computacional. 2. Litologia. 3. Caracterização de Reservatório. I. da Fonseca, Leonardo Goliatt, orient. II. Pereira, Egberto, coorient. III. Título.

**Camila Martins Saporetti**

**Integração de dados petrofísicos, petrográficos e de técnicas de inteligência computacional para a caracterização litológica de reservatórios de petróleo**

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutor em Modelagem Computacional. Área de concentração:

Aprovada em 12 de Agosto de 2020

BANCA EXAMINADORA

---

Prof. D.Sc. Leonardo Goliatt da Fonseca - Orientador
Universidade Federal de Juiz de Fora

---

Prof. D.Sc. Egberto Pereira - Coorientador
Universidade do Estado do Rio de Janeiro

---

Prof[a]. D.Sc. Luciana Conceição Dias Campos
Universidade Federal de Juiz de Fora

---

Prof. D.Sc. Heder Soares Bernardino
Universidade Federal de Juiz de Fora

---

D.Sc. Leonardo Costa de Oliveira
Petrobras

---

D.Sc. Eduardo Krempser da Silva
Fundação Oswaldo Cruz

Dedico este trabalho à minha mãezinha. Por todos estes anos de muito amor, carinho e cuidados conosco. Obrigada.

## AGRADECIMENTOS

Agradeço à Deus por me amparar nos momentos difíceis e por me dar força interior para superar as dificuldades.

Agradeço a todos aqueles que de alguma forma contribuíram para que esse trabalho se concretizasse.

Ao Leonardo Goliatt agradeço a orientação, a confiança, a dedicação e claro a paciência durante esses oito anos. Obrigada pelo incentivo, por sempre me mostrar as possibilidades e por ser este exemplo de profissional.

Ao Egberto Pereira pela coorientação e a disponibilidade em ajudar. Seu ponto de vista foi fundamental para o andamento desta tese e análise dos resultados.

Ao Leonardo Oliveira pela colaboração e por ter iniciado esse projeto e acreditado que ele poderia evoluir.

Ao Reginaldo, técnico do laboratório, por sempre consertar nosso PC's e propor melhorias. Muito obrigada pela disponibilidade.

As secretárias do PGMC: Maíra, Samantha, Renata, Adriana, Natália e demais que passaram pelo PGMC no período que estive por lá. Obrigada por sempre atenderem e resolverem as nossas solicitações da melhor maneira possível.

A Marilene, Renata e Rosana por manter nosso ambiente de trabalho em condições e pelas conversas nos corredores. Vou sentir saudades.

À ANP por fornecer dados que colaboraram com o desenvolvimento deste trabalho.

À UFJF agradeço pelos auxílios disponibilizados, que foram fundamentais para meu desenvolvimento acadêmico e pessoal.

À CAPES pela bolsa que possibilitou o andamento deste projeto.

À minha mãe e irmã, agradeço pela paciência, pelo apoio e por acreditarem que eu conseguiria. Vocês são fundamentais na minha vida.

À Francislaine, agradeço por ter feito da sua família a minha. Obrigado por sempre estar disponível quando preciso, por me escutar e por ter me dado apoio quando mais precisei.

À Ariane, Francislaine e a Jacqueliny agradeço pelos conselhos e pela convivência. Vocês se tornaram minhas irmãzinhas.

A Janaína e Rodrigo pela amizade, vocês foram fundamentais em diversos momentos. Muito obrigada!!!

A Daniele que se mostrou grande amiga durante o doutorado. A nossa aproximação foi muito importante para tornar os dias mais agradáveis.

Ao Marcus pela amizade e confiança depositada em mim por todo este período.

Você foi uma das boas surpresas que o doutorado me reservou.

Às doidonas dos créditos, nosso grupinho da graduação, Anna Claudia, Bárbara, Isis, Letícia, Liliane e Stephanie, agradeço pela amizade e por me escutarem mesmo longe. Vocês foram mais que fundamentais para que este trabalho se realizasse.

Aos amigos de pós-graduação pela ajuda e pelos momentos de descontração.Vocês são demais.

Por último, não menos importante, gostaria de agradecer ao ex-presidente Luiz Inácio Lula da Silva e a ex-presidente Dilma Rousseff pelas políticas públicas empregadas nas universidades federais as quais me permitiram chegar até aqui.

"Que os vossos esforços desafiem as impossibilidades, lembrai-vos de que as grandes coisas do homem foram conquistadas do que parecia impossível."(Charles Chaplin).

## RESUMO

A litologia é a descrição das características físicas e mineralógica de uma unidade rochosa ou formação rochosa. Sua definição em poços de petróleo por meio de múltiplos perfis elétricos e geofísicos tem um papel importante no processo de caracterização do reservatório. A partir da litologia, pode-se gerar modelos que serão a base através da qual cálculos petrofísicos são feitos, e em seguida, podem ser usados em simuladores de fluxo para compreender o comportamento de um campo de petróleo. A identificação pode ser realizada por métodos diretos e indiretos, mas nem sempre são viáveis devido ao custo ou imprecisão dos resultados. Modelos preditivos de distribuição de heterogeneidades e qualidade em reservatórios de hidrocarbonetos são fundamentais para exploração e otimização da produção de campos de óleo e gás. As heterogeneidades são determinadas por meio das distintas petrofácies, um conjunto de características petrográficas que especificam um grupo de rochas. O procedimento de identificar petrofácies geralmente é longo, o que faz com que a automatização seja necessária para agilizar o processo, e assim a análise seja concluída rapidamente. Através de sua determinação pode-se obter informações sobre as rochas reservatório, tais como: sua história deposicional e diagenética, estrutura do poro e mineralogia. Nesse contexto, técnicas de inteligência computacional aparecem como uma alternativa para discriminar litologia e petrofácies. Este trabalho objetiva o desenvolvimento de uma metodologia capaz de auxiliar na caracterização de reservatórios petrolíferos. A litologia e petrofácies foram derivadas do reconhecimento de padrões de características petrofísicas e petrográficas respectivamente. As características foram analisadas por meio de Análise de Componentes Principais. Métodos supervisionados foram empregados para classificar amostras e avaliar como novas amostras serão distribuídas. Para encontrar os classificadores ótimos, o método de evolução diferencial foi aplicado. Técnicas para aumentar a dimensionalidade foram utilizadas como uma forma de avaliar o comportamento dos métodos utilizados. Foi utilizado a Análise Filogenética como uma ferramenta para entender o processo de diagênese que ocorre durante o processo de litificação da rocha sedimentar e identificação dos eventos que ocorreram durante este processo. A metodologia apresentada surge como uma alternativa para auxiliar o geólogo/petrólogo na caracterização de um reservatório de petróleo.



Palavras-chave: Inteligência Computacional. Litologia. Caracterização de Reservatório.

## ABSTRACT

Lithology is the description of the physical and mineralogical characteristics of a rock unit or rock formation. Its definition in oil wells through multiple electrical and geophysical profiles has an important role in the reservoir characterization process. From the lithology, models can be generated based on which petrophysical calculations are made. Then they can be used in flow simulators to understand the behavior of an oil field. Direct and indirect methods can carry out the identification, but they are not always feasible due to the results' cost or imprecision. Predictive models for the distribution of heterogeneities and quality in hydrocarbon reservoirs are fundamental for exploring and optimizing the production of oil and gas fields. Heterogeneities are determined employing different petrofacies, a set of petrographic characteristics that specify a group of rocks. The procedure of identifying petrofacies is generally long, which makes automation necessary to speed up the process, and thus the analysis is completed quickly. It is possible to obtain information about the reservoir rocks through its determination, such as their depositional and diagenetic history, pore structure, and mineralogy. In this context, computational intelligence techniques appear as an alternative to discriminate lithology and petrofacies. This work aims to develop a methodology capable of assisting in the characterization of oil reservoirs. Lithology and petrofacies were derived from the recognition of patterns of petrophysical and petrographic characteristics, respectively. The characteristics were analyzed through Principal Component Analysis. Supervised methods were used to classify samples and evaluate how new samples will be distributed. To find the optimal classifiers, the differential evolution method was applied. Techniques to increase dimensionality were used as a way to evaluate the behavior of the methods used. Phylogenetic Analysis was used to understand the process of diagenesis that occurs during the lithification process of sedimentary rock and the identification of the events that occurred during this process. The presented methodology appears as an alternative to assist the geologist/petrologist in characterizing an oil reservoir.



Keywords: Computational Intelligence. Lithology. Reservoir Characterization.

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| AIE | Agência Internacional de Energia |
| ANN | Redes Neurais Artificiais |
| ANP | Agência Nacional do Petróleo, Gás Natural e Biocombustíveis |
| BAMPETRO | Banco de Dados Ambientais para a Indústria do Petróleo |
| BDEP | Banco de Dados de Exploração e Produção |
| BDIEP | Base de Dados Integrada de Exploração e Produção |
| BNDG | Banco Nacional de Dados Gravimétricos |
| bpd | barris por dia |
| CALI | Diâmetro da perfuração |
| CNPC | Corporação Nacional de Petróleo da China |
| CNJ | Clustered Neighbor-Joining |
| CNST | Espessura estratigráfica verdadeira |
| CPRM | Serviço Geológico do Brasil |
| DGF | Campo de Gás Daniudui |
| DRHO | Densidade Aparente Corrigida |
| EPE | Empresa de Pesquisa Energética |
| EUA | Estados Unidos da América |
| E,D&P | Exploração, Desenvolvimento e Produção |
| E&P | Exploração e Produção |
| FCNL | Registro de Neutron Compensado (detector longe) |
| FFDC | Registro de Formação Compensada (detector longe) |
| FINEP | Financiadora de Estudos e Projetos |
| FN | Redes Funcionais |
| FPP | Fuzzy Petroleum Prediction |
| GB | Gradient Boosting |
| GPR | Radar de Penetração no Solo |
| GR | Raio Gama |
| GRNN | Rede Neural de Regressão Generalizada |
| GTB | Gradient Tree Boosting |
| HGF | Campo de Gás Hangjinqi |
| K | módulo de Elasticidade |
| KF | K-Fold |
| KNN | K-vizinhos mais próximos |
| LAS | Log ASCII Standard |
| LSSVM | Máquina de Vetor de Suporte pelo Mínimo Quadrado |
| MLP | Multilayer Perceptron |
| NCNL | Registro de Neutron Compensado (detector próximo) |
| NB | Naive Bayes |
| NFDC | Registro de Formação Compensada (detector próximo) |
| NPHI | Porosidade de Neutron |

| | |
|---|---|
| Opep | Países Exportadores de Petróleo |
| PB | Pentabytes |
| P&D | Pesquisa e Desenvolvimento |
| PIB | Produto Interno Bruto |
| PSO | otimização de enxame de partículas |
| RF | Florestas Aleatórias |
| RHOB | Densidade Aparente |
| RNA | Redes Neurais Artificiais |
| RSM | Modelo de Superfície de Resposta |
| SC | Coeficiente de Silhueta |
| SIRP3D | Simulador de Reservatórios de Petróleo 3D |
| SVM | Máquinas de Vetores de Suporte |
| TB | Terabytes |
| TENS | Tensão do cabo |
| TSPD | Perfuração de Sonda Sísmica de Túnel |
| VP | Velocidade sísmica da onda primária |
| VS | Velocidade sísmica da onda secundária |
| VSP | Perfil Sísmico Vertical |

## SUMÁRIO

# 1 INTRODUÇÃO

Neste capítulo serão apresentados a contextualização do tema proposto, a motivação onde está descrito o que levou a escolha do tema, os objetivos do trabalho e a estrutura do texto.

## 1.1 CONTEXTUALIZAÇÃO

Nesta seção o tema será apresentado e o contexto em que ele se insere. O problema é delimitado e a necessidade de resolução é exposta.

### 1.1.1 **Exploração de Recursos Energéticos Naturais**

Os recursos naturais são elementos existentes na natureza que são fundamentais ao ser humano, seja no processo de desenvolvimento da civilização ou na sobrevivência e conforto da sociedade em geral. São explorados para servir de matéria ou energia às pessoas. A evolução tecnológica é a grande responsável por tornar parte desses recursos acessíveis para a sociedade em geral.

Os recursos naturais energéticos são os mais utilizados pela população e atividades industriais, sendo dividido em renováveis e não-renováveis, sendo que os não-renováveis representam mais de 80% dos consumo de energia mundial. O petróleo é a fonte de energia mais consumida, fundamental na produção de energia elétrica e de combustíveis para transportes e máquinas industriais. O aumento do seu preço está relacionado com a diminuição das reservas mundiais (4).

Os maiores produtores mundiais de petróleo localizam-se no Oriente Médio, (65% das reservas mundiais), e incluem a Arábia Saudita, os Emirados Árabes Unidos, o Kuwait, o Irã e o Iraque. A estes juntam-se outros grandes produtores, como o Níger e a Namíbia (5). A América Latina e o Caribe produzem 1/3 da produção mundial de bioetanol, cerca de 25% de biocombustíveis e 13% de petróleo. No entanto, entre 2000 e 2010 a exploração petrolífera nessa região não acompanhou a alta dos preços, afastando-se da tendência mundial. Contudo, a renda estimada do setor de hidrocarbonetos durante o período de expansão 2004-2009 (7,1% do PIB) duplicou a média apresentada entre 1990-2003 (3,6% do PIB) (6).

Tendo em vista o que foi exposto acima observa-se que é necessário o aumento do investimento em pesquisa e desenvolvimento (P&D) em petróleo e gás, a fim de diminuir gastos e danos à natureza através da inovação dos métodos para exploração e explotação.

#### 1.1.1.1 *Investimentos em Recursos Naturais Energéticos no Mundo*

O investimento da China em exploração de petróleo e gás natural cresceu 12,6% durante o período 2012-2016 ao se comparar com os cinco anos anteriores. Com o aumento constante em investimento, a China teve crescimento contínuo nas reservas de petróleo nesse período. Houve a descoberta de mais de dez campos petrolíferos na escala 700 milhões de barris (7). Em 2018, o investimento total na busca de petróleo e gás recebeu um aumento de 8,9% em termos anuais e o investimento em extração de petróleo e gás aumentou 24,7%.

Segundo o Instituto de Pesquisa de Economia e Tecnologia da Corporação Nacional de Petróleo da China (CNPC), a China importou 65% de petróleo bruto utilizado em 2016. O país planeja aumentar a produção nacional do recurso para 1340 milhões de barris até 2020, enquanto a de gás natural deve ultrapassar 2260,8 bilhões de barris. Nos primeiros onze meses de 2016, a produção de petróleo bruto da China somou 1214,1 milhões de barris e a de gás natural atingiu 761,69 bilhões de barris (8).

Nos Estados Unidos da América (EUA), a produção de petróleo bruto deve atingir o nível de 12,1 milhões de barris por dia (bpd) em 2023, segundo a Agência Internacional de Energia (AIE). Hoje em dia, a Rússia produz 11 milhões de bpd liderando a produção. O documento da AIE apresenta que os EUA irão atingir novos marcos na grande expansão de sua produção de petróleo e gás, devido aos avanços tecnológicos, melhora da eficiência dentre outros fatores que encorajam produtores de óleo de xisto (um tipo de petróleo não convencional) a aumentar suas atividades de exploração (9).

Os EUA estão se aproximando de atingir a meta de produzir petróleo bruto suficiente para suprir a demanda doméstica por produtos refinados, conseguindo diminuir a dependência de importações do Oriente Médio. A expectativa é que a produção do Brasil, Canadá e Noruega, todos países que não pertencem à Organização dos Países Exportadores de Petróleo (Opep), aumente. Os EUA juntamente com essas nações irão acrescentar barris suficientes para satisfazer a progressista demanda até o final da década. Dos países inclusos na Opep, a previsão é que apenas os países do Oriente Médio tenha algum aumento na produção, visto que outros participantes, como por exemplo a Venezuela, encaram problemas internos.

A produção da Arábia Saudita poderá atingir 12,3 milhões de bpd em 2023, ou seja, poderá concorrer a hegemonia mundial com os EUA. No entanto, os sauditas geralmente produzem abaixo de seu potencial para manter sua postura de fornecedores-chave, sendo capazes de expandir ou reduzir a produção de acordo com as necessidades do mercado.

Tendo em vista todos os derivados líquidos, segundo a AIE, a produção dos EUA deverá desenvolver-se ao ponto de aproximadamente 17 milhões de bpd nos próximos cinco anos, perante os cerca de 13 milhões de bpd atuais, superando a Arábia Saudita ou a Rússia.

#### 1.1.1.2 *Investimentos no Brasil*

A Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) estima que o investimento em pesquisa e desenvolvimento (P&D) em produção de petróleo e gás no Brasil deverá somar aproximadamente 26 bilhões de reais nos próximos 4 anos (10).

Do início de 1998 até o segundo trimestre de 2013 foram gerados 8 bilhões de reais para investimentos em P&D, devido a cláusula que consta nos contratos de concessão para exploração e produção (E&P) de petróleo e gás natural no Brasil. A cláusula de P&D determina que concessionários devem fazer despesas em pesquisa, desenvolvimento e inovação nas áreas de utilidade para o setor de petróleo. Este valor é referente a 1% da receita bruta dos campos em que é devido o pagamento da participação especial (10).

Segundo a cláusula de P&D, no mínimo 50% dos recursos gerados devem ser direcionados a universidades ou institutos de pesquisa credenciados pela ANP, para projetos adotados pela ANP. No Brasil, a Empresa de Pesquisa Energética (EPE) é encarregada por estudos referentes à avaliação da reserva e potencial de áreas para E&P de petróleo e gás natural, que originam na previsão de produção de petróleo e gás natural, estimativas de investimentos e de equipamentos.

Desde o início da operação a produção de petróleo e gás nos reservatórios do pré-sal tem tido sucessivos aumentos, passando de uma média 55,8 mil bpd para 1.017,9 mil em 2016. Isso foi resultado da alta produtividade dos poços, assinalando a grande viabilidade econômica da área e o foco em produzir nesses reservatórios, proporcionando forte concentração de investimentos nessa área (11). A produção do Pré-sal em Dezembro de 2019, oriunda de 114 poços, foi de 2,117 milhões bpd de petróleo e 537 milhões bpd de gás natural, totalizando 2,654 milhões de barris de óleo equivalente por dia. Houve um aumento de 2,6% em relação ao mês anterior e um aumento de 40,6% se comparada ao mesmo mês de 2018 (12).

### 1.1.2 **Bases de dados na indústria do Petróleo e Gás**

Bases de dados ligadas à indústria do petróleo e gás apresentam grandes volumes de informações (mapas, dados sísmicos, perfis de poços, sensoriamentos e topografias), gerando conhecimentos geofísicos, geológicos e geoquímicos empregados na exploração. Os dois principais bancos de dados no setor de pesquisa e desenvolvimento de petróleo são o da Petrobras e o da ANP, que possui o monopólio dos dados gerados nas bacias sedimentares brasileiras (13).

#### 1.1.2.1 *Banco de Dados de Exploração e Produção - BDEP*

O Banco de Dados de Exploração e Produção (BDEP) possui cerca de 3,2 petabytes (PB) em informações. O mesmo foi criado em 2000, pela ANP em convênio com o Serviço

Geológico do Brasil (CPRM). A base de dados do BDEP armazena dados sísmicos (2D, 3D, dados de navegação, relatório do observador), dados de poço (perfis digitais, perfis compostos, pasta de poço, curvas em LAS) e dados de métodos não-sísmicos (gravimetria, magnetometria, entre outros). Os dados estão disponíveis conforme nome do levantamento sísmico, nome do poço e nome do projeto de métodos potenciais. Pelo sistema BDEP Web Maps, os dados de livre acesso no BDEP podem ser filtrados por localização geográfica, através de um procedimento manual, sendo possível identificar os nomes dos levantamentos conforme a área selecionada (14).

#### 1.1.2.2 *Banco Nacional de Dados Gravimétricos - BNDG*

Dados gravimétricos são empregados para determinar o geoide e tornam possível o estudo aprofundado da estrutura das bacias sedimentares. O geoide é uma determinada superfície equipotencial do campo da gravidade: aquela que mais se aproxima do nível médio dos mares; nos continentes e ilhas acha-se no interior da crosta (15). A comunidade científica brasileira executa levantamentos gravimétricos há mais de cem anos, porém os dados existentes eram dispersos em vários grupos de pesquisa. No intuito de reunir, organizar e padronizar os dados de gravimetria no país, a ANP implantou o projeto do Banco Nacional de Dados Gravimétricos (BNDG).

#### 1.1.2.3 *Base de Dados Integrada de E & P - BDIEP/PETROBRAS*

A Base de Dados Integrada de Exploração e Produção (BDIEP), da Petrobras contempla uma ampla gama de assuntos de E & P: informações geofísicas, geológicas, de reservatório, de produção de hidrocarbonetos, de interpretação exploratória, de engenharia de poço, de concessões, de divisões geográficas da terra, entre outros. A BDIEP tem um volume de dados com aproximadamente de 5 terabytes (TB), não levando em consideração os dados sísmicos, as informações de sensores remotos de produção e as informações gerais de E & P (documentos e arquivos). O volume de dados sísmicos soma mais de 120 TB de dados processados e mais de 7 PB de dados brutos, adquiridos pela Petrobras no Brasil e no exterior.

#### 1.1.2.4 *Banco de Dados Ambientais para a Indústria do Petróleo - BAMPETRO*

O Banco de Dados Ambientais para a Indústria do Petróleo (BAMPETRO) armazena informações ambientais multidisciplinares e georreferenciadas das bacias sedimentares terrestres e marinhas brasileiras, agrupadas nas seguintes áreas: geofísica, geologia, oceanografia física, meteorologia, química, biologia e socioeconomia. O BAMPETRO é fomentado pela Financiadora de Estudos e Projetos (FINEP) com recursos do fundo CT-Petro, e foi desenvolvido como um projeto em rede de P & D, com a participação de profissionais de várias instituições de ensino do Rio de Janeiro e de consultores de outras universidades

brasileiras. Contou em sua implantação com uma equipe multidisciplinar que incluiu cartógrafos, biólogos, químicos, oceanógrafos, geólogos, geógrafos, geofísicos, e tecnologistas da informação (16).

Metodologias vem sendo desenvolvidas para analisar os dados disponibilizados por estas bases citadas anteriormente. Nesta tese foram utilizadas bases de dados petrofísicas cedidas pelo Banco de Dados de Exploração e Produção (BDEP) para analisar a metodologia proposta.

### 1.1.3 **Ferramentas Computacionais para o Gerenciamento de Dados**

Analisar e armazenar dados pode ser novidade para alguns setores da indústria, entretanto, as empresas de petróleo e gás já lidam há algumas décadas com uma enorme quantidade de dados. Programas sísmicos, visualização de dados e dispositivos de computação com sensores que coletam e transmitem dados progridem na geração de novas possibilidades dentro do setor (13).

#### 1.1.3.1 *Visualização de Dados*

A visualização de dados de maneira que melhore a comunicação entre profissionais e partes interessadas envolvidas desde o desenvolvimento de campo até a tomada de decisões é importante para os diferentes estágios de Exploração, Desenvolvimento e Produção (E, D & P). O grupo *Interactive Modeling, Visualization & Analytics R&D* da Universidade de Calgary desenvolve protótipos de software que fornecem representações visuais e análises que reflitam e expressem as informações disponíveis, o nível de incerteza e requisitos de visualização para processos em diferentes estágios de E, D & P.

O software DigitalROCK (17) produz imagens em 3D da escala dos poros das rochas por meio de um scanner micro-CT de alta resolução. A partir dessas imagens consegue-se avaliar a permeabilidade relativa da saída de um reservatório e de formações rochosas. Essas imagens podem gerar uma simulação de fluxo de fluido e prever digitalmente como a mistura fluida irá fluir através da rocha, e como os fluidos do reservatório irão interagir com a formação rochosa.

O PetroVisual (18) é um software de análise e visualização de dados que trabalha com bancos de dados de petróleo e gás, como ARIES, PHDWin e OGRE. O PetroVisual permite que questionamentos sobre os dados de óleo e gás sejam feitos e que se analise os resultados da consulta através de uma vasta coleção de mapas, gráficos, grades, tabelas dinâmicas e árvores de pesquisa.

#### 1.1.3.2 *Extração e Geração de Conhecimento*

Empresas de desenvolvimento de softwares terceirizam ferramentas que possibilitam realizar a extração e geração de conhecimento através de imagens, dados petrográfi-

cos/petrofísicos, lâminas, dentre outras entradas. Estes softwares já existentes possuem bom desempenho e uma gama de possibilidades de análises, porém o custo é elevado para obter uma licença. Abaixo estão listados alguns desses desenvolvedores.

A Endeeper é uma provedora de software e serviços para gestão de conhecimento e integração de dados geológicos. Possui uma variedade de softwares disponíveis à disposição com intuito de aprimorar o procedimento de análise geológica a partir de características petrológicas e sedimentológicas de diferentes tipos de rocha. Como exemplo de software pode-se destacar o PETROLEDGE (análise petrográfica sistemática de reservatórios de petróleo clásticos, carbonáticos e outras rochas sedimentares) e o STRATALEDGE (sistema para descrição de testemunhos de rocha).

A PETREC é uma empresa de base tecnológica atuante no seguimento de tecnologia e inovação em serviços de aquisição, processamento e interpretação conjunta de dados geofísicos e geológicos. Através de parcerias, a PETREC consegue suprir as diferentes etapas do processo de exploração, desenvolvimento e produção de um campo, que envolve desde o condicionamento e processamento de dados geofísicos até a certificação de reservas. Determinação de características petrofísicas através de microtomografia, modelagem sísmica, e processamento sísmico terrestre e marítimo são alguns exemplos de serviços prestados.

O Simulador de Reservatórios de Petróleo 3D (SIRP3D) é um simulador que pode ser utilizado por profissionais de diversas áreas devido a sua interface gráfica ¨amigável¨. O usuário pode ativar e desativar poços, fornecer vazões de petróleo e água, estabelecer características do solo, como porosidade e permeabilidade. Ademais, possui um sistema de visualização que mostra em três dimensões o movimento dos fluidos no reservatório.

O WellRes é um aplicativo capaz de simular o escoamento tridimensional multifásico (água, óleo e gás) em reservatórios de petróleo e, simultaneamente, o escoamento trifásico no interior dos poços. Permite configurar problemas de forma fácil, assim como visualizar campos 3D e diversas curvas de produção/injeção do reservatório e das propriedades do escoamento no interior dos poços. Possibilita a importação de malhas e propriedades de rocha de softwares como IMEX e ECLIPSE.

As ferramentas citadas acima auxiliam em uma série de pesquisas, porém não fornecem um módulo para analisar a diagênese através das amostras e dos constituintes. A metodologia proposta, além de realizar análises que são realizadas como classificação e agrupamento e ser *open source*, atende essa necessidade de explorar os dados para auxiliar no processo de entendimento da diagênese.

## 1.2 MOTIVAÇÃO

O desenvolvimento e otimização da produção de campos de petróleo está sujeita à definição e a distribuição de heterogeneidades de reservatórios de hidrocarbonetos, o que faz com que estas tarefas sejam primordiais. Para uma rocha ser classificada como um reservatório apropriado para exploração e extração, ela deve apresentar além de uma extensão aceitável, boa porosidade, uma apreciável permeabilidade e um eficiente fator de recuperação, entre outros fatores. Tais propriedades, denominadas petrofísicas, estão ligadas ao percurso deposicional das unidades deposicionais, em especial às condições de sedimentação e ao processo diagenético, sendo as mesmas de extrema importância para definição da qualidade do reservatório. As heterogeneidades são caracterizadas por meio de diversas petrofácies sedimentares, um conjunto de características petrográficas que individualizam um grupo de rochas.

O estudo da diagênese das rochas, que pode ser definido como conjunto de processos químicos e físicos sofridos pelos sedimentos desde a sua deposição até a sua consolidação, vem sendo incentivado pelas empresas petrolíferas, com intuito de entender a distribuição da porosidade em reservatórios. O interesse vem do fato que estas rochas podem originar bons reservatórios de hidrocarbonetos. No decorrer do processo de diagênese, minerais podem precipitar-se como cimento nos poros da rocha, o que resulta na diminuição de sua porosidade e da permeabilidade, prejudicando seu potencial de armazenamento (19).

A partir da análise petrográfica é possível identificar os constituintes de uma rocha. Dessa forma, pode-se realizar uma avaliação das implicações futuras de suas propriedades sobre o comportamento do reservatório de petróleo (20). Essa análise ocorre com o uso de microscópio petrográfico, onde o geólogo/petrólogo descreve as lâminas discriminando seus aspectos geológicos. Uma base de dados é criada através das observações realizadas e pode-se, com isso, agrupar os dados em diferentes petrofácies.

O mapeamento da distribuição de heterogeneidades pode ser realizada através da quantificação dos constituintes minerais e diagenéticos da rocha em questão. Este processo usualmente é muito longo, pois envolve o processo de amostragem, geração dos dados e posterior interpretação destes (21). Além disso, devido à grande quantidade de dados, nem toda informação obtida pode ser adequadamente aproveitada no procedimento manual.

A definição de litologia em poços de petróleo por meio de múltiplos perfis petrofísicos tem um papel importante no processo de caracterização do reservatório. A partir da litologia, pode-se gerar modelos que, por sua vez, serão a base a partir da qual alguns cálculos são feitos e, em seguida, podem ser utilizados em simuladores de fluxo para compreender e estudar o comportamento de um dado campo. Através de sua determinação pode-se obter informações sobre a heterogeneidade do reservatório, um fator importante para avaliar o seu potencial.

A identificação de litologia pode ser feita por métodos diretos e indiretos (22). Métodos diretos são realizados através da obtenção de amostras petrofísicas do reservatório. Esta é a maneira mais segura de determinar a litologia, no entanto, obter essa amostra física nem sempre é uma tarefa fácil. Métodos indiretos fazem uso de registros de poços que medem as propriedades físicas de formações geológicas e fluidos fornecendo a maioria dos dados subterrâneos disponíveis para um geólogo de exploração. Além de sua importância nas decisões finais, eles também são ferramentas inestimáveis para mapear e identificar litologias. Embora, não geram o mesmo desempenho que os métodos diretos (23).

Tendo em vista que não há uma metodologia que unifique e atenda a realização das tarefas descritas acima, e que identificar os eventos diagenéticos é algo ainda feito de forma manual, nota-se a necessidade de automatizar os procedimentos de caracterização do reservatório com o objetivo de melhorar a produtividade ou a interpretabilidade dos dados. Assim, como as características petrográficas possuem relação com a história diagenética das rochas (20), será realizado um estudo do emprego de análise filogenética (24) para auxiliar no entendimento da diagênese e na caracterização e individualização das petrofácies. Nesse contexto, técnicas de Inteligência Computacional aparecem como um mecanismo útil para auxiliar na caracterização de reservatórios petrolíferos.

## 1.3 OBJETIVOS

Nesta seção serão apresentados o Objetivo Geral e os Objetivos Específicos que foram trabalhados durante o desenvolvimento do trabalho proposto.

### 1.3.1 **Objetivo Geral**

O objetivo principal deste trabalho é desenvolver uma metodologia computacional, por meio do uso de técnicas de inteligência computacional, capaz de extrair informações de dados petrográficos e petrofísicos de forma que auxilie o geólogo/petrólogo na caracterização de reservatórios de petróleo.

### 1.3.2 **Objetivos Específicos**

Os objetivos específicos são enumerados a seguir:

1. Estudar e implementar técnicas supervisionadas e não supervisionadas otimizando os parâmetros através do método de Evolução Diferencial e coeficiente de silhueta, respectivamente;

2. Aplicar o Bootstrap como uma alternativa para encontrar os parâmetros do método de agrupamento;

3. Analisar a geração de características polinomiais no procedimento não supervisionado;

4. Utilizar a Análise de Componentes Principais para obter informação a respeito da influência das características petrográficas/petrofísicas no processo de caracterização de reservatórios petrolíferos;

5. Introduzir um estudo do uso de Análise Filogenética como um meio de auxiliar no entendimento do processo diagenético.

## 1.4 ESTRUTURA DO TEXTO

Esse trabalho subdivide-se em sete capítulos. No Capítulo 1 encontra-se uma apresentação dessa tese que é composta pela contextualização, motivação e os objetivos.

No Capítulo 2 é exibido o contexto geológico, onde são descritos os principais conceitos geológicos importantes para o desenvolvimento desse trabalho. Informações sobre sedimentologia, diagênese, petrofácies, litologia, permeabilidade, porosidade serão encontradas nesse capítulo.

No Capítulo 3 é exibida uma revisão bibliográfica dos trabalhos que foram publicados nos últimos anos.

Os dados utilizados e os métodos de Inteligência Computacional aplicados são apresentados no Capítulo 4. Nesse capítulo estão contidos conceitos relacionados a classificação, análise de agrupamento, validação cruzada, evolução diferencial e métricas para a seleção de modelos.

No Capítulo 5 estão descritos os procedimentos realizados para o desenvolvimento do projeto em questão, os resultados e discussões.

O Capítulo 6 apresenta a conclusão com base nos resultados obtidos.

A perspectiva de trabalhos futuro encontra-se no Capítulo 6.1 e no Apêndice A encontram-se as bases de dados utilizadas para testar o desempenho das técnicas de Inteligência Computacional.

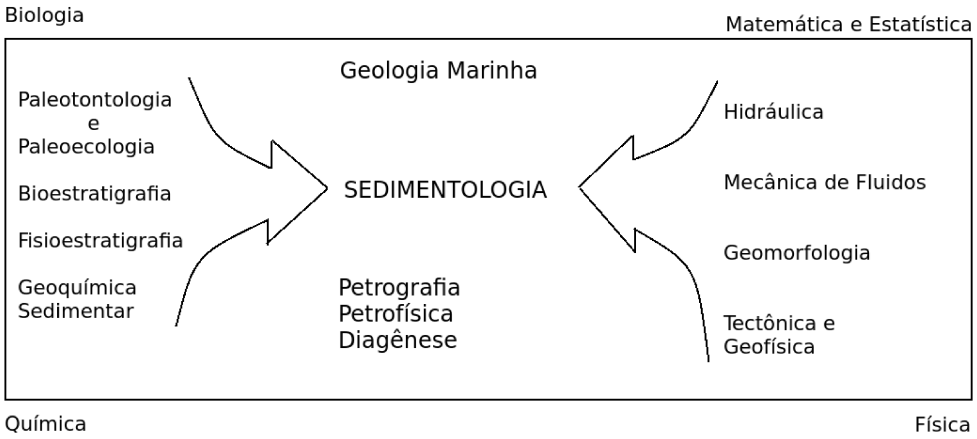
# 2 CONTEXTO GEOLÓGICO

Neste capítulo será exposto conceitos relacionados a geologia que são considerados importantes para o entendimento do trabalho.

## 2.1 SEDIMENTOLOGIA

A sedimentologia é o estudo dos depósitos sedimentares e suas origens. Pode ser aplicada em diversos tipos de depósitos (e.g. antigos ou modernos, marinhos ou continentais), sendo útil para análises minerais, de texturas e estruturas, diagenéticas e na evolução temporal e espacial de pacotes sedimentares (25). Com base em observação e descrição das feições em sedimentos inconsolados e consolidados se ocupa da reconstrução dos paleoambientes de sedimentação em termos estratigráficos e tectônicos. Faz-se uso de métodos de vários ramos das geociências e das ciências afins. A Figura 1 ilustra a relação entre a sedimentologia e algumas disciplinas.

Figura 1 - Sedimentologia e disciplinas relacionadas

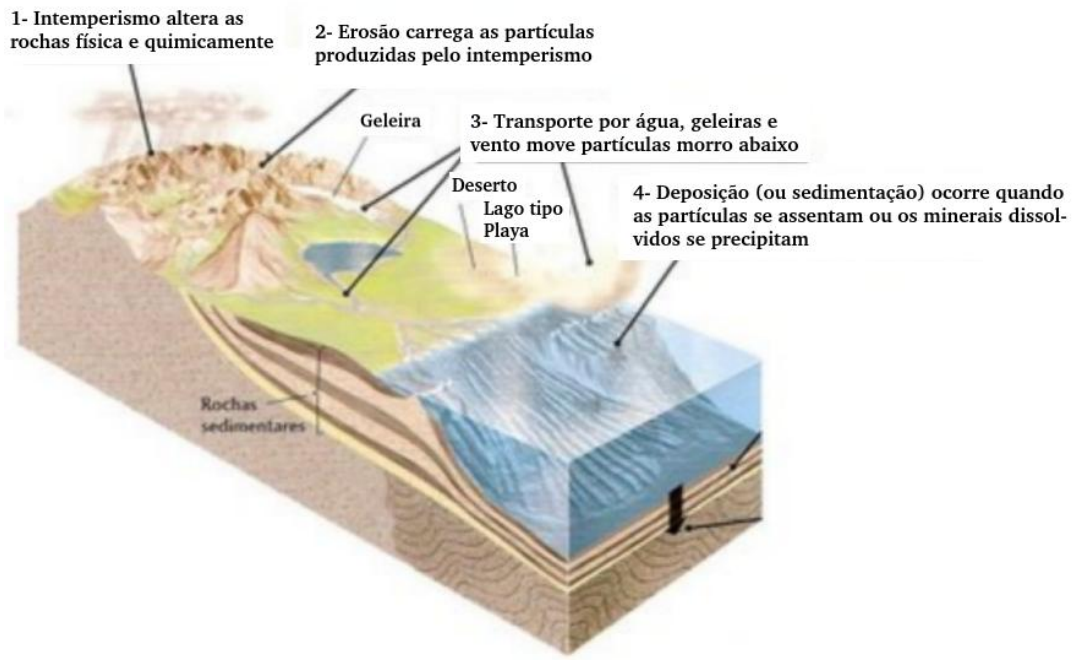



Fonte: Modificado de (25).

Os ramos de interesse para este trabalho são a Petrografia Sedimentar e a Petrologia Sedimentar que visam o estudo microscópio dos sedimentos. As propriedades petrofísicas (permeabilidade e porosidade) e a diagênese, que são processos químicos e físicos sofridos pelos sedimentos desde a sua deposição até a sua consolidação. Entender a diagênese é fundamental pois, ela interfere na variação permo-porosa de uma rocha sedimentar, e, consequentemente, na sua capacidade de armazenamento e no fluxo de fluidos, como água, petróleo e gás (25).

O processo de formação de uma rocha sedimentar ocorre a partir da atuação do intemperismo e da erosão, que promovem a desintegração de diferentes litotipos e posterior transporte, sedimentação e litificação dos detritos para uma bacia de deposição (i.e bacia sedimentar). (Figura 2). A seguir são descritas algumas dessas ações.

Figura 2 - Intemperismo, Erosão, Transporte e Deposição




Fonte: Retirado de (26).

### 2.1.1 **Intemperismo**

O intemperismo é o processo no qual as rochas são destruídas na superfície terrestre.A partir desse processo são gerados diferentes partículas granulométricas (eg. argila, silte, areia, seixo), os solos e outras substâncias, que são transportadas principalmente pelos rios para os oceanos. O intemperismo pode ser causado por processos físicos, químicos e biológicos.

#### 2.1.1.1 *Intemperismo Físico*

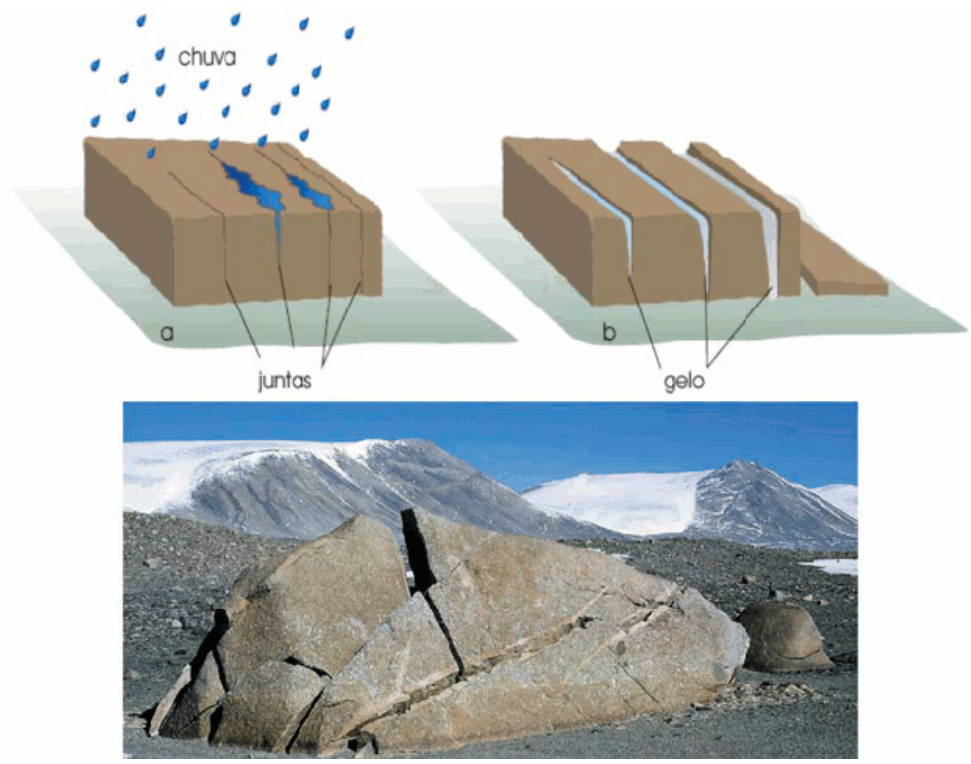
O intemperismo físico é constituído pelos processos que originam a desagregação das rochas, antes unidos e com sua fragmentação, transformando-as em material descontínuo e friável (27). As extensas alternâncias de temperaturas que ocorrem diariamente nas regiões frias e temperadas, seguidas de congelamento e descongelamento, levando à fragmentação dos grãos minerais. Ademais, os minerais com distintos coeficientes de dilatação térmica, procedem-se de maneira diferenciada às mudanças climáticas, o que causa deslocamento relativo entre os cristais, deteriorando a ligação inicial entre os grãos.

No caso da expansão térmica acontecer devido à insolação, ocorre em regiões com grandes variações térmicas entre o dia e a noite. Esta característica é comum em regiões desérticas. Assim as rochas se expandem e se contraem, estabelecendo um gradiente de temperatura entre a superfície e o interior da rochas quando a rocha é submetida ao aquecimento. Isso se dá pelo fato de que a maior parte das rochas possuem condutibilidade

térmica baixa. Consequentemente a superfície da rocha expande mais que seu interior, criando um esforço que ocasionaria uma desagregação (28).

O congelamento da água nas fissuras das rochas, conduzido por um aumento de volume, exerce pressão nas paredes, acarretando esforços que resultam por amplificar as fraturas e fragmentar a rocha (27). A Figura 3 exemplifica o processo descrito.

Figura 3 - Fragmentação pela ação do gelo




Fonte: Retirado de (27).

### 2.1.1.2 *Intemperismo Químico*

O intemperismo químico acontece quando os minerais de uma rocha são quimicamente modificados ou dissolvidos. A deterioração ou esmaecimento de inscrições gravadas em lápides ou monumentos antigos é causado sobretudo pelo intemperismo químico. Na maior parte dos ambientes da superfície terrestre as principais reações do intemperismo são: dissolução, hidratação, hidrólise e oxidação (29).

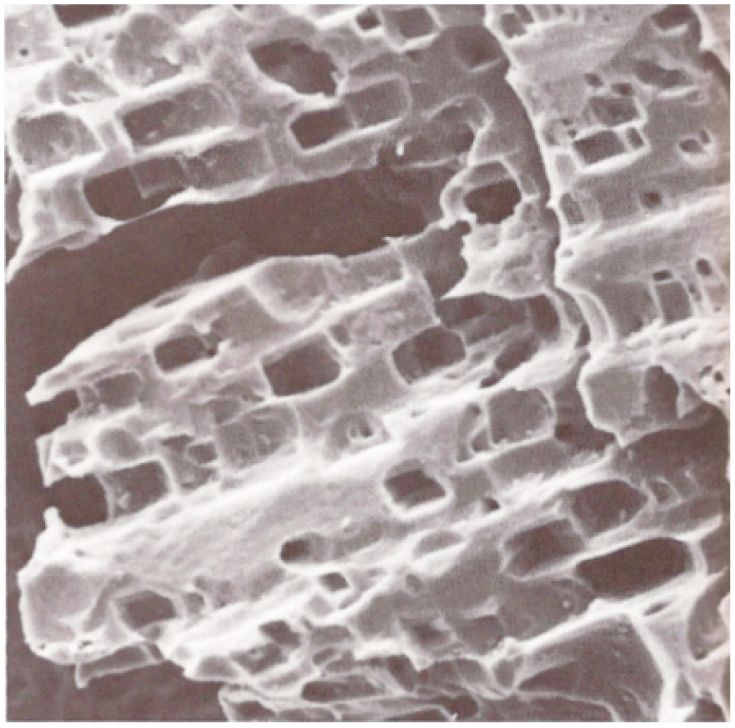
A dissolução usualmente representa o primeiro estágio do processo de intemperismo químico. A quantidade de material dissolvido provém da quantidade e da qualidade da água abrangida e da solubilidade do mineral (25).

A hidratação constitui a adição de água num mineral sem que aconteça nenhuma reação química. Na hidratação, os minerais expandem-se o que pode exercer pressões com efeitos semelhantes àqueles apurados no decorrer do congelamento da água. A hidrólise baseia-se na reação química entre o mineral e a água. A decomposição dos silicatos,

feldspatos, micas, hornblenda, augita dentre outros, realiza-se através da hidrólise, isto é, da ação da água dissociada.

A oxidação é uma das reações dominantes que acontecem durante o intemperismo químico. Quando a água com oxigênio dissolvido entra no subsolo, a oxidação processa-se inicialmente nos primeiros metros superficiais, parando totalmente no lençol freático. No processo de oxidação, o oxigênio reage com os minerais, especialmente com aqueles que contêm ferro, manganês e enxofre. A oxidação é beneficiada pela existência de umidade. A Figura 4 mostra feldspato após o intemperismo químico no solo.

Figura 4 - Fotomicrografia, obtida por microscópio eletrônico de varredura, de um feldspato marcado e corroído pelo intemperismo químico no solo.



Fonte: Retirado de (30).

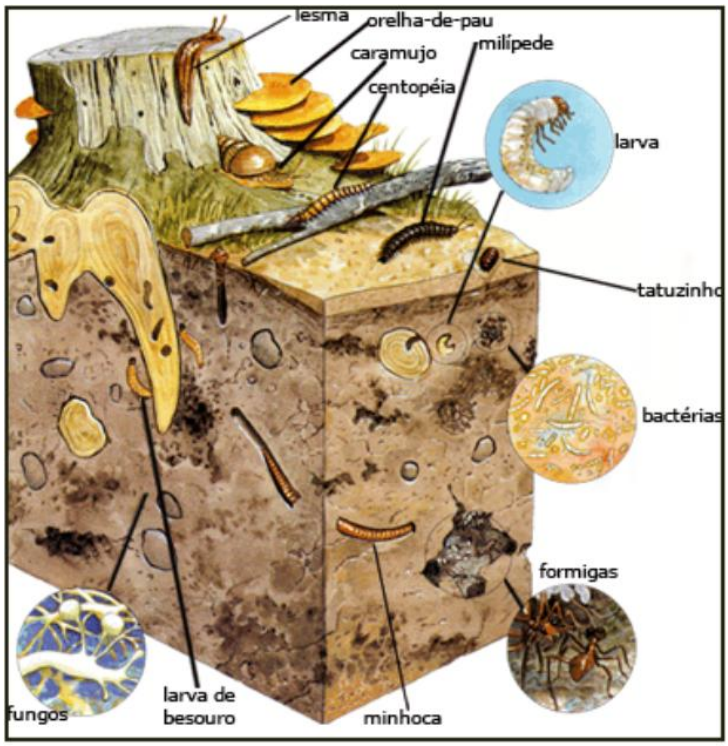
#### 2.1.1.3 *Intemperismo Biológico*

O intemperismo biológico é o processo de transformação das rochas através da ação de seres vivos, como bactérias, animais ou vegetais. Incluem-se nesse processo as raízes das árvores, as ações de bactérias, a decomposição de organismos ou excrementos, entre outros.

A ação dos organismos vivos é fundamental na formação do solo, tanto nos aspectos da criação dos horizontes superficiais orgânicos do solo, na qual possui um conglomerado de restos animais e/ou vegetais que são decompostos por microorganismos, como as bactérias. Esta camada de solo é encarregada pela preservação de vários ecossistemas que

precisam destas quantidades orgânicas para se desenvolverem. A Figura 5 mostra a ação dos organismos vivos no solo.

Figura 5 - A ação dos organismos vivos no solo que geram o intemperismo biológico.




Fonte: Retirado de (27).

### 2.1.2 **Erosão**

A erosão é um processo de deslocamento de terra ou de rochas de uma superfície. A erosão pode ocorrer devido a ação de fenômenos da natureza ou do ser humano. Os tipos de erosão que podem ser listados de acordo com o tipo de agente erosivo atuante, como a água, os ventos e os seres vivos (31).

A erosão pluvial é provocada pelas águas das chuvas. Em geral, todo desgaste do solo decorrente pelas precipitações pode ser denominado erosão pluvial. Em áreas pouco protegidas pela vegetação e outros elementos, os efeitos da ação da água podem ser sentidos com maior intensidade.

Erosão fluvial é o trabalho de remodelamento do relevo exercido pelos rios nas vertentes e interflúvios (25). O clima modifica a descarga fluvial e, consequentemente, os regimes dos rios e a forma de atuação do tipo de erosão em questão.

A erosão marinha geralmente é provocada pelas ondas nas regiões litorâneas. As falésias marinhas ativas são os indícios mais perceptíveis da ação desse fenômeno. O efeito

da dissolução pode se apresentar em rochas mais solúveis como calcário sendo também uma ação considerada erosão fluvial.

A erosão eólica é originada pela ação dos ventos, que vão aos poucos removendo as partículas dos solos. A erosão glacial é a causada pela ação do gelo. Geralmente ocorre devido as variações de temperatura que congelam e descongelam a água, que se dilata e se comprime, afetando as rochas e os solos.

### 2.1.3 **Transporte**

À medida que uma rocha sofre com as ações do intemperismo, os resíduos minerais são liberados do arcabouço rochoso e passam a constituir o manto de intemperismo. Os elementos soltos ficam susceptíveis à energia potencial em razão da aceleração gravitacional, sendo mais cedo ou mais tarde transportadas declive abaixo (25).

Podem ser identificados diversos tipos de transporte de acordo com os agentes envolvidos, que basicamente são os mesmos que agem na erosão. Dessa forma, podem ser identificados entre os substanciais, os transportes por águas pluviais e fluviais, ventos, geleiras e movimentos de massa (25).

Águas pluviais e fluviais são os principais agentes que atuam nas áreas continentais. Os elementos sedimentares incorporados a esses meios, a partir das atividades mecânicas e hidráulicas, podem ser transportadas por distintos processos.

Ventos acarretam o deslocamento de material sedimentar, de barlavento (lado de onde sopra o vento) para sotavento (lado oposto ao lado do qual sopra o vento), tanto a favor quanto contra o declive do terreno. Esse tipo de transporte é mais comum em desertos ou planícies costeiras e mais raro em planícies aluviais e de regiões periglaciais.

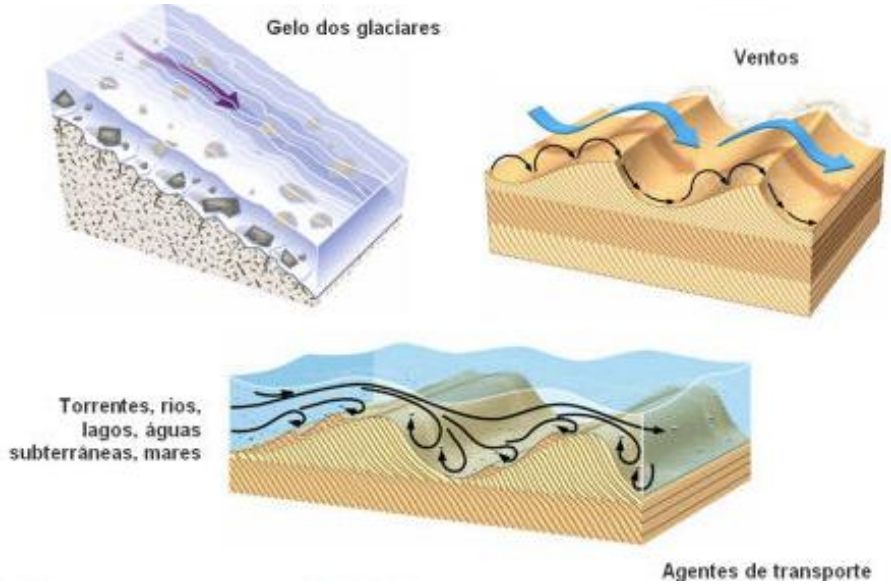
As geleiras favorecem o deslocamento do material sedimentar declive abaixo a partir de um vale glacial. Usualmente o material de transporte glacial é individualizado pela grande heterogeneidade granulométrica e composicional e, ademais, os fragmentos são bastante angulosos.

Os movimentos de grandes massas, também denominados fluxos gravitacionais, referem-se aos mecanismos de transporte de sedimentos paralelamente ao substrato, com maior ou menor atuação da gravidade. (32). Os movimentos de massa são de vários tipos, tanto em relação às escalas temporais e espaciais em que se procedem os fenômenos. Ademais, os processos e os produtos ligados a esses fenômenos são de grande importância para a geologia, geomorfologia e geotecnia. A Figura 6 ilustra alguns agentes de transporte.

### 2.1.4 **Deposição**

Os elementos sedimentares depositam-se quando o vento se acalma, as correntes de água se retardam, ou os bordos das geleiras se unem. Esses elementos formam camadas de

Figura 6 - Agentes de Transporte




Fonte: Retirado de (33).

sedimentados nos continentes ou no leito marinho. No oceano ou nos ambientes aquáticos continentais são formados precipitados químicos e conchas fraturadas de organismos mortos que são depositados (26).

## 2.2 FORMAÇÃO DE BACIAS SEDIMENTARES

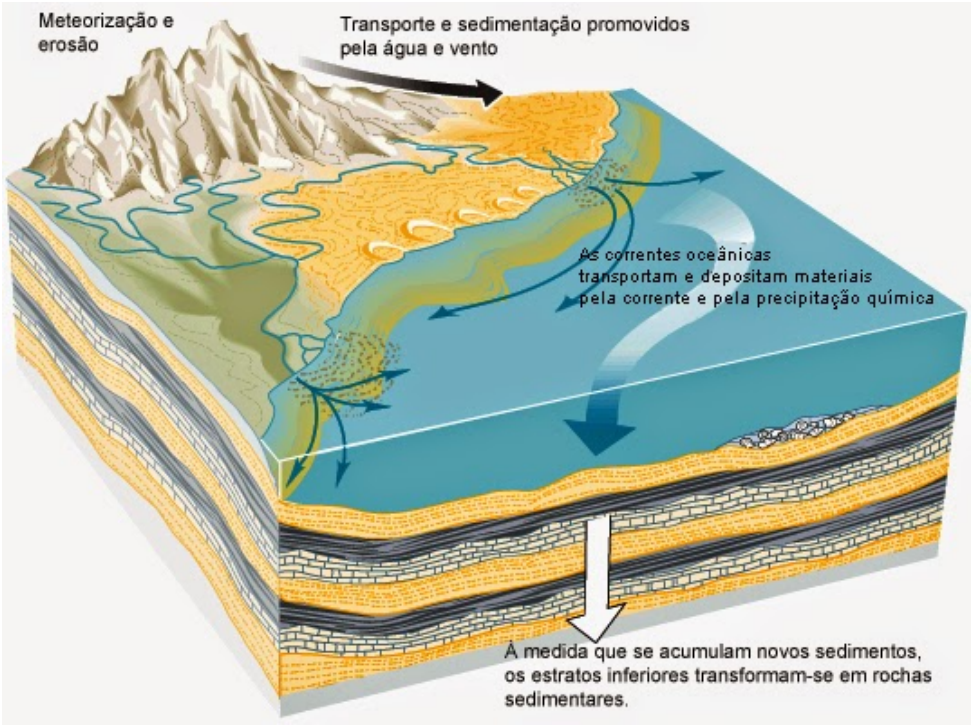
As rochas sedimentares são formadas através da desintegração e decomposição de rochas preexistentes (magmáticas, metamórficas ou sedimentares), devido à ação de intemperismo. O intemperismo desintegra a rocha em partículas menores, que são transportadas pela erosão, sendo depositadas em camadas de sedimentos nas margens continentais. A precipitação bioquímica produz outro tipo de sedimento, como a formação dos recifes de corais. Ao mesmo tempo em que as camadas acumulam-se e vão sendo gradativamente soterradas, elas litificam, consolidando até tornar-se uma rocha sedimentar. A Figura 7 exibe esse processo.

As bacias sedimentares são áreas de extensão considerável, onde ocorre deposição sedimentar. Tais feições geológicas são formadas através de processos tectônicos diversos, onde, em geral, há uma fase inicial de subsidência mecânica, sucedida por processos termais de subsidência, controlada por contínua sedimentação, gerandoi uma espessa acumulação de sedimentos e rochas sedimentares.

## 2.3 DIAGÊNESE E LITIFICAÇÃO

A diagênese pode ser definida como um conjunto de transformações químicas, físicas e biológicas pelas quais passaram os sedimentos desde a sua deposição inicial até

Figura 7 - Formação de Rochas Sedimentares



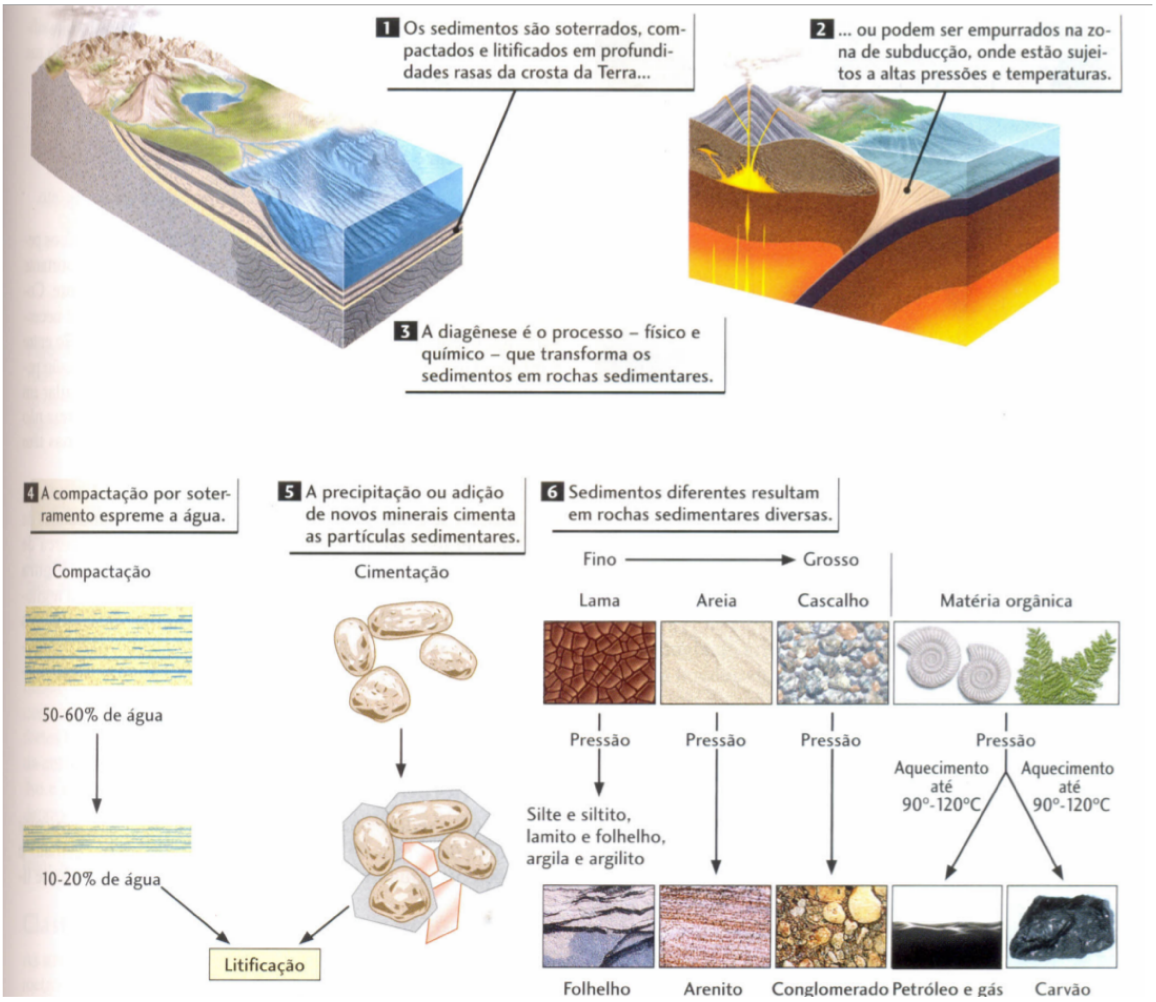

Fonte: Retirado de (26).

após a litificação (34).

Os principais processos durante a diagênese são a compactação, a dissolução, a cimentação e a recristalização diagenética. A compactação é a diminuição do volume e porosidade de um sedimento em função da pressão exercida pelos sedimentos superpostos em uma bacia. A dissolução atinge constituintes característicos ou camadas de sedimentos específicas. A eliminação seletiva de minerais componentes de um sedimento com o tempo pela ação de fluidos intersticiais é um caso de dissolução. A cimentação está associada à precipitação química de diversas substâncias, que preenchem os poros de sedimentos. É um dos processos diagenéticos mais importantes, que transformam um sedimento inconsolidado em rocha sedimentar. A recristalização é a modificação mineralógica e da textura cristalina de componentes sedimentares pela ação de soluções intersticiais em condições de soterramento. A litificação é um processo resultante da compactação e cimentação que consiste na transformação do depósito sedimentar inconsolidado em rocha. Alguns dos processos diagenéticos podem ser visto na Figura 8 (34).

Os fenômenos químicos são os mais comuns da diagênese, compreendendo reações que frequentemente sucedem na precipitação dos minerais autigênicos. Assim, os processos diagenéticos podem também ser controlados pela temperatura, pressão e assembléia de minerais. (35) estabelece três grandes regimes hidrológicos para as águas intraformacionais, relacionados à movimentação e o local de ocorrência em uma bacia: regime meteórico, que abrange as porções mais rasas da bacia, onde a água presente nos sedimentos é oriunda

Figura 8 - Processos Diagenéticos




Fonte: Retirado de (26).

de infiltração superficial até o lençol freático e pode atingir até 2 km de profundidade; regime compactacional, que está relacionado a expulsão da água dos poros em função da compactação e regime termobárico, relacionado as partes mais profundas da bacia, onde o fluido intersticial pode ser gerado a partir da desidratação de alguns argilo-minerais e sais.

Os processos diagenéticos podem ocorrer em várias profundidades e distintas condições em uma bacia sedimentar. A necessidade de se estabelecer uma relação genética dos processos diagenéticos com a profundidade e condições em que estes se desenvolvem fizeram com que (36) sugerissem a divisão do campo diagenético em três estágios principais: eodiagênese, onde os processos diagenéticos são próximos à superfície de sedimentação e a química da água intersticial é controlada pelo ambiente de superfície precedente ao soterramento; mesodiagênese, que retrata o regime de subsuperfície, onde os processos diagenéticos ocorrem no decorrer do soterramento efetivo, com o fluido intersticial já separado da influência superficial, e telodiagênese, que compreende os processos que agem na superfície de erosão, ou nas suas proximidades, em sedimentos que previamente

passaram pelo estágio mesodiagenético, mas que foram expostos por soerguimentos e/ou erosão de camadas suprajacentes.

## 2.4 TIPOS DE ROCHAS SEDIMENTARES

As rochas sedimentares são constituídas, principalmente, por três componentes: terrígenos, aloquímicos e ortoquímicos (1). Estes componentes podem estar misturados em diversas proporções.

**a) Componentes Terrígenos**: São substâncias minerais provenientes da erosão de uma área fora da bacia de sedimentação. Exemplos: quartzo, feldspato, minerais pesados, etc.

**b) Componentes Aloquímicos**: São compostos minerais derivados do retrabalhamento de substâncias químicas precipitadas no interior da própria bacia de sedimentação. Exemplos: conchas de moluscos, oólitos, pisólitos, etc.

**c) Componentes Ortoquímicos**: São os precipitados químicos normais, produzidos na bacia de sedimentação e sem evidências consideráveis de transporte ou agregação.

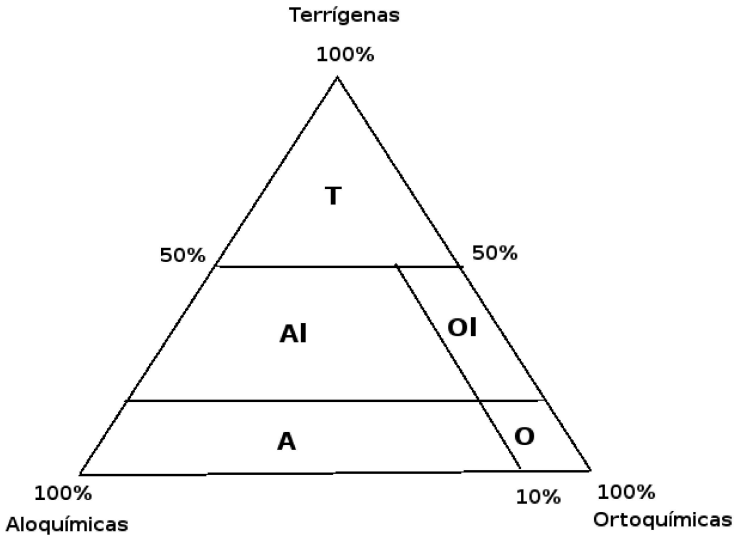
Baseado nos três componentes, as rochas sedimentares podem ser classificadas em (1) (Figura 9):

- **Rochas Terrígenas (T)**: Correspondem de 65% a 75% das seções estratigráficas. Exemplos: folhelhos e arenitos.
- **Rochas Aloquímicas Impuras (AI)**: Abrangem de 10% a 15% das seções estratigráficas aflorantes. Exemplos: folhelos muito fossilíferos e calcário arenoso muito fossilíferos.
- **Rochas Aloquímicas (A)**: Compreendem de 8% a 15% das seções estratigráficas. Exemplos: calcários oóliticos e calcários fossilíferos.
- **Rochas Ortoquímicas Impuras (OI)**: Refazem de 2% a 5% das seções estratigráficas. Exemplo: calcários microcristalinos argilosos.
- **Rochas Ortoquímicas (O)**: Compreendem de 2% a 8% das seções estratigráficas. Exemplos: calcários microcristalinos e dolomitos microcristalinos.

## 2.5 SISTEMAS PETROLÍFEROS

Durante anos de exploração, a indústria petrolífera foi gradativamente constatando que para encontrar reservatórios de petróleo com potencial para exploração era necessário que alguns requisitos geológicos acontecessem simultaneamente em bacias sedimentares.

Figura 9 - Diagrama triangular de classificação geral das rochas sedimentares segundo (1)

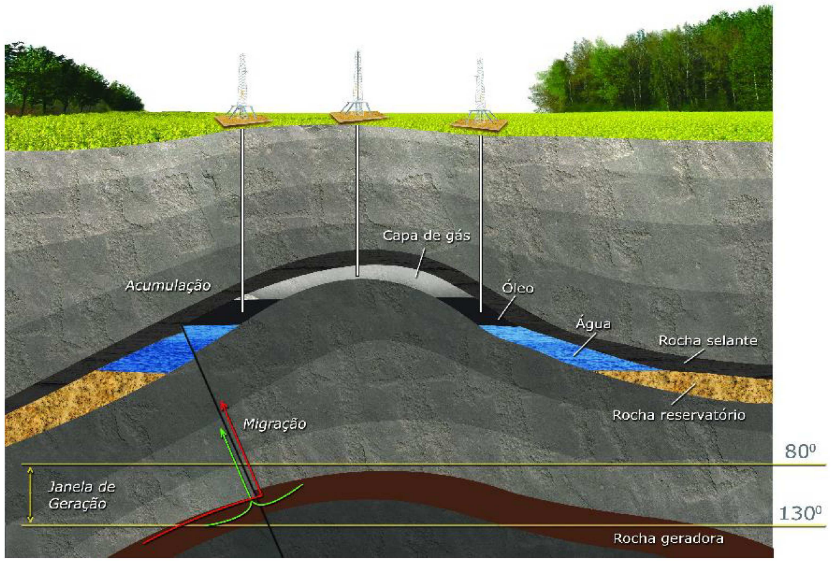



Fonte: Elaborada pelo autor (2020).

O estudo dessas características juntamente com a simulação introdutória de condições ótimas foi denominada sistema petrolífero (37).

Um sistema petrolífero (Figura 10) operante consiste na existência e no funcionamento síncronos de quatro elementos (rochas geradoras maturas, rochas reservatório, rochas selantes e trapas) e dois fenômenos geológicos dependentes do tempo (migração e sincronismo) (37).

Figura 10 - Sistema Petrolífero




Fonte: Retirado de (38).

### 2.5.1 **Rochas Geradoras**

Uma rocha é denominada geradora quando possui grandes volumes de matéria orgânica de qualidade adequada. São estas rochas que, submetidas a apropriadas temperaturas e pressões, geram o petróleo em subsuperfície (37).

### 2.5.2 **Migração**

Visto que o petróleo foi gerado, ele passa a preencher um volume maior do que o querogênio primário na rocha geradora. A rocha se torna supersaturada em hidrocarbonetos e a alta pressão dos mesmos faz com que a rocha-fonte se frature de maneira intensa, proporcionando a expulsão dos fluidos para zonas de pressão mais baixa. O caminho percorrido pelos fluidos petrolíferos, a partir de várias rotas pela subsuperfície, até um local portador de espaço poroso, selado e aprisionado, pronto para armazená-los, é o fenômeno denominado migração (37).

### 2.5.3 **Trapa ou Armadilha**

Os fluidos petrolíferos quando estão em movimento são dirigidos para áreas de pressão mais baixas que os arredores, geralmente posicionadas em situações estruturalmente mais altas que as vizinhanças. As configurações geométricas das estruturas das rochas sedimentares que permitem a focalização dos fluidos migrantes nos arredores para locais elevados são denominadas de trapas ou armadilhas (37). As trapas não permitem que os fluidos petrolíferos escapem futuramente, obrigando-os a se acumularem lá.

### 2.5.4 **Rochas Reservatórios**

Uma rocha é definida reservatório se a porosidade e permeabilidade são adequadas à acumulação de hidrocarbonetos. As rochas reservatórios dividem-se em dois grandes grupos: carbonáticos e siliciclásticos (39).

#### 2.5.4.1 *Siliciclásticas*

As rochas reservatórios siliciclásticas são usualmente arenitos e conglomerados, que evidenciam antigos ambientes sedimentares de alta energia. Os carbonatos são um dos cimentos diagenéticos mais atuantes nas rochas siliciclásticas, o que é decisivo para identificar a qualidade destas rochas como reservatórios. O cimento é o material que precipita-se quimicamente, ocupando frações ou todos os espaços porosos, atingindo os valores porosidade e a permeabilidade das rochas (39).

#### 2.5.4.2 *Carbonáticas*

As rochas reservatórios carbonáticas são formados principalmente por carbonatos provindos de processos biológicos e bioquímicos, ou seja, de origem ou influência orgânica, apesar da precipitação inorgânica de carbonato de cálcio ($CaCO_3$) a partir da águas marinhas também seja um importante processo.

### 2.5.5 **Rochas Selantes**

Quando os fluidos petrolíferos estão no interior de uma trapa eles devem encontrar uma situação de impermeabilização tal que os impeça de escaparem. Normalmente, esta condição é proporcionada por rochas selantes, localizadas acima das rochas reservatório, que impedem a migração vertical dos fluidos, o que faz com que se forme uma acumulação petrolífera (39).

### 2.5.6 **Sincronismo**

Sincronismo é o processo que faz com que as rochas geradoras, reservatórios, selantes, trapas e migração se originem e se desenvolvam em uma escala de tempo apropriada para a geração de acumulações de petróleo. Dessa forma, uma vez iniciada a formação de hidrocarbonetos dentro de uma bacia sedimentar, depois um soterramento apropriado, o petróleo expulso da rocha geradora deve buscar rotas de migração já existentes. Assim, a trapa já deve estar gerada para aproximar os fluidos migrantes, os reservatórios porosos já devem estar depositados e pouco soterrados para perderem seus atributos permo-porosos primitivos, e as rochas selantes já devem existir para impermeabilizar a trapa (39).

## 2.6 POROSIDADE E PERMEABILIDADE

A porosidade é uma propriedade petrofísica das rochas sedimentares e está relacionada com a porcentagem (em volume) de vazios (poros) de uma rocha e expressa a fração do volume total de uma rocha que pode ser ocupada por fluidos. Na maior parte dos reservatórios a porosidade varia de 10% a 20% (40). A quantidade, tamanho, geometria e grau de conectividade dos poros controlam de forma direta a produtividade do reservatório. A porosidade de uma rocha é calculada diretamente, em amostras de testemunho, ou de forma indireta, por meio de perfis elétricos. A porosidade de uma rocha pode ser classificada como insignificante ($0 - 5\%$), pobre ($5 - 10\%$), regular ($10 - 15\%$), boa ($15 - 20\%$), ou muito boa ($> 20\%$) (40).

A permeabilidade é uma propriedade petrofísica das rochas sedimentares que mede a capacidade da rocha de transmitir fluidos a partir dos seus poros, sem deformar sua estrutura ou acarretar deslocamento relativo de suas partes. A permeabilidade é expressa em Darcys ($D$) ou milidarcys ($md$). Supervisionada pela quantidade, geometria e grau

de conectividade dos poros, a permeabilidade de uma rocha é calculada diretamente, em amostras de testemunho e pode ser classificada como baixa ($< 1md$), regular ($1 - 10md$), boa ($10 - 100md$), muito boa ($100 - 1000md$) e excelente ($> 1000md$). A maior parte dos reservatórios possui permeabilidades de 5 a 500md (40).

## 2.7 AMBIENTES DE SEDIMENTAÇÃO, LITOLOGIA, FÁCIES E PETROFÁCIES SEDIMENTARES

Nessa seção serão apresentados os conceitos referentes ambientes de sedimentação, litologia, fácies e petrofácies sedimentares que objetos de estudo nos quais a metodologia tem interesse em identificar.

### 2.7.1 **Ambientes de Sedimentação**

Os ambientes de sedimentação podem ser definidos como partes da superfície terrestre com propriedades físicas, químicas e biológicas bem definidas e distintas das apresentadas pelas áreas vizinhas (25). Estas propriedades compreendem uma gama de variáveis que se interagem, determinando os atributos dos distintos ambientes de sedimentação.

O estudo das sequências sedimentares é parte de uma pesquisa mais ampla na análise de uma bacia sedimentar. A identificação de ambientes de sedimentação não é só de grande interesse para pesquisadores, mas também na prospecção de recursos naturais associados às rochas sedimentares, tais como petróleo, carvão, calcário, fosfato, entre outros que ocorrem em ambientes sedimentares específicos. Maiores detalhes do processo de formação desses recursos podem ser encontrados em (41) e (25).

### 2.7.2 **Litologia**

O termo litologia refere-se à composição ou tipo de rocha. Compreende a descrição de rochas em afloramento ou amostra de mão, baseada em características como a cor, textura, estrutura, composição mineralógica ou granulometria (42). Sua identificação é fundamental para a caracterização do reservatório devido às propriedades físicas e químicas da rocha e afeta a resposta de cada instrumento utilizado para medir as propriedades de formação.

A identificação da litologia ocorre através de métodos diretos ou indiretos. Os métodos diretos são realizados pela obtenção de uma amostra física do reservatório. Este é o caminho mais preciso para determinar a litologia, mas para chegar a essa amostra física nem sempre é fácil.

Métodos indiretos fazem uso de perfis de poços que medem as propriedades físicas de formações geológicas e fluidos que fornecem a maioria dos dados de subsuperfície. Além

de sua importância na tomada de decisões, eles também são ferramentas inestimáveis para mapeamento e identificação de litologias. No entanto, os métodos indiretos não possuem a mesma eficácia que os métodos diretos.

### 2.7.3 **Fácies Sedimentares**

Gressly (43) percebeu, enquanto trabalhava na região dos Alpes, que litologias e fósseis diferentes poderiam ocorrer na mesma época. A partir dessa observação, ele propôs o termo *fácies* para definir unidades de rochas caracterizadas por propriedades litológicas (composição, textura, estruturas sedimentares e cor) e paleontológicas (conteúdo e registro fossilífero) semelhantes.

Fácies sedimentar pode ser definida como uma parte restrita em área de uma determinada unidade estratigráfica, que exibe características diferentes significantes das demais partes da unidade (44).

### 2.7.4 **Petrofácies**

As petrofácies podem ser definidas como uma técnica para o reconhecimento das heterogeneidades de um reservatório auxiliando na análise da evolução diagenética do mesmo. Segundo De Ros (45), petrofácies são caracterizadas pela combinação de estruturas específicas de deposição, texturas e composição primária, com processos diagenéticos dominantes. A combinação de aspectos texturais primários e composicionais com processos e produtos diagenéticos específicos correspondem a variação de valores definidos de porosidade e permeabilidade, bem como as logs características e as assinaturas sísmicas.

O reconhecimento de petrofácies (45), inicia com uma petrografia detalhada de amostras representativas da área estudada. Uma análise quantitativa através da contagem de 300 ou mais pontos é importante, mas não é sempre essencial para o reconhecimento de petrofácies, pois, em alguns casos, os padrões principais podem ser identificados diretamente a partir de uma descrição qualitativa. As amostras são separadas em grupos, primeiro de acordo com estruturas sedimentares e textura. As amostras devem ser assim agrupadas considerando-se a superposição de atributos de deposição (estrutura e textura) com as principais categorias de composição primária e com a distribuição dos processos diagenéticos mais influentes. Os atributos com maior impacto sobre a porosidade e permeabilidade são reconhecidos, e petrofácies preliminares são atribuídas. O agrupamento de amostras nas mesmas petrofácies assume que elas exibem comportamento petrofísico semelhante. As petrofácies preliminarmente definidas são confrontadas com parâmetros quantitativos petrofísicos e petrográficos, utilizando ferramentas estatísticas e redes neurais (45). Os valores limites são então definidos para os atributos texturais e composicionais influentes que restringem as petrofácies.

## 2.8 GEOLOGIA SEDIMENTAR APLICADA

As tradicionais aplicações da geologia sedimentar estão relacionadas a prospecção de combustíveis fósseis (petróleo e carvão mineral) e depósitos de minerais. As motivações econômicas em torno da busca e exploração de combustíveis fósseis são responsáveis pelo avanço das pesquisas em geologia sedimentar. A partir da década de 50 observou-se um crescimento relevante dos grupos de pesquisas de empresas petrolíferas, que constataram o quão era necessário melhorar as técnicas de interpretação que levassem ao prognóstico mais rápido e preciso das tendências de distribuição da permoporosidade e dos reservatórios de subsuperfície (25).

A geologia sedimentar encontra vasta aplicação como fonte de subsídios na prospecção e exploração de recursos naturais não-renováveis, como petróleo e o gás natural, e recentemente em geologia ambiental e em engenharias, chegando em pesquisas criminalísticas (25).

### 2.8.1 **Petróleo e Gás Natural**

O petróleo originou-se através da matéria orgânica soterrada juntamente com sedimentos lacustres ou marinhos. Possui estado físico oleoso e normalmente densidade menor do que da água. Sua composição química é formada por combinação de moléculas de hidrocarbonetos. Seu uso teve início na Antiguidade, nas formas de betume, asfalto, entre outras. Era muito utilizado como impermeabilizante ou inflamável com finalidades bélicas. No século XIX, o querosene passou a ser empregado para substituir o óleo de baleia na iluminação pública e, dando início ao uso comercial do petróleo e a produção por um poço de 21 metros de profundidade perfurado na Pensilvânia (4).

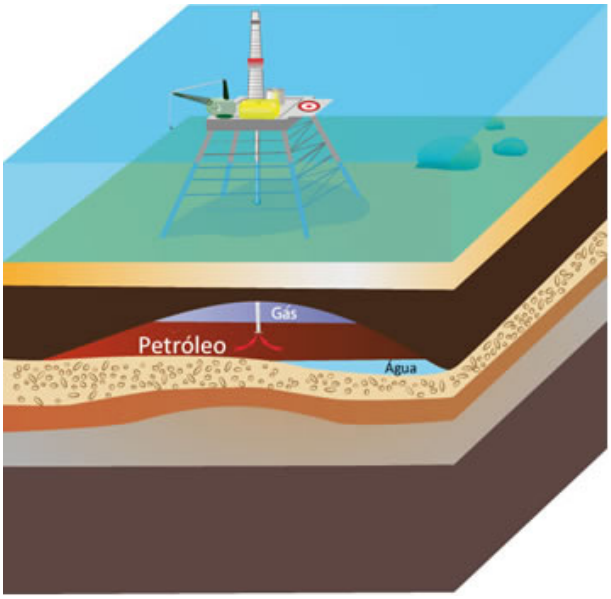
Com o desenvolvimento dos motores a gasolina e diesel, o uso do petróleo passou por um processo de intensificação que vem até os dias atuais. No Brasil, o primeiro direito de extração foi concedido em 1858, para exploração de mineral betuminoso, na Bahia. Pesquisas e perfurações de poços foram realizadas de maneira esparsa na Bahia, Alagoas e São Paulo até que encontrou-se petróleo a 210 metros de profundidade em Lobato, Bahia. O primeiro campo comercialmente viável foi descoberto em 1941, em Candeias (46). Em 1954 foi criada a Petrobras – Petróleo Brasileiro S.A instituindo o monopólio estatal. Na década de 70 foram feitas descobertas na Bacia de Campos que é uma das bacias que lidera a produção até os dias de hoje. No ano de 1997 tem-se o fim do monopólio estatal na E,P&D e a implantação da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), órgão vinculado ao Ministério das Minas e Energia que responde pela política nacional para o setor.

No final de 2007, a Petrobras anunciou a descoberta de uma reserva de 5 a 8 bilhões de barris de petróleo abaixo de uma camada de rocha salina (pré-sal) no campo de Tupi, na Bacia de Santos. O óleo encontrado é considerado de boa qualidade. A partir desse

marco aumentou-se ainda mais os investimentos em P&D na área de Petróleo e Gás (4).

O gás natural é um combustível fóssil não renovável composto por uma mistura de hidrocarbonetos, principalmente metano ($CH_4$). O gás natural é encontrado em jazidas ou depósitos subterrâneos, que em geral estão associados ao petróleo, uma vez que esses dois combustíveis fósseis passam pelo mesmo processo de formação e se acumulam no mesmo tipo de ambiente. Esse combustível gasoso, após ser tratado e processado, apresenta grande teor energético, sendo muito aproveitado nas indústrias para a geração de energia elétrica. O gás natural é utilizado também como combustível industrial, veicular e doméstico, como matéria-prima nas indústrias siderúrgica, química e de fertilizantes (4). A Figura 11 mostra o posicionamento do petróleo e do gás.

Figura 11 - Posicionamento Gás e Petróleo




Fonte: Retirado de (38).

# 3 REVISÃO BIBLIOGRÁFICA

Técnicas inteligentes como redes neurais, raciocínio difuso e computação evolucionista para análise e interpretação de dados são ferramentas cada vez mais relevantes para extrair informações através de dados e transformar estas informações em conhecimento. Na indústria de petróleo e gás, essas técnicas inteligentes podem ser usadas para análise de incertezas, avaliação de risco, fusão e mineração de dados, análise e interpretação de dados e descoberta de conhecimento, a partir de dados diversos como sísmica 3D, dados geológicos, log de poços, e dados de produção. Além disso, essas técnicas podem ser uma chave para a localização e produção das reservas de petróleo e gás remanescentes. Técnicas podem ser usadas como uma ferramenta para: redução do risco de exploração, redução do custo de exploração e produção, melhoria da recuperação através de uma produção mais eficiente, estendendo a vida útil dos poços produtores (47).

Nos últimos anos, na indústria do petróleo, pode-se assistir a um grande aumento no volume de dados. Isso é causado pelo aumento da taxa de amostragem, maior *offset* e maior aquisição de registros, levantamentos multicomponentes, sísmica 4D e, mais recentemente, à possibilidade de gravação contínua em "campos de petróleo instrumentados". Assim, serão necessárias, além de técnicas eficientes para processar grandes volumes de dados, técnicas automatizadas para refinamento das informações, selecionando os tipos de eventos desejados ou interpretações automatizadas. A lógica fuzzy e as redes neurais mostraram ser ferramentas eficazes para tais aplicações (48). Os métodos de redução de dimensionalidade de dados têm uma importância crescente para esses grandes volumes de dados, tanto para processamento, análise e visualização eficientes de transmissão rápida de dados e armazenamento de dados econômico. O maior impacto dos avanços nas técnicas de redução de dimensionalidade será realizado quando os pesquisadores tiverem a capacidade de processar e analisar completamente os dados no domínio compactado. Isso possibilitará o processamento intensivo de grandes volumes de dados em uma fração do tempo, resultando em enormes reduções de custos. A mineração de dados é outra alternativa que ajuda a identificar a parte em que as informações importantes dos grandes volumes de dados se concentra. Em estudos recentes foi demonstrado que métodos de aprendizagem de máquina combinados com métodos mais convencionais, como o autovalor ou a análise de componentes principais, são muito úteis (47).

Aminzadeh (49) mostrou uma evolução dos métodos de detecção inteligentes aplicados na indústria do petróleo, usando dados geológicos e geofísicos, destacando aplicações de *soft computing* e inteligência artificial. Realizou uma discussão a respeito da inteligência humana e da inteligência de máquina e citou a necessidade de combinar inteligência humana e de máquina. Enfatizou o papel dos métodos de *soft computing* (redes neurais, lógica fuzzy e computação evolucionista) que podem desempenhar no estabelecimento de inteligência híbrida para abordar problemas de exploração e produção.

## 3.1 SISTEMAS E FERRAMENTAS PARA A CARACTERIZAÇÃO DE RESERVATÓRIOS

Uma metodologia integrada foi desenvolvida, por (50), para identificar relações não-lineares e mapeamento entre dados sísmicos 3D e dados de registro de produção. O método utiliza técnicas convencionais, como reconhecimento de padrões geo estatísticos e clássicos, em conjunto com técnicas modernas, como *soft computing*. Utilizou técnicas de agrupamento para reconhecer a localização ideal de um novo poço com base em dados sísmicos 3D e dados de registro de produção. Nikravesh (51) propôs uma ferramenta de caracterização inteligente de reservatórios, denominada IRESC, baseada em *soft computing* que é um conjunto de metodologias inteligentes de computação usando neuro computação, raciocínio fuzzy e computação evolucionista.

Abel *et al.* (52) descreveram o sistema PetroGrapher, uma aplicação de banco de dados para dar suporte a análise petrográfica, interpretação de rochas de reservatórios de petróleo, e gerenciamento de dados relevantes, usando recursos da tecnologia de sistemas baseados em conhecimento e da tecnologia de banco de dados.

De Ros e Goldberg (53) desenvolveram uma ferramenta para caracterização e predição de qualidade, chamada *Reservoir Petrofacies*, onde a determinação de petrofácies de reservatórios é iniciada pelo reconhecimento de petrofácies preliminares através de uma descrição metódica de características em amostras coletadas a partir de distribuição representativa, seguida de reconhecimento de características com maior impacto na porosidade e permeabilidade. As petrofácies preliminares são então comparadas com parâmetros petrofísicos e petrográficos quantitativos usando uma rede neural artificial.

Fuzzy Petroleum Prediction (FPP) é um sistema especializado que foi projetado por (54). São utilizados cinco fatores de previsão de petróleo: temperatura, pressão, densidade do petróleo bruto, gravidade e densidade do gás. O sistema especialista de previsão de petróleo aplica-se a trinta poços no campo petrolífero de Daqing e outras fontes de petróleo.

Tabela 1 – Resumo do levantamento dos trabalhos relacionados - Sistemas/Ferramentas para caracterização de reservatórios.

| Trabalho | Ano | Sistema/Ferramenta | Utilidade |
|---|---|---|---|
| (50) | 2001 | técnicas convencionais e *soft computing* | identificar relações não-lineares e mapeamento entre dados sísmicos 3D e dados de registro de produção |
| (51) | 2004 | IRESC | caracterização inteligente de reservatórios |
| (52) | 2004 | PetroGrapher | dar suporte a análise petrográfica, interpretação de rochas de reservatórios de petróleo, e gerenciamento de dados relevantes |
| (53) | 2007 | Reservoir Petrofacies | caracterização e predição de qualidade de um reservatório de petróleo |
| (54) | 2013 | Fuzzy Petroleum Prediction | prever temperatura, pressão, densidade do petróleo bruto, gravidade e densidade do gás de um reservatório. |

Fonte: Elaborada pelo autor (2020).

## 3.2 CLASSIFICAÇÃO LITOLÓGICA

Cracknell e Reading (55) realizaram uma comparação de cinco algoritmos de aprendizagem de máquina: *Naive Bayes* (NB), *K-Nearest Neighbors* (K-vizinhos mais próximos - KNN), *Random Forest* (Florestas Aleatórias - RF), *Support Vector Machines* (Máquinas de Vetores de Suporte - SVM) e *Artificial Neural Network* (Redes Neurais Artificiais - ANN) para classificação litológica supervisionada usando dados geofísicos de detecção remota amplamente disponíveis e espacialmente restritos.

Oliveira (56) investigou o uso de técnicas de classificação automática aplicadas ao problema de classificação de litofácies em amostras de perfis de três poços de uma reserva da Amazônia. Utilizou os métodos SVM, KNN, Multilayer Perceptron e Regressão Logística. Comparou o desempenho de classificadores individuais frente à combinação dos mesmos através do voto majoritário. Concluiu que combinar classificadores em um sistema de votação majoritário apresenta desempenho superior ao uso de classificadores individuais.

*Support Vector Machines* (Máquinas de Vetores de Suporte - SVM) foi utilizado por (57) para prever a litologia a partir de dados de atributos sísmicos invertidos e registros de poços baseados em estudos petrográficos de testemunhos litológicos em um reservatório carbonático heterogêneo no Irã.

Horrocks *et al.* (58) apresentaram um estudo onde algoritmos de aprendizado de máquina para classificação de litologia no contexto da exploração de carvão. Utilizaram dados de sete poços contendo 19 registros comuns para comparar três algoritmos de aprendizado de máquina para classificação litológica: uma ANN e uma SVM , que foram reportadas como tendo um bom desempenho em registros de rede, e um classificador NB, que tem baixa complexidade computacional.

Um estudo foi introduzido por (59) onde a Análise Discriminante com *kernel* Fisher e uma Análise Discriminante Linear melhorada foram aplicadas para identificar litologia.

A confiabilidade da previsão do tipo de rocha usando densidade de um poço, resposta de raios gama e medidas de susceptibilidade magnética foi avaliada em Victoria, Sudbury, Ontário (60). Uma rede neural supervisionada, treinada usando informações litológicas do furo FNX1182, resultou em 64% dos tipos de rochas sendo classificados corretamente quando comparados com a classificação produzida por geólogos durante o registro do testemunho.

Xie *et al.* (2) realizaram uma comparação de cinco algoritmos de aprendizado de máquina NB, SVM, ANN, RF e *Gradient Tree Boosting* (GTB). Esses algoritmos possuem como entrada dados do registro do poço de dois campos de gás na Bacia de Ordos para identificar as classes litológicas.

Saporetti *et al.* (61) integraram o *Gradient Boosting* (GB) com a Evolução Diferencial (DE) para identificação da litologia de formação usando dados do campo de gás Daniudui e do campo de gás Hangjinqi.

Para identificar sua litologia, Min *et al.* (62) propuseram utilizar o deep learning para estabelecer a *Deep Belief Network* (DBN) por meio de logs de poços.

He *et al.* (63) combinaram uma Rede Neural Profunda (DNN) com uma abordagem de superamostragem denominada MAHAKIL para identificar litologia usando dados de logs de um reservatório de gás natural de arenito.

Liu *et al.* (64) propuseram um método de aprendizado de transferência denominado máquina de aprendizado extremo de adaptação à articulação de desvio de dados (DDJA-ELM) para aumentar a precisão do ELM aplicado a novos poços.

Tabela 2 – Resumo do levantamento dos trabalhos relacionados - Métodos de Classificação.

| Referência | Ano | Técnica(s) Utilizada(s) | Melhor Resultado Apresentado (Acurácia) |
|---|---|---|---|
| (55) | 2014 | NB, KNN, RF, SVM, RNA | 76,2% |
| (56) | 2014 | KNN, Regressão Logística, Perceptron, SVM, Tree(J48), Votação | 79,5% |
| (57) | 2015 | SVM | 75,4% |
| (58) | 2015 | SVM, ANN, NB | 73,2% |
| (59) | 2016 | Análise Discriminante | 80,7% |
| (60) | 2016 | ANN | 64% |
| (2) | 2018 | NB, SVM, ANN, RF, GTB | 85% |
| (61) | 2019 | GB+DE | 82,8% e 85,8% |
| (62) | 2020 | DBN | 94,8% |
| (63) | 2020 | DNN | 70% |
| (64) | 2020 | DDJA-ELM | 96,9% |

Fonte: Elaborada pelo autor (2020).

## 3.3 ANÁLISE DE AGRUPAMENTO EM DADOS PETROGRÁFICOS E PETROFÍSICOS

Horrocks *et al.* (65) desenvolveram uma técnica denominada Clustered Neighbor-Joining (CNJ) que é baseada no emprego de métodos de agrupamento seguido de uma técnica para construir árvores filogenéticas. Neste trabalho foram utilizados dados Y_DNA haplogroups.

Ouafi *et al.* (66) apresentaram um método baseado em clusterização para prever fácies. A técnica é baseada em um algoritmo de determinação de fácies seguido de codificação de eletrofácies usando dados petrofísicos, especialmente (GR, RHOB, THOR, POTH) descrições de testemunhos detalhadas.

Mahmoodi e Smith (67) analisaram a densidade do poço, a radioatividade gama e as medidas de susceptibilidade magnética em cinco furos de sonda na localidade de Victoria (localizada na faixa sul da bacia de Sudbury) para identificar unidades físicas homogêneas. O algoritmo de clusterização Fuzzy K-means foi usado para classificação não supervisionada dos dados.

Silva (68) propôs uma metodologia para caracterização do reservatório carbonático Albiano do Campo B da Bacia de Campos a partir de dados de perfis de poço e dados de petrofísica básica de laboratório. Através da estimação de valores mais confiáveis de porosidade, permeabilidade e saturação de água de um reservatório antes da realização

de testes em laboratório. Para atingir este objetivo utilizou o módulo "Cluster Analysis for Rock Typing"do software Interactive Petrophysics para dividir o poço em eletrofácies. Para cada uma destas eletrofácies foi determinada uma equação através da técnica de regressão linear múltipla para encontrar a porosidade e a permeabilidade usando os dados de laboratório como saída e os dados de perfis de poço como entrada.

Methe *et al.* (69) empregaram análise de agrupamento para obter informações sobre litologia. Testaram algoritmos de agrupamento com diferentes abordagens (agrupamento hierárquico de Ward, K-Means, Mean-Shift e DBSCAN) em dados geofísicos de uma perfuração.

Oloso *et al.* (70) propuseram uma solução híbrida de agrupamento K-Means e Redes Funcionais (FN) para prever as propriedades pressão-volume-temperatura do petróleo bruto. K-Means é usado para gerar grupos do conjunto de dados de entrada antes de usar redes funcionais para executar a previsão das variáveis de destino reais.

Wang *e*t al. (71) usaram o método de agrupamento KNN otimizado com base na distância de cosseno ponderada para identificar litologia em estratos mesozóicos do campo de Gaoqing, depressão de Jiyang.

Saporetti *e*t al. (72) analisaram o uso de abordagens de agrupamento para auxiliar a análise de dados petrográficos e ferramentas de análise filogenética para entender o processo diagenético que ocorreu durante a formação de rochas sedimentares.

Abdideh e Ameri (73) aplicaram um método de agrupamento baseado em gráficos de alta resolução (MRGC) em parâmetros petrofísicos e geológicos para a separação de eletrofácies de uma sequência de reservatório de gás carbonato.

Tabela 3 – Resumo do levantamento dos trabalhos relacionados - Métodos de Agrupamento.

| Referência | Ano | Técnica Utilizada | Utilidade |
|---|---|---|---|
| (65) | 2012 | Clustered Neighbor-Joining | construir árvores filogenéticas |
| (66) | 2014 | —— | prever fácies |
| (67) | 2015 | Fuzzy K-Means | identificar unidades físicas homogêneas |
| (68) | 2016 | K-Means | identificar eletrofácies |
| (69) | 2017 | Ward, K-Means, Mean-Shift, DBSCAN | obter informações sobre litologia |
| (70) | 2017 | K-Means e redes funcionais (FN) | prever as propriedades pressão-volume-temperatura do petróleo bruto |
| (71) | 2018 | KNN | identificar litologia |
| (72) | 2018 | análise de dados petrográficos e estudo de diagênese. | |
| (73) | 2019 | MRGC | identificar eletrofácies |

Fonte: Elaborada pelo autor (2020).

## 3.4 MÉTODOS INTELIGENTES EM SISTEMAS DE CARACTERIZAÇÃO

Mohaghegh *et al.* (74) demostraram que o desenvolvimento de logs de imagens magnéticas virtuais razoavelmente precisos é possível. Utilizaram redes neurais artificiais para esse procedimento.

Rodolfo *et al.* (75) apresentaram uma metodologia que visa resolver o problema inverso da previsão de propriedades de reservatórios em intervalos/poços, utilizando técnicas de *soft computing* e análise estatística multivariada para obter modelos petrofísicos melhores. O primeiro passo desta metodologia é pré-processar os dados, para os quais gráficos Q-Q são usados com elipses de confiança de 95% para controle de qualidade. Depois disso, o número ideal de variáveis independentes são identificados usando componentes principais, análise fatorial e conceitos de lógica Fuzzy. Em seguida, as redes neurais são

aplicadas para modelar quaisquer variáveis de destino.

Mohaghegh *et al.* (76) apresentou uma metodologia de inversão sísmica inteligente para alcançar uma correlação desejável entre sinais sísmicos de frequência relativamente baixa e os dados de log de linhas fixas de frequência muito maior. A rede neural de regressão generalizada (GRNN) foi utilizada para construir dois modelos de correlação independentes entre: 1) Perfil sísmico de superfície e perfil sísmico vertical (VSP), 2) VSP e logs de poço. Depois de gerar VSP's virtuais a partir da superfície sísmica, os logs de poço são previstos usando a correlação entre os logs do VSP e do poço.

Anifowose e Abdulraheem (77) demonstraram o desempenho de dois modelos híbridos como ferramentas de Inteligência Computacional na previsão da porosidade e permeabilidade de reservatórios de óleo e gás. A modelagem híbrida foi baseada na combinação de três técnicas existentes de Inteligência Artificial: Redes Funcionais, Sistema Lógico Fuzzy Tipo 2 e Máquinas de Vetores de Suporte, utilizando seis conjuntos de dados.

Araújo (78) desenvolveu um sistema inteligente, capaz de obter automaticamente a porosidade efetiva, em camadas sedimentares, a partir de um banco de dados construído com informações do Radar de Penetração no Solo (GPR). O sistema inteligente foi construído para modelar a relação entre a porosidade e os atributos eletromagnéticos do GPR. Estimou-se a porosidade utilizando modelo de rede neural artificial (Multilayer Perceptron (MLP)) e regressão linear múltipla.

Panja *et al.* (79) usaram três modelos de Inteligência Artificial (Máquina de Vetor de Suporte pelo Mínimo Quadrado (LSSVM), Redes Neurais Artificiais (RNA) e Modelo de Superfície de Resposta (RSM)) para prever a produção de hidrocarbonetos a partir de poços fraturados hidraulicamente. A rotina de otimização de enxame de partículas (PSO) foi usada nos modelos para obter os parâmetros ótimos.

O modelo petrofísico de Xu-Payne é um método comumente usado para prever a $V_s$ (velocidade da onda S). Zhang *e*t al. (80) empregaram o método de aprendizado de máquina Rede Neural de Memória de Longo Prazo (LSTM) para melhorar o fluxo de trabalho petrofísico tradicional.

Tabela 4 – Resumo do levantamento dos Métodos inteligentes em sistemas de caracterização.

| Referência | Ano | Utilidade |
|---|---|---|
| (74) | 2001 | desenvolveu logs de imagens magnéticas virtuais |
| (75) | 2002 | utilizou técnicas de *soft computing* e análise estatística multivariada para obter modelos petrofísicos melhores. |
| (76) | 2005 | desenvolveu uma metodologia de inversão sísmica inteligente para alcançar uma correlação desejável entre sinais sísmicos de frequência relativamente baixa e os dados de log de linhas fixas de frequência muito maior |
| (77) | 2011 | demonstrou o desempenho de dois modelos híbridos como ferramentas de Inteligência Computacional na previsão da porosidade e permeabilidade de reservatórios de óleo e gás. |
| (78) | 2013 | desenvolveu um sistema inteligente, capaz de obter automaticamente a porosidade efetiva, em camadas sedimentares |
| (79) | 2017 | usou modelos de Inteligência Artificial para prever a produção de hidrocarbonetos a partir de poços fraturados hidraulicamente |
| (80) | 2020 | usou LSTM para melhorar a previsão de $V_s$ |

Fonte: Elaborada pelo autor (2020).

# 4 MATERIAIS E MÉTODOS

Neste capítulo será apresentados os materiais utilizados, que são as bases de dados que serviram de entrada para a metodologia proposta e os métodos que compõem essa metodologia.

## 4.1 BASES DE DADOS UTILIZADAS

As bases de dados utilizadas para testar a metodologia proposta são divididas em petrofísicas e petrográficas. Esses conceitos e as respectivas bases serão apresentadas a seguir.

### 4.1.1 **Dados Petrofísicos**

O conhecimento das características petrofísicas das rochas reservatório faz-se necessário para compor uma interpretação geofísica e geológica adequadas, fornecendo assim melhores perspectivas de previsões de funcionamento dos reservatórios.

A definição de litologia em poços de petróleo por meio de múltiplos perfis de análise geofísica tem um papel importante no processo de caracterização do reservatório. A partir da litologia, pode-se gerar modelos que, por sua vez, são a base a partir da qual cálculos petrofísicos são feitos, e em seguida, podem ser utilizados em simuladores de fluxo para compreender e estudar o comportamento de um campo de petróleo.

Para as empresas petrolíferas arquivos com a formatação ASCII é mais interessante, então os dados geralmente possuem este padrão. Como característica pode-se citar o fato de terem caracteres letras, números e símbolos legíveis, de acordo com um padrão aceito mundialmente. Seu formato binário representa um padrão aberto, o que faz com que sejam facilmente lidos e manipulados. Os arquivos ASCII não possuem um formato próprio, eles são um modelo aberto de dados. Os arquivos *Log ASCII Standard* (LAS) têm extensão .las e são arquivos ASCII. Programas de uso geral podem ler e manipular estes arquivos, não existindo a necessidade de aplicativos específicos ou proprietários. Dessa forma, arquivos ASCII podem ser a base sobre onde são adicionadas semântica e sintaxe, para a construção de um padrão de troca de dados particulares, como o que acontece com o padrão LAS.

O LAS foi apresentado em 1990 pela Sociedade Canadense de Log de Poços (Canadian Well Logging Society) para facilitar a troca de informações digitais de poços entre distintas companhias e clientes. Por ser um arquivo em formato ASCII, os dados em LAS podem ser importados e exportados para qualquer plataforma. Embora construído para complementar os formatos mais robustos já existentes da época, o LAS se tornou desde então o formato mais usado para transferência digital de dados de logs de poços e um verdadeiro padrão da indústria (81).

Cada arquivo LAS inicia com uma linha de título, marcada pelo uso do ~ no início da linha. Além de conter uma grande quantidade de seções já definidas, sendo possível construir novas seções estabelecidas pelo usuário. Algumas seções são obrigatórias, como ~Version, ~Well, ~Curve e ~ASCII, sendo ainda fundamental que ~Version e ~Well sejam as primeiras seções do documento, respectivamente (81). O caractere # é utilizado para comentário. Todos os mnemônicos (como STRT, STOP, STEP, etc) são obrigatórios em uma seção de poço. Contudo, apenas os mnemônicos STRT, STOP e STEP têm a obrigatoriedade de ter um valor associado. Para compreender melhor a estrutura, são apresentados os tipos primários de seções, bem como os tipos de linhas das seções. A Figura 12 mostra um exemplo de um arquivo LAS.

Figura 12 - Exemplo de um arquivo LAS.

```
~Version
#MNEM .UNIT                 VALUE  : DESCRIPTION
 VERS .                       2.0  : CWLS LOG ASCII STANDARD - VERSION 2.0
 WRAP .                        NO  : ONE LINE PER DEPTH STEP
~Well
#MNEM .UNIT                     VALUE  : DESCRIPTION
 STRT  .FT                       30.0  : Start Depth
 STOP  .FT                      594.0  : Stop Depth
 STEP  .FT                        0.1  : Step
 NULL  .                       -999.25  : Null Value
 COMP  .                       KGS-OGS  : Company
 WELL  .                     CURRENT # 1  : Well
 FLD   .                                  : Field
 SEC   .                            17  : Section
 TOWN  .                            5S  : Township (e.g. 42S)
 RANG  .                           13E  : Range  (e.g. 25E)
 LOC   .                      5S-13E-17  : Location (Sec Town Range)
 LOC1  .          1515' FSL & 1195' FEL  : Location 1 (quarter calls)
 LOC2  .                         SWNESE  : Location 2 (footages)
 PROV  .                                  : Province
 CTRY  .                             US  : Country
 STAT  .                       OKLAHOMA  : State
 CNTY  .                       PONTOTOC  : County
 API   .                  35-123-23588  : API-Number
 UWI   .                                  : Unique Well ID
 SRVC  .                                  : Service Company
 LIC   .                                  : Licence Number
 DATE  .                     06/10/2008  : Date preferred format is MM/DD/YYYY
 LATI  .                        N34.706  : Latitude
 LONG  .                        W96.638  : Longitude
 GDAT  .                          NAD27  : Geodetic Datum
 X     .                      244102.82  : X or East-West coordinate
 Y     .                     4388921.03  : Y or North South coordinate
 HZCS  .                            UTM  : Horizontal Co-ordinate System
 UTM   .                           15.0  : UTM Location
 STUS  .                                  : Well Status
~Parameter
#MNEM .UNIT           VALUE  : DESCRIPTION
 EGL   .M               775  : Ground Level Elevation
 EKB   .M                    : Kelly Bushing Elevation
 EDF   .F                    : Derrick Floor Elevation
 ERT   .F                    : Rotary Table Elevation
 TDL   .F            3837.0  : Total Depth Logger
 TDD   .              593.6  : Total Depth Driller
 CSGL  .               8.62  : Casing Bottom Logger
 CSGD  .                     : Casing Bottom Driller
 CSGS  .IN                   : Casing Size
 CSGW  .LB                   : Casing Weight
 BS    .IN             3.00  : Bit Size
 DFT   .              WATER  : Mud type
 MSS   .                 NA  : Mud Sample Source
 DFD   .gm/cc            9.0  : Mud Density
 DFV   .S                    : Mud Viscosity (Funnel)
 DFL   .cc                   : Fluid Loss
 PH    .                     : PH
 RM    .                 NA  : Resistivity of Mud
 MST   .                 NA  : Temperature of Mud
 RMF   .                     : Resistivity of Mud Filtrate
 MFT   .                     : Temperature of Mud Filtrate
 RMC   .                 NA  : Resistivity of Mud Cake
 MCST  .                     : Temperature of Mud Cake
 BHT   .DEG-F         116.0  : Maximum Recorded Temperature
 RMB   .OHM-M                : Resitivity @ BHT
 TIMC  .DATE             NA  : Date/Time Circulation Stopped
 TIML  .DATE                 : Date/Time Logger Tagged Bottom
 UNIT  .                401  : Logging Unit Number
 BASE  .                     : Home Base of Logging Unit
 ENG   .            RUNNELS  : Recording Engineer
 WIT   .                     : Witnessed By
```

Fonte: Extraído de (82).

#### 4.1.1.1 *Poço A*

A área de estudo localiza-se na Bacia de Campos, na região do Campo de Marlim. Três poços dessa região foram selecionados para empregar a metodologia proposta, com base na disponibilidade de perfis elétricos naquele intervalo estratigráfico. Os poços selecionados foram denominados de A, B e C, e não terão suas localizações (coordenadas) exatas ou nomes explicitados. O arquivo possui 11 informações (a profundidade e 10 características petrofísicas) para 2297 entradas analisadas. As características e as descrições encontram-se na Tabela 5.

As características petrofísicas apresentadas nas Tabelas 5, 6 e 5 são descritas a seguir. Registro de raios gama (GR) é a medida da intensidade da radioatividade natural das rochas. Os registros de raios gama são particularmente úteis porque xistos e arenitos normalmente têm assinaturas de raios gama diferentes que podem ser correlacionadas prontamente entre os poços (83).

Registro do calibrador (CALI) é uma ferramenta para medir o diâmetro e a forma de um furo de sondagem. Os registros do calibrador são geralmente medidos mecanicamente, com apenas alguns usando dispositivos sônicos. As ferramentas medem o diâmetro em uma corda específica do poço. Como os poços são geralmente irregulares, é importante ter uma ferramenta que mede o diâmetro em vários locais diferentes simultaneamente. Engenheiros de perfuração usam a medição do calibrador como uma indicação qualitativa da condição do poço e do grau em que o sistema de lama manteve a estabilidade do poço. Os dados do calibrador são integrados para determinar o volume do poço aberto, que é então usado no planejamento das operações de cimentação (83).

O Registro de Neutron compensados (CNL) é usado para medir a porosidade da rocha. A massa do átomo de hidrogênio e do nêutron é quase a mesma. Quando os neutron são bombardeados na formação, eles colidem com o hidrogênio, se na formação o número de átomos de hidrogênio for maior (água) então diminui a velocidade do neutron como resultado, pouco número de neutron será recebido no receptor e resultado oposto será obtido no caso de Hidrocarboneto (Gás). Para realizar esse procedimento utiliza-se detector longe e perto (84).

Perfuração de Sonda Sísmica de Túnel (TSPD) é um método que usa dados de vibração de perfuração chamado para estimar a distribuição da velocidade da onda elástica à frente da face do túnel (84).

Os perfis de densidade (DRHO,RHOB) são o indicador chave de porosidade na maioria dos poços, fornecendo um valor de densidade aparente da formação. Isso é obtido usando uma ferramenta de perfilagem de densidade. O dispositivo de perfilagem carrega uma fonte de raios gama e dois detectores. Formações densas absorvem muitos raios gama, enquanto formações de baixa densidade absorvem menos. Consequentemente, taxas de contagem alta nos detectores ilustram formações de baixa densidade, enquanto taxas de

contagem baixas identificam formações de alta densidade (85).

Espessura estratigráfica verdadeira (CNST) é a espessura de um corpo de rocha após correção para o mergulho do corpo e o desvio do poço que o penetra. Os valores da CNST em uma área podem ser plotados e contornos desenhados para criar um mapa *isopach* (uma linha em um mapa ou diagrama conectando pontos abaixo dos quais um determinado estrato ou grupo de estratos tem a mesma espessura) (85).

A tensão do cabo (TENS) mede a tensão no cabo que segura as ferramentas. É utilizado para detectar se a ferramenta está presa (84).

Porosidade de Neutron (NPHI) refere-se a um perfil de porosidade com base no efeito da formação em nêutrons rápidos emitidos por uma fonte. O hidrogênio tem, de longe, o maior efeito em desacelerar e capturar nêutrons. Visto que o hidrogênio é encontrado principalmente nos fluidos dos poros, o registro da porosidade do nêutron responde principalmente à porosidade. O registro é calibrado para ler a porosidade correta assumindo que os poros são preenchidos com água doce (83).

Os Registros de Densidade de Formação Compensada (FFDC, NFDC) determinam a porosidade medindo a densidade das rochas. Como esses registros superestimam a porosidade das rochas que contêm gás, eles resultam no "cruzamento"das curvas dos registros quando combinadas com os registros de neutrons (CNL) (83).

Tabela 5 – Logname e descrições - Poço A.

| Logname | Descrição |
|---|---|
| CALI | Diâmetro da perfuração |
| CNST | Espessura estratigráfica verdadeira |
| DRHO | Densidade Aparente Corrigida |
| FCNL | Registro de Neutron Compensado (detector longe) |
| FFDC | Registro de Formação Compensada (detector longe) |
| GR | Raio Gama |
| NCNL | Registro de Neutron Compensado (detector próximo) |
| NFDC | Registro de Formação Compensada (detector próximo) |
| NPHI | Porosidade de Neutron |
| RHOB | Densidade Aparente |

Fonte: Elaborada pelo autor (2020).

#### 4.1.1.2 *Poço B*

O arquivo possui 6 informações (a profundidade e 5 características petrofísicas) para 1601 entradas analisadas. As características e as descrições encontram-se na Tabela 6.

Tabela 6 – Logname e descrições - Poço B.

| Logname | Descrição |
|---|---|
| CALI | Diâmetro da perfuração |
| DRHO | Densidade Aparente Corrigida |
| GR | Raio Gama |
| NPHI | Porosidade de Neutron |
| RHOB | Densidade Aparente |

Fonte: Elaborada pelo autor (2020).

#### 4.1.1.3 *Poço C*

O arquivo possui 8 informações (a profundidade e 7 características petrofísicas) para 1476 entradas analisadas. As características e as descrições encontram-se na Tabela 7.

Tabela 7 – Logname e descrições - Poço C.

| Logname | Descrição |
|---|---|
| CNST | Espessura estratigráfica verdadeira |
| FCNL | Registro de Neutron Compensado (detector longe) |
| FFDC | Registro de Formação Compensada (detector longe) |
| NCNL | Registro de Neutron Compensado (detector próximo) |
| NFDC | Registro de Formação Compensada (detector próximo) |
| TENS | Tensão do cabo |
| TSPD | Perfuração de Sonda Sísmica de Túnel |

Fonte: Elaborada pelo autor (2020).

#### 4.1.1.4 *Daniudui e Hangjinqi*

Os dados petrofísicos usados foram obtidos de 5 poços no campo de gás Daniudui (DGF), que é localizado na parte oriental do declive Yishan da Bacia Ordos e sete poços no campo de gás Hangjinqi (HGF), que é localizado no norte da Bacia Ordos (Figura 13).

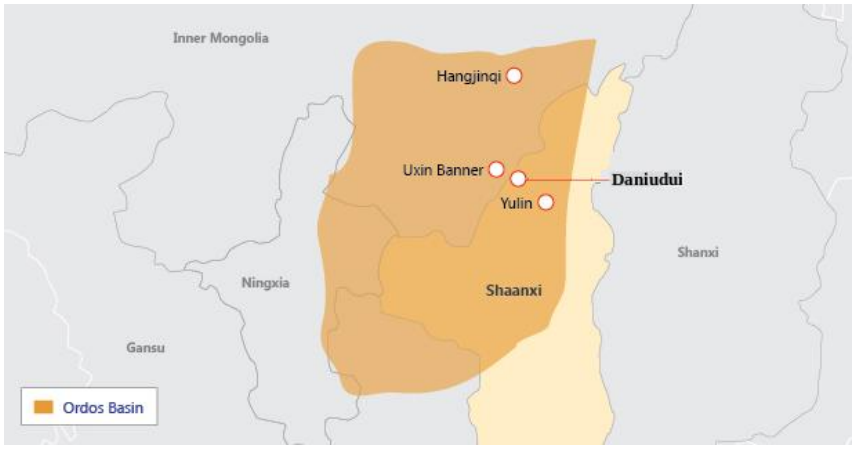
A base de dados apresenta informações de 12 poços e 2153 lâminas. Possui dados referentes a 7 propriedades (Tabela 8) (2).

Tabela 8 – Registros analisados - Base de Dados DGF e HGF.

| | |
|---|---|
| Raio Gama (API) | Latero profundo (Ωm) |
| Acústico ($\mu$s/m) | Latero raso (Ωm) |
| Densidade (g/cm$^3$) | Caliper |
| Nêutron compensado (%) | |

Fonte: Elaborada pelo autor (2020).

Figura 13 - Localização dos dos campos de gás Daniudui e Hangjinqi.




Fonte: Adaptado de (86).

As rochas clásticas são divididas segundo o tamanho dos grãos detríticos em arenito pedregoso (>1 mm), arenito grosso (0.5 - 1 mm), arenito médio (0.25 - 0.5 mm), arenito fino (0.01 - 0.25 mm), siltito (0.005 - 0.05) e lamito (<0.005 mm), (87). Oito classes litológicas foram identificadas: rocha carbonática (CR), carvão (C), arenito pedregoso (PS), arenito grosso (CS), arenito médio (MS), arenito fino (FS), siltito (S) e lamito (M).

A Tabela 9 apresenta a distribuição das amostras na base de dados. Nota-se que ocorre um desbalanceamento nos dados, uma vez que há classes com muitas amostras e outras com poucas.

Tabela 9 – Classes DGF e HGF x $N^o$ de amostras

| Base de Dados | Classes | $N^o$ de amostras | Percentual |
|---|---|---|---|
| DGF | C | 104 | 11,3% |
| | CR | 48 | 5,3% |
| | CS | 114 | 12,5% |
| | FS | 132 | 14,5% |
| | M | 133 | 14,5% |
| | MS | 211 | 23,1% |
| | PS | 120 | 13,1% |
| | S | 53 | 5,7% |
| HGF | C | 14 | 1,2% |
| | CS | 207 | 16,8% |
| | FS | 146 | 11,7% |
| | M | 248 | 20,0% |
| | MS | 206 | 16,6% |
| | PS | 370 | 29,9% |
| | S | 47 | 3,8% |

Fonte: Elaborada pelo autor (2020).

### 4.1.2 **Dados Petrográficos**

A Petrografia Sedimentar investiga as amostras de rochas extraídas de testemunhos de poços ou áreas de exploração de petróleo, com o propósito de definir, através da análise de um fragmento de rocha, as características de um reservatório de petróleo. Esse conhecimento estabelece a capacidade econômica do reservatório e os métodos de explotação empregáveis. A análise das amostras é feita a partir da descrição minuciosa das feições da rocha observadas por meio de um microscópio ótico de luz polarizada ou amostras de mão. A interpretação das feições é realizada utilizando tipicamente raciocínio de imagens. A solução dessa tarefa depende do conhecimento de um especialista (88).

A caracterização petrográfica é uma das bases para a avaliação e previsão da qualidade dos reservatórios, tanto na investigação por novos reservatórios de petróleo, quanto no desenvolvimento e produção de campos já conhecidos. A petrografia de rochas-reservatório é uma tarefa especializada e intensiva cuja quantidades extensas de dados são obtidas, do tipo qualitativo e quantitativo. Entretanto, apesar da relevância das informações petrográficas para a exploração de petróleo, o conhecimento agregado geralmente é pouco gerenciado, e frequentemente dados e informações de enorme valor são perdidos (89).

Três bases de dados petrográficos foram utilizadas: Tibagi, Paleosul e La Ciotat-1. A escolha dessas bases se deu pela disponibilidade na literatura e o interesse em investigar a abrangência da metodologia propostas. As bases estão disponibilizadas no Apêndice A.

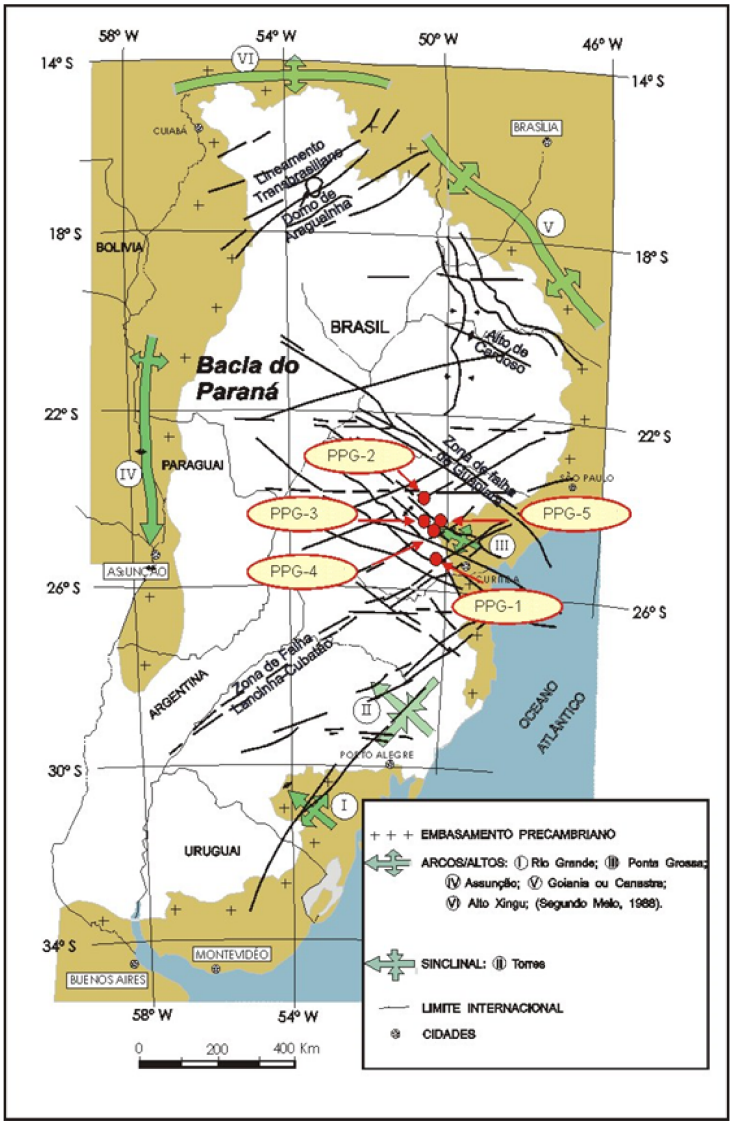
#### 4.1.2.1 *Tibagi*

Os dados analisados são lâminas petrográficas obtidas a partir de amostras coletadas em furos de sondagem pertencentes ao Membro Tibagi, Devoniano da Bacia do Paraná (Projeto Paleosul). A Figura 14 apresenta a localização dos furos de sondagem.

A base de dados apresenta informações de 5 poços e 44 lâminas ao todo. A análise quantitativa das lâminas foi feita através da contagem (em microscópio petrográfico de luz transmitida) de 300 pontos em cada lâmina petrográfica com espaçamento de 0.3 mm. Detalhes do processo de obtenção dos dados podem ser encontrados em (90). Nas amostras provenientes destes poços, foram contabilizadas, para cada lâmina, as porcentagens de 22 constituintes (Tabela 10).

A Tabela 11 apresenta as petrofácies e os respectivos número de amostras. A partir da observação dessa tabela, nota-se um desbalanceamento nos dados, a petrofácies PT-1 possui a maioria das amostras (53.49%).

Figura 14 - Localização dos Furos de Sondagem PPG-1, PPG-2, PPG-3, PPG-4 e PPG-5




Fonte: Extraído de (21).

#### 4.1.2.2 *Paleosul*

Os dados analisados são lâminas petrográficas obtidas a partir de amostras coletadas em 3 furos de sondagem pertencentes ao Devoniano da Bacia do Paraná (Projeto Paleosul) da Formação Ponta Grossa. A Figura 15 apresenta a localização dos furos de sondagem.

A base de dados apresenta informações de 3 poços e 60 lâminas ao todo. Nas amostras provenientes destes poços, foram contabilizadas, para cada lâmina, as porcentagens de 25 constituintes (Tabela 12). Detalhes do processo de obtenção dos dados podem ser encontrados em (41).

A Tabela 13 apresenta a distribuição das amostras na base de dados. Nota-se que ocorre um desbalanceamento nos dados, uma vez que há petrofácies com muitas amostras e outras com poucas.

Tabela 10 – Petrográficos analisados - Base de Dados Tibagi

| | |
|---|---|
| Bioclasto | Quartzo |
| Crescimento Secundário de Quartzo | Feldspato |
| Caolinita | Muscovita |
| Ilita/Smectita | Opaco |
| Pirita | Turmalina |
| Siderita | Zircão |
| Cimento Carbonático | Rutilo |
| Cimento Silicoso | Glauconita |
| Cimento Ferruginoso | Clorita |
| Porosidade Intergranular | Pseudo Matriz |
| Porosidade Intragranular | Litoclasto |

Fonte: Elaborada pelo autor (2020).

Tabela 11 – Petrofácies Tibagi x Nº de amostras

| Petrofácies | $N^o$ de amostras | Percentual |
|---|---|---|
| PT-2 | 1 | 2.32% |
| PT-3 | 2 | 4.65% |
| I-2 | 5 | 11.63% |
| PT-4 | 5 | 11.63% |
| I-1 | 7 | 16.28% |
| PT-1 | 23 | 53.49% |

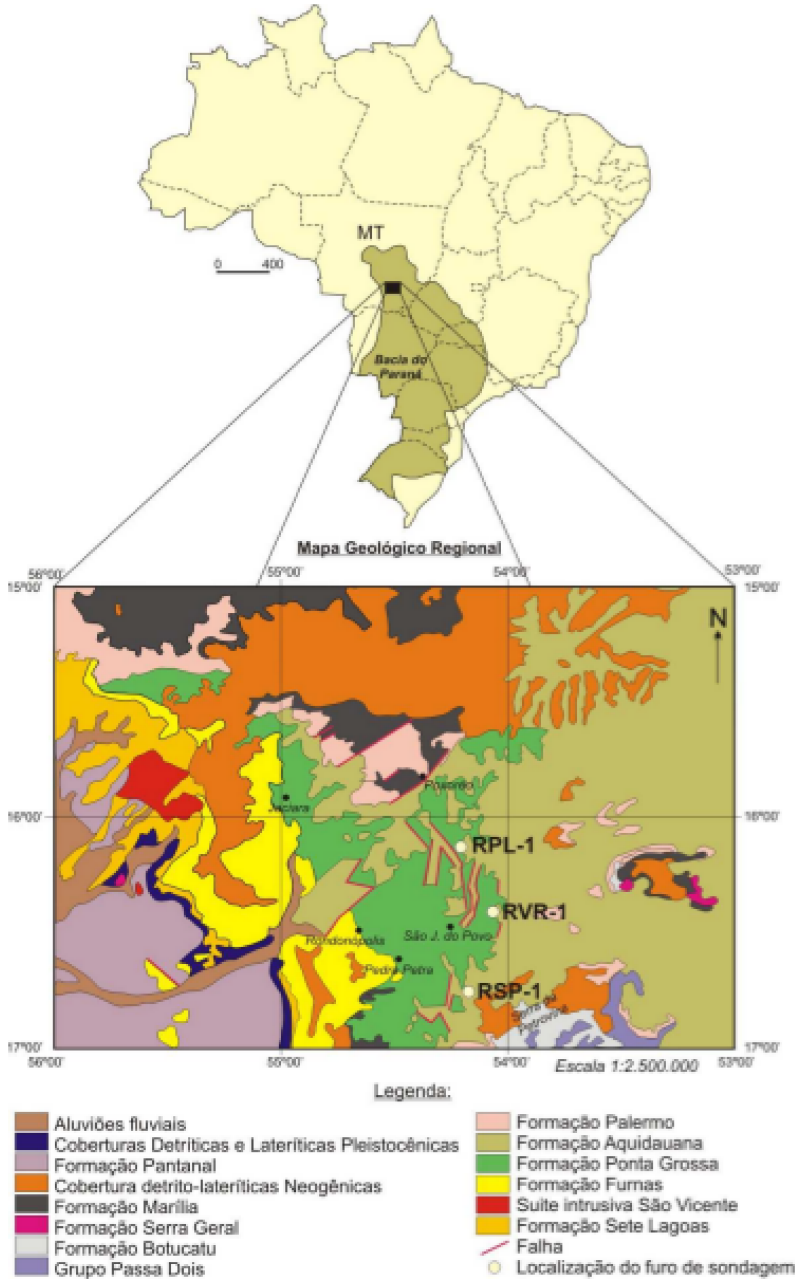
Fonte: Elaborada pelo autor (2020).

Tabela 12 – Petrográficos analisados - Base de Dados Paleosul

| | |
|---|---|
| Quartzo Monocristalino | Nódulos de Siderita |
| Quartzo Policristalino | Siderita |
| K-Feldspato Detrítico | Pirita |
| Plagioclásio Detrítico | Caolinita |
| Mica | Super Crescimento Quartzo |
| Minerais Pesados | Albitização do Feldspato |
| Por bioturbação | Carbonato |
| Glauconita | Outros (Ti-óxidos, super cres. F) |
| Bioclasto | Porosidade Intergranular |
| Ooide Goetita | Porosidade Intragranular |
| Ooide Bertierina | Oversized |
| Substituição de grão Bertierina (F, M) | Moldic |
| Argilomineral não-identificado | |

Fonte: Elaborada pelo autor (2020).

Figura 15 - Localização dos Furos de Sondagem RSP-1, RVR-1 e RPL-1




Fonte: Extraído de (41).

Tabela 13 – Petrofácies Paleosul x Nº de amostras

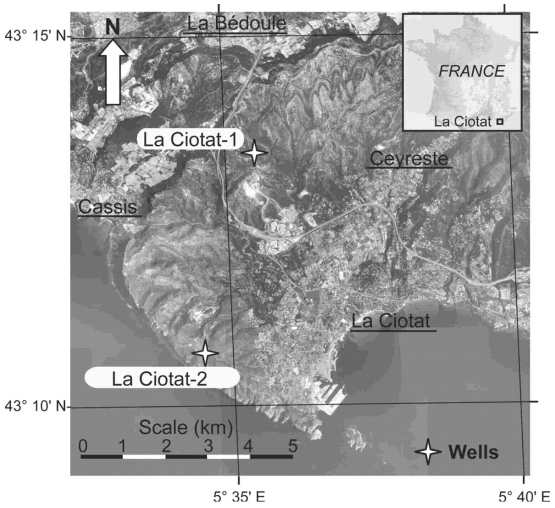
| Petrofácies | Número de amostras | Percentual |
|---|---|---|
| PG-3 | 1 | 1.67% |
| PG-2 | 3 | 5.00% |
| PG-1 | 56 | 93.33% |

Fonte: Elaborada pelo autor (2020).

#### 4.1.2.3 *La Ciotat-1*

O poço estudado é localizado na Bacia Provença do Sul nas proximidades de Cassis e La Ciotat, chamado La Ciotat-1 (Figura 16).

Figura 16 - Localização dos Furos de Sondagem La Ciotat-1 e La Ciotat-2




Fonte: Extraído de (3).

A base de dados apresenta informações de 1 poço e 40 lâminas. Possui dados referentes a 19 propriedades (Tabela 14). Os dados foram separados em sete classes petrográficas de acordo com (3) com mostra a Tabela 15. Detalhes do processo de obtenção dos dados podem ser encontrados em (3).

Tabela 14 – Petrográficos analisados - Base de Dados La Ciotat-1

| | |
|---|---|
| Porosidade (%) | Fração de carbonato (%) |
| Densidade a granel (g/cm3) | Calcita (%) |
| VP (m/s) | Dolomita (%) |
| VS1 (m/s) | Quartzo (%) |
| VS2 (m/s) | Ortoclásio (%) |
| K1 (GPa) | Albita (%) |
| K2 (GPa) | Argilas (%) |
| m1 (GPa) | Pirita (%) |
| m2 (GPa) | $\alpha$ |
| Tamanho do Grão | |

Fonte: Elaborada pelo autor (2020).

A Tabela 16 apresenta a distribuição das amostras na base de dados. Nota-se que ocorre um desbalanceamento nos dados, uma vez que há classes com muitas amostras e outras com poucas. Observas-se que C5 possui apenas uma amostra.

Tabela 15 – Classes petrográficas e suas descrições segundo (3).

| Classes | Descrições |
|---|---|
| C1 | Calcário com textura grainstone (quartzo $< 5\%$) |
| C2 | Calcário com textura wackestone-packstone (quartzo $< 5\%$) |
| C3 | Calcário rico em quartzo com espaço intergranular esparítico/microesparítico: textura grainstone ou wackestone-packstone com matriz recristalizada (quartzo $5\% - 50\%$) |
| C4 | Calcário rico em quartzo com espaço intergranular micrítico: textura wackestone-packstone (quartzo $5\% - 50\%$) |
| C5 | Pedra calcária argilosa, rica em quartzo, com textura wackestone-packstone (quartzo $5\% - 50\%$ e argila $2\% - 5\%$) |
| C6 | Arenito cimentado limpo (quartzo $> 50\%$) |
| C7 | Arenito com matriz micrítica carbonática (quartzo $> 50\%$) |

Fonte: Elaborada pelo autor (2020).

Tabela 16 – Classes La Ciotat-1 x Nº de amostras

| Classes | Número de amostras | Percentual |
|---|---|---|
| C1 | 5 | 12.50% |
| C2 | 8 | 20.00% |
| C3 | 6 | 15.00% |
| C4 | 9 | 22.50% |
| C5 | 1 | 2.50% |
| C6 | 4 | 10.00% |
| C7 | 7 | 17.50% |

Fonte: Elaborada pelo autor (2020).

## 4.2 METODOLOGIA PROPOSTA

A metodologia empregada se divide em duas partes: supervisionada e não supervisionada. A supervisionada engloba métodos de classificação, Evolução Diferencial, validação cruzada, métricas para avaliação dos classificadores. A parte não supervisionada envolve Bootstrap, Características Polinomiais, K-Means, Análise de Silhueta, Análise de Componentes Principais (ACP) e Análise Filogenética. Os métodos utilizados estão nos pacotes *S*cikit-Learn e *S*cikit-Bio em Python (91). A metodologia proposta difere das apresentadas no Capítulo 3 por além de ter as análises comumente utilizadas na literatura, apresenta uma alternativa para auxiliar na compreensão do processo diagenético por meio de Análise Filogenética.

Para as bases de dados petrofísicas (Poço A, Poço B e Poço C), que não possuem classificação prévia, empregou-se a abordagem não supervisionada, uma vez que não há conhecimento prévio de alguma classificação que já tenha sido realizada. Duas estratégias foram aplicadas:

1. K-Means + ACP.
2. Características Polinomiais + K-Means.

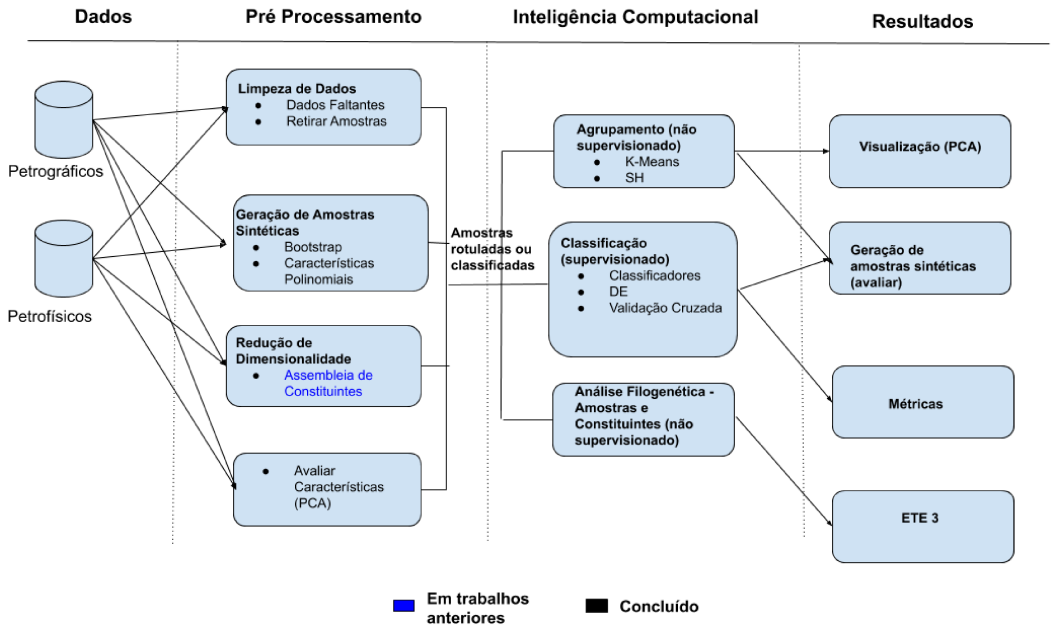
Para as bases de dados petrográficas Tibagi e Paleosul utilizou-se a abordagem não supervisionada com a inclusão do Bootstrap, pois para utilizar esta técnica há a necessidade de se conhecer as classes.

1. K-Means + ACP.
2. Características Polinomiais + K-Means.
3. Bootstrap + K-Means.
4. Análise Filogenética.

Para as bases de dados La Ciotat-1 e Daniudui e Hangjinqi foi aplicado a abordagem supervisionada. Fez-se uso de uma técnica de validação cruzada para dividir os dados em treinamento e teste, posteriormente empregou-se o método de Evolução Diferencial (DE) no classificador com os dados de treinamento para encontrar os parâmetros ótimos. Após ter encontrado o melhor modelo, o mesmo foi executado no conjunto de teste. Seis classificadores foram utilizados Redes Neurais Artificiais (ANN), Árvore de Decisão (DT), Máquina de Aprendizado Extremo (ELM), Gradient Boosting (GB), K-Vizinhos Mais Próximos (KNN) e Máquina de Vetor Suporte (SVM).

O esquema apresentado na Figura 17 ilustra a metodologia proposta de forma geral.

Figura 17 - Esquema ilustrando a metodologia proposta.




Fonte: Elaborada pelo autor (2020).

## 4.3 MÉTODOS COMPUTACIONAIS

O uso de técnicas de Inteligência Computacional para resolver problemas envolve algumas etapas de acordo com o objetivo: classificar (supervisionada) ou agrupar (não supervisionada). Se o problema for de classificação há divisão da base de dados em treinamento e teste, pode-se utilizar métodos para encontrar os parâmetros ótimos dos classificadores como *Grid-Search* e Evolução Diferencial, e métricas para a seleção de modelos como a acurácia. No caso de agrupamento pode-se utilizar métodos para encontrar os melhores parâmetros, métricas para a seleção de modelos como o coeficiente de silhueta e Análise de Componentes Principais para visualizar os agrupamentos encontrados. Estas etapas serão descritas nas próximas seções.

### 4.3.1 **Bootstrap**

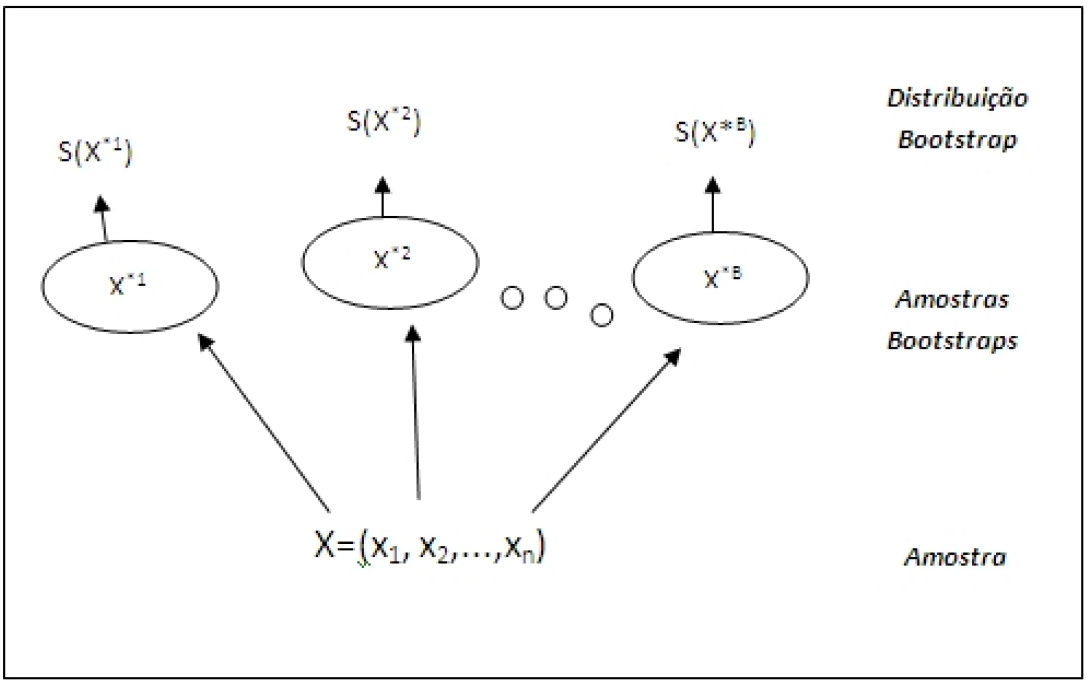
O Bootstrap (92) é uma técnica estatística não paramétrica de reamostragem, que tem como objetivo conseguir informações de características da distribuição de alguma

variável aleatória. Para tal, aproxima-se uma distribuição de probabilidade a partir de uma função empírica obtida de uma amostra finita.

O método baseia-se em realizar amostragens de tamanho igual ao da amostra inicial com reposição da mesma, isto é, n sorteios são realizados, sendo n o número de observações disponíveis na amostra, com reposição da amostra original, o que origina uma amostra bootstrap. Este procedimento deve ser repetido B vezes para que sejam geradas B amostras bootstrap. Posteriormente, é possível construir uma distribuição bootstrap da variável de interesse. Essa distribuição é utilizada para realizar inferências e tirar informações sobre o parâmetro em estudo.

Na Figura 18, $B$ amostras bootstrap da amostra original foram geradas. Cada amostra bootstrap possui tamanho $n$ e é obtida através de $n$ sorteios com reposição da amostra original. A distribuição bootstrap é obtida ao se calcular o desvio padrão $S(x^{*i})$ de cada uma das $B$ amostras bootstraps obtidas. Após o cálculo dos desvios padrões para cada amostra bootstrap $(S(x^{*1}), S(x^{*2}), \cdots, S(x^{*B}))$ pode-se obter a distribuição boostrap do parâmetro de interesse. Esse foi um dos primeiros exemplos da aplicação do Bootstrap (93).

Figura 18 - Esquema do Processo Bootstrap




Fonte: Retirado de (94).

O Bootstrap será aplicado como uma forma de determinar os centroides que serão parâmetro do algoritmo K-Means para a análise de agrupamento nos dados originais (95, 96, 97).

### 4.3.2 **Características Polinomiais**

Os recursos de entrada para uma tarefa de modelagem estatística, geralmente, interagem de maneiras inesperadas e frequentemente não lineares.

Essas interações podem ser identificadas e modeladas por um algoritmo de Inteligência Computacional. Outra abordagem é projetar novos recursos que expõem essas interações e ver se elas melhoram o desempenho do modelo. Além disso, transformações como elevar variáveis de entrada a uma potência podem ajudar a expor melhor as relações importantes entre variáveis de entrada e a variável de destino.

Esses recursos são chamados de recursos de interação e polinomiais e permitem o uso de algoritmos de modelagem mais simples, já que parte da complexidade de interpretar as variáveis de entrada e seus relacionamentos é devolvida ao estágio de preparação de dados. Às vezes, esses recursos podem resultar em melhor desempenho de modelagem, embora ao custo de adicionar milhares ou até milhões de variáveis de entrada adicionais (98).

Recursos polinomiais são aqueles criados ao elevar os recursos existentes a um expoente. Por exemplo, se um conjunto de dados tinha um recurso de entrada X, então um recurso polinomial seria a adição de um novo recurso (coluna) onde os valores foram calculados ao elevar ao quadrado os valores em $X$, por exemplo, $X^2$. Se uma amostra de entrada é bidimensional e da forma $[a, b]$, as características polinomiais de grau 2 são $[1, a, b, a^2, ab, b^2]$. Esse processo pode ser repetido para cada variável de entrada no conjunto de dados, criando uma versão transformada de cada uma.

O grau do polinômio é usado para controlar o número de recursos adicionados, por exemplo, um grau de 3 adicionará duas novas variáveis para cada variável de entrada. Normalmente, um pequeno grau é usado, como 3 ou 4.

De modo geral, é incomum usar de maior que 3 ou 4 porque para grandes valores a curva polinomial pode se tornar excessivamente flexível e assumir algumas formas muito estranhas (98).

Também é comum adicionar novas variáveis que representam a interação entre os recursos, por exemplo, uma nova coluna que representa uma variável multiplicada por outra. Isso também pode ser repetido para cada variável de entrada, criando uma nova variável de "interação"para cada par de variáveis de entrada.

Uma versão ao cubo ou a quarta potência de uma variável de entrada mudará a distribuição de probabilidade, separando os valores pequenos e grandes, uma separação que é aumentada com o tamanho do expoente.

Essa separação pode ajudar alguns algoritmos de Inteligência Computacional a a fazer melhores previsões ou encontrar melhores agrupamentos e é comum para tarefas de modelagem preditiva de regressão e geralmente tarefas que têm variáveis de entrada

numéricas.

### 4.3.3 **Análise de Agrupamento**

Uma questão básica que muitos pesquisadores de várias áreas enfrentam é como organizar dados observados em estruturas que agrupem subconjuntos semelhantes, isto é, como criar ou desenvolver taxionomias. O conceito de Análise de Agrupamento surgiu como uma forma de automatizar esse processo (99).

As técnicas de análise multivariada tornam possível realizar a avaliação de um conjunto de atributos, considerando as similaridades existentes. A Análise de Agrupamento é uma técnica multivariada que tem como intuito encontrar uma ou várias partições na base de dados, ou seja, grupos, segundo algum critério de classificação, de tal maneira que exista homogeneidade intra grupo e heterogeneidade entre os grupos (99).

Na literatura existem diversas medidas de similaridade/dissimilaridade propostas que têm sido bastante utilizadas em Análise de Agrupamento. A escolha entre essas medidas depende da preferência do usuário em relação a aproximação adotada pela medida.

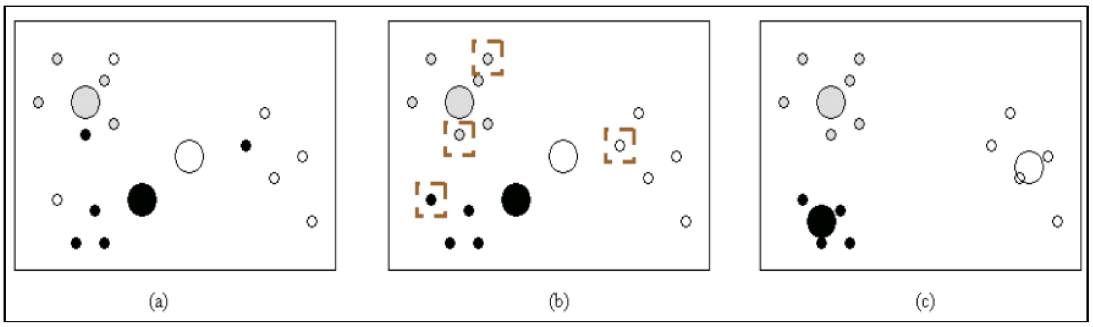
Para adotar um método de agrupamento, deve-se decidir qual é o mais adequado, entre os vários existentes, de acordo com intuito do trabalho. O uso de diferentes técnicas podem resultar em distintas soluções, daí a importância na escolha do método de agrupamento. A Análise de Agrupamento, em geral, compreende seis etapas:

1. A seleção de indivíduos que serão agrupados;
2. A definição de um conjunto de variáveis a partir das quais serão obtidas informações necessárias ao agrupamento dos indivíduos;
3. A definição de uma medida de semelhança ou distância entre os indivíduos;
4. A escolha de um algoritmo de agrupamento;
5. Escolha dos parâmetros do algoritmo;
6. Validação dos resultados.

Neste trabalho utilizou-se o método K-Means (100), um dos algoritmos mais utilizados para realizar agrupamentos. A partir da escolha de centroides iniciais, de forma aleatória, cada amostra é atribuída ao centroide mais próximo. O próximo passo é atualizar os centroides tomando o valor médio de todas as amostras designadas para cada centroide anterior. Calcula-se a diferença entre os antigos e os novos centroides, repetindo o processo a partir da atribuição de amostras aos centroides, até que este valor seja inferior a um limiar. O número de agrupamentos a ser gerados é passado como parâmetro. A medida

de dissimilaridade utilizada foi a distância Euclidiana, pois é comumente utilizada na literatura. A Figura 19 ilustra o procedimento de agrupamento para K= 3.

Figura 19 - Exemplo de execução do K-Means. (a) Cada elemento foi distribuído para um dos três agrupamentos, de maneira aleatória, e os centroides foram calculados para cada grupo (representados pelos círculos maiores). (b) Os elementos foram destinados para os grupos que possuem centroides mais próximos (c) Os centroides foram recalculados. Os grupos já estão em sua forma final. Caso não estivessem, seria repetido os passos (b) e (c) até que estivessem




Fonte: Extraído de (101).

Para avaliar a qualidade de um agrupamento, critérios de validação são adotados. A escolha do critério depende do problema a ser tratado. Existem alguns estudos tratando desse assunto e comparando critérios, como em (102). Foi utilizado o coeficiente de silhueta (Silhuette Coefficient - SC) que será descrito a seguir.

#### 4.3.3.1 *Silhueta*

A análise de silhueta (103) é um método geométrico baseado na compactação e separação de agrupamentos com o intuito de analisar a qualidade dos agrupamentos formados. O número ótimo de agrupamentos é definido pelo maior coeficiente de silhueta resultante da análise de silhueta. Para cada amostra $i$ o valor $s_i$ é definido pela seguinte fórmula:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{4.1}$$

Considerando que a amostra $i$ pertença ao agrupamento $A$, $a_i$ é descrito como a dissimilaridade média da amostra $i$ em relação a todas as outras amostras do agrupamento $A$. Seja $B$ um agrupamento diferente de $A$, $b_i$ é a dissimilaridade média da amostra $i$ em relação a todas as amostras de $B$. O coeficiente de silhueta de um conjunto de dados é dado pela média dos coeficientes individuais das amostras

$$SC = \frac{\sum_{i=1}^{N} s_i}{N} \tag{4.2}$$

onde $N$ é o número de amostras do conjunto de dados. A métrica utilizada para o cálculo da dissimilaridade é a distância Euclideana. O valor de $SC$ varia de -1 a 1. Resultado próximo de 1, indica que os objetos estão melhor agrupados.

#### 4.3.3.2 *Kruskal Wallis*

Um teste de análise de variância ou uma *ANalysis Of VAriance* (ANOVA) é uma generalização dos testes $t$ para mais de dois grupos. A hipótese nula ($H_0$) afirma que existem médias iguais nas populações das quais os grupos de dados foram amostrados. O nível de significância ($p$-valor $< 0.05$) é o nível de risco máximo aceitável para rejeitar a hipótese nula quando ela é verdadeira. Quando o $p$-valor é menor do que o nível de significância de 5%, a interpretação é que os resultados são estatisticamente significativos, e $H_0$ é rejeitada, prevalecendo $H_1$. Um resultado são estatisticamente significativo quando a estatística da amostra é atípica o suficiente em relação à hipótese nula para que podemos rejeitar a hipótese nula para toda a população. O teste ANOVA é um teste paramétrico, então segue os seguintes pressupostos: distribuição normal, homogeneidade dos dados e variáveis intervalares e contínuas. Aqui será utilizado o teste de *Kruskal-Wallis* que é a versão não paramétrica do ANOVA que não coloca nenhuma restrição sobre a comparação.

O teste de *Kruskal-Wallis* (104) é aplicado quando estão em comparação três ou mais grupos independentes e a variável deve ser de mensuração ordinal. Procedimentos para a realização do teste:
a) Dispor, em postos, as observações de todos os grupos em uma única série, atribuindo de 1 a N.
b) Determinar o valor de R (soma dos postos) para cada um dos grupos de postos.
c) Determinar $H_{cal}$ (valor real do teste) através de:

$$H_{cal} = \frac{\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)}{1 - \frac{\sum T}{N^3 - N}} \tag{4.3}$$

onde a soma é sobre todos os grupos e $T = (t-1)t(t+1) = t^3 - t$ para cada grupo de empates, sendo t o número de observações empatadas no grupo.
d) O valor teórico $H_{tab}$ é obtido através de uma tabela da distribuição de Qui-quadrado.
e) Por último, comparar o valor real $H_{cal}$ com o valor teórico de $H_{tab}$, segundo o nível de significância de 5%. Se H calculado for menor que $H_{tab}$ tabelado não se pode rejeitar a hipótese nula.

A $H_0$ é que as médias das características intergrupos são iguais e $H_1$ considera que as médias são significativamente diferentes.

- $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_n$
- $H_1$: $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_n$

### 4.3.4 **Classificação**

A tarefa de classificação consiste em aprender com dados previamente rotulados para auxiliar no planejamento e tomada de decisões. A classificação está especificamente voltada à atribuição de uma das classes pré-definidas pelo analista a novos fatos ou objetos submetidos à classificação.

O procedimento de classificação de dados é dividido em dois passos. No primeiro, definido como treinamento, ocorre a criação de um modelo que descreve um conjunto predeterminado de classes de dados. Essa criação é realizada através da análise das amostras de uma base de dados, na qual as amostras são descritas por características e cada uma delas pertence a uma classe definida anteriormente, identificada por uma das características. O conjunto de amostras usadas neste passo é o conjunto de treinamento.

Geralmente representa-se os padrões aprendidos no primeiro passo por regras de classificação, árvores de decisão ou formulações matemáticas. Este padrão pode ser aplicado para prever as classes de futuras amostras desconhecidas, além de possibilitar um maior entendimento sobre a base de dados.

No segundo passo, testa-se o modelo criado, ou seja, utiliza-se o modelo para classificação de um novo conjunto de amostras, separadas das utilizadas no treinamento, chamado conjunto de teste. Este conjunto, do mesmo modo, possui as classes conhecidas, então depois da classificação, usualmente calcula-se o percentual de acertos, comparando as classes preditas pelo modelo com as classes conhecidas (105).

#### 4.3.4.1 *Redes Neurais Artificiais*

Uma Rede Neural Artificial (*Artificial Neural Network*, ANN) (106) é uma ferramenta matemática, que apresenta um bom desempenho, que pode ser utilizada para uma série de aplicações, como predição e agrupamento. Dentre as diferentes topologias neurais considera-se que as redes perceptron multicamadas são àquelas implementadas com maior frequência devido à facilidade de implementação e sintaxe menos robusta. As redes neurais, em geral, são modelos estatísticos flexíveis utilizados para a modelagem de problemas não-lineares de alta complexidade.

A escolha da arquitetura apropriada da ANN e seus parâmetros podem levar a ganhos significativos na solução do problema. No entanto, é difícil criar uma arquitetura projetada à mão que melhore o desempenho da melhor maneira possível. A construção automática de arquiteturas de redes neurais tem sido do interesse de vários pesquisadores, usando diferentes tipos de técnicas como Aprendizagem por Reforço, Otimização Bayesiana e Algoritmos Evolutivos. Neste trabalho, o número de camadas e o número de neurônios em cada uma delas é encontrado através do algoritmo de Evolução Diferencial (107).

A codificação ANN é descrita nas Tabelas 39 e 40. A Figura 20 exibe alguns

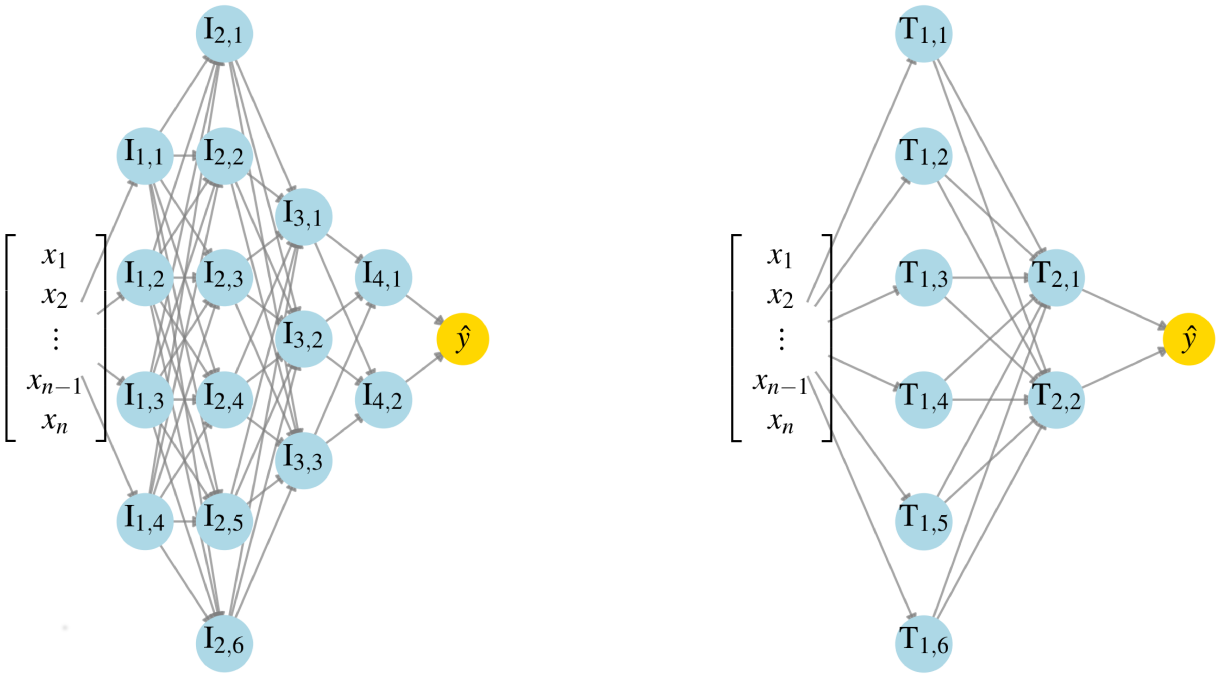
exemplos de soluções candidatas usando essa codificação. Usando essa codificação, é possível realizar uma busca no espaço de parâmetros, incluindo a função de ativação (Tabela 17), algoritmo de treinamento, coeficiente de regularização da função de perda usada para treinar a rede, o número de camadas e o número de neurônios por camada.

Tabela 17 – Funções de ativação de saída usadas na ANN.

| ID | Nome | Função de Ativação ($\varphi$) |
|---|---|---|
| 0 | Identity | $\mathrm{I}(x) = x$ |
| 1 | Logistic | $\mathrm{L}(x) = 1/(1 + e^{-x})$ |
| 2 | Tanh | $\mathrm{T}(x) = \tanh(x)$ |
| 3 | ReLU | $\mathrm{R}(x) = \max(0, x_i; i = 1, \cdots, n)$ |

Fonte: Elaborada pelo autor (2020).

Figura 20 - Exemplos de espaço de busca de ANN: soluções candidatas implementando o algoritmo de treinamento *Stochastic Gradient Descent* (SGD) e usando um coeficiente de regularização igual a 0.05, conforme descrito na Tabela 39. Esquerda: solução candidata $\boldsymbol{\theta}$ = [0,1,0.05,4,4,6,3,2, -], que representa uma rede neural com função de ativação de identidade e 4 camadas ocultas com 4, 6, 3 e 2 neurônios, respectivamente. Direita: solução candidata $\boldsymbol{\theta}$ = [2, 1,0,05, 2,6,2, -, -, -], função de ativação da tangente hiperbólica, seis neurônios no primeiro camada oculta e dois neurônios no segundo



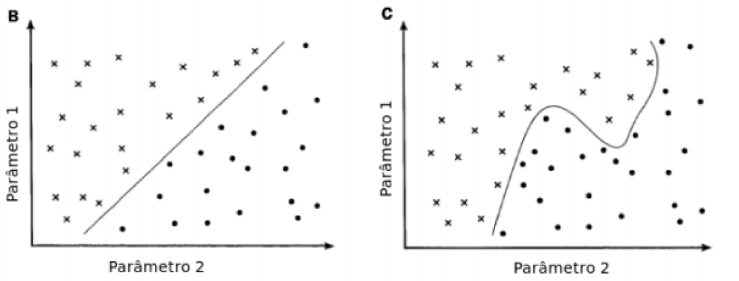

Fonte: Elaborada pelo autor (2020).

#### 4.3.4.2 *Máquinas de Vetor Suporte*

As Máquinas de Vetor Suporte (*Support Vector Machines* - SVM) (108) é um método de aprendizagem supervisionado usado para estimar uma função que classifique dados de entrada em duas classes (normalmente, mas é multiclasses). O objetivo do treinamento através de SVM é a obtenção de hiperplanos que dividam as amostras de tal forma que sejam otimizados os limites de generalização. Os resultados o uso desta técnica são comparáveis aos gerados por outros métodos de aprendizado, como as Redes Neurais Artificiais (ANN) (109), e em algumas aplicações têm se mostrado superiores, tal como em Bioinformática (110) e em Geologia (111).

Em problemas reais os padrões encontrados na maioria dos casos são complexos e não-lineares. Para expandir a SVM linear a resolução de problemas não-lineares foram inseridas funções reais, que mapeiam o conjunto de treinamento em um espaço linearmente separável, o espaço de características. Um conjunto de dados é definido como não-linearmente separável, quando não é possível separar os dados com um hiperplano. A Figura 21 mostra um conjunto linearmente e outro não-linearmente separável.

Figura 21 - Problema não linearmente separável e um linearmente separável.




Fonte: Extraído de (112).

A SVM não-linear realiza uma transformação de dimensionalidade, através das funções Kernel, para tornar um problema de classificação linear, e dessa forma fazer uso do hiperplano ótimo.

Seja o conjunto de entrada $S$ representado pelos pares $(x_1, y_1), ..., (x_n, y_n)$, com $y_i$, $i = 1, 2, .., n$ o rótulo de cada padrão $i$, o conjunto de dados de treinamento. O espaço de característica é um espaço de dimensionalidade mais alta no qual serão mapeados o conjunto de entrada $S$, por meio de uma função $\phi$, com intuito de obter um novo conjunto de dados $S'$ linearmente separável, representado por $(\phi(x_1), y_1), ..., (\phi(x_n), y_n)$.

Com os dados de treinamento mapeados para o espaço de características, o SVM é comumente formulado como um problema de otimização da seguinte forma:

$$Maximizar \sum_{i=1}^{n} \alpha_i \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(z_i).\phi(z_j) \rangle \tag{4.4}$$

$$\text{sujeito a } \begin{cases} \alpha_i > 0, i = 1, .., n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

onde $\alpha_i$ são os multiplicadores de Lagrange e $n$ é o número de amostras.

O hiperplano de decisão ótimo é determinado como

$$(w\phi(x)) + b = 0 \tag{4.5}$$

Os valores ótimos de $(w, b)$, sendo $w$ e $b$, o vetor peso e o bias respectivamente, que será representado por $(w*, b*)$, pode ser calculado como segue:

$$w* = \sum_{i=1}^{n} \alpha_i * y_i \phi(x_i) \tag{4.6}$$

o valor de $b*$ pode ser estimado empregando as equações de Karush-Kuhn-Tucker (KKT)

$$\alpha_i* = (y_i \langle w * \phi(x_i) \langle + b*) - 1) = 0, i = 1, ..., n \tag{4.7}$$

Dado um vetor $\phi(x_j)$, pode-se obter $b*$ através da condição de KKT

$$b* = y_j - \langle w * \phi(x_j) \langle . \tag{4.8}$$

O problema de classificação não-linear de um novo padrão $z$ é resolvido calculando

$$sgn(\langle w * \phi(z) \langle + b*) \tag{4.9}$$

O produto interno $\phi(z_i)\phi(z_i)$ (Eq. 4.4 é substituído pelo kernel $K(z_i, z_j)$ que tem algumas propriedades especiais. Existem diferentes tipos de kernel, os mais utilizados são apresentados na Tabela 18.

Tabela 18 – Tipos de Kernel

| Tipo de Kernel | $K(z_i, z_j)$ |
|---|---|
| Polinomial | $(\langle z_i . z_j \rangle + 1)^p$ |
| RBF | $e^{(-\gamma \|z_i - z_j\|^2)}$ |
| Sigmoidal | $\tanh(\beta_0 \langle z_i . z_j \rangle) + \beta_1$ |

Fonte: Elaborada pelo autor (2020).

#### 4.3.4.3 *K-Nearest Neighbors*

K-Vizinhos mais próximos (K-Nearest Neighbors, KNN) é um classificador no qual o aprendizado é baseado na aproximação. O conjunto de treinamento é constituído por vetores $n$-dimensionais e cada amostra deste conjunto representa um ponto no espaço $n$-dimensional.

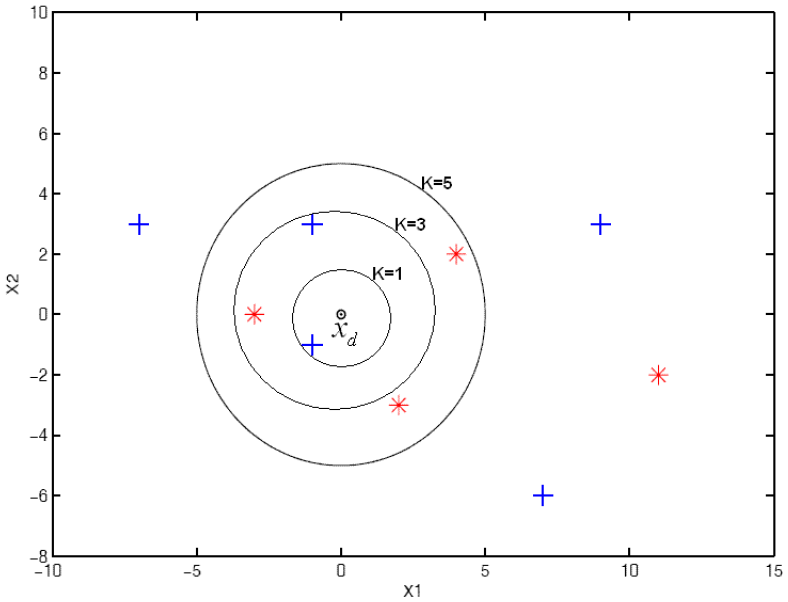
A classe de um elemento desconhecido ao conjunto de treinamento é determinada da seguinte forma: o classificador busca K amostras do conjunto de treinamento que estejam mais próximos da amostra desconhecida, isto é, que tenham as menores distâncias.

Estas K amostras são denominadas de K-vizinhos mais próximos. Examina-se quais são as classes desses K vizinhos e a classe que mais se repete será atribuída à amostra desconhecida. A métrica utilizada aqui para o cálculo da distância entre dois pontos é a distância Euclidiana. Seja $X = (x_1, x_2, ..., x_n)$ e $Y = (y_1, y_2, ..., y_n)$ dois pontos do $R^n$. A distância Euclidiana entre X e Y é dada por

$$D(X, Y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

O KNN possui como parâmetro livre o número de K-vizinhos, que pode ser monitorado pelo usuário com o intuito de melhorar o desempenho do classificador. A classificação pode ser computacionalmente intensiva quando se trata de um conjunto de dados com grande dimensão. Na Figura 22 tem-se um exemplo de classificação pelo KNN.

Figura 22 - Classificação pelo método KNN. Para uma amostra desconhecida $x_d$ entre amostras da classe 1 e 2. Dependendo do número de vizinhos mais próximos, $x_d$ pode ser classificada como segue: se $K = 1$, $x_d$ é classificado como "+", se $K = 3$, $x_d$ é classificado como "+", se $K = 5$, $x_d$ é classificado como "*"




Fonte: Elaborado pelo autor (2020).

#### 4.3.4.4 *Árvore de Decisão*

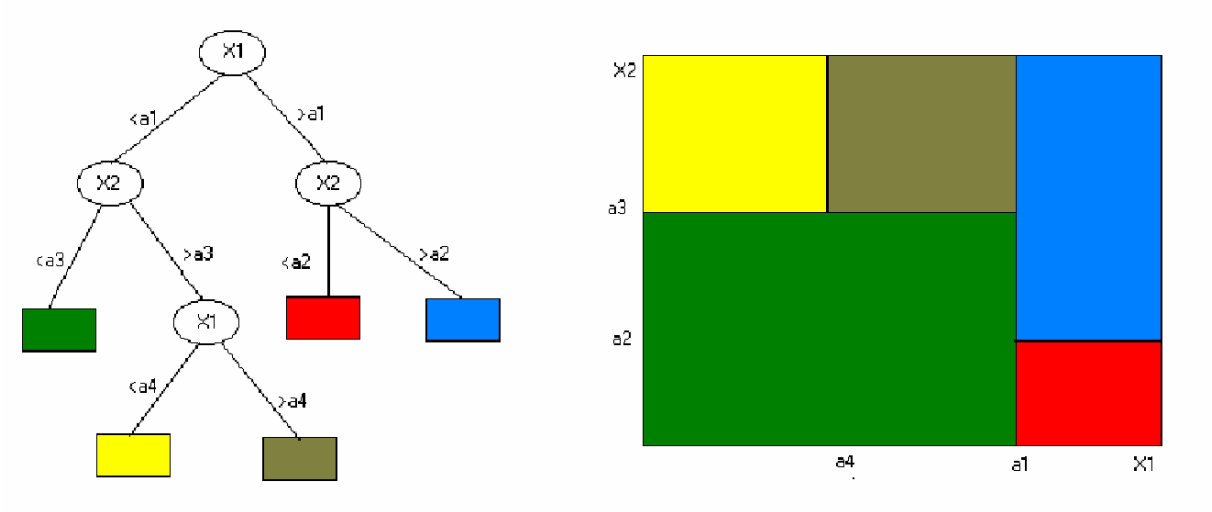
A Árvore de Decisão (Decision Tree, DT) é um modelo estatístico em que na sua criação utiliza-se de um conjunto de treinamento constituído por entradas e classes. Este classificador faz uso da abordagem dividir para conquistar, ou seja, um problema complexo é dividido em problemas menores e mais simples (subproblemas) e de maneira recursiva este modelo é empregado a cada subproblema. Da árvore de decisão pode-se retirar regras do tipo se-então-senão que são compreendidas naturalmente. A habilidade de diferenciação de uma árvore vem da divisão do espaço de características em espaços menores (subespaço) e a cada subespaço é conectado a uma classe (113).

A Figura 23 apresenta uma árvore de decisão na qual cada nó de decisão possui um teste para certa característica, cada ramo procedente equivale a um possível valor desta característica, o conjunto de ramos são diferentes, cada folha indica uma classe e, cada caminho percorrido da árvore, da raiz à folha corresponde uma regra de classificação. No espaço de características, cada folha equivale a um hiper retângulo no qual a interseção destes é vazia e a união é todo o espaço (114).

O critério utilizado para realizar as partições é o da utilidade do atributo para a classificação. Nos casos em que a árvore é usada para classificação, os critérios de partição mais conhecidos são baseados na entropia e índice Gini.

A Entropia é definida como o cálculo do ganho de informação levando em consi-

Figura 23 - Árvore de decisão e sua respectiva exibição no espaço




Fonte: Modificado de (115).

deração uma medida aplicada na teoria da informação. A entropia descreve a impureza dos dados: em uma base de dados, mede a falta de homogeneidade dos dados de entrada em relação a sua classificação. Quando o conjunto de dados é heterogêneo a entropia é máxima (igual a 1) (116). Dado um conjunto de entrada (X) que pode ter C classes diferentes, a entropia de X será dada pela Eq. 4.10.

$$\text{Entropia}(X) = \sum_{i=1}^{C} -p_i \log_2 p_i \tag{4.10}$$

onde $p_i$ é a proporção de dados de X que pertencem à classe i.

O índice Gini ($IG$) mede o grau de heterogeneidade dos dados, podendo ser usado para medir a impureza de um nó. Para um determinado nó é dado pela Eq. 4.11.

$$IG = 1 - \sum_{i=1}^{C} p_i^2 \tag{4.11}$$

onde $p_i$ é a frequência relativa de cada classe em cada nó e C é o número de classes.

O nó é puro quando este índice é igual a zero e quando ele se aproxima do valor um, o nó é impuro. Nas árvores de classificação com partições binárias quando se usa o critério de Gini tende-se a isolar num ramo os registros que representam a classe mais frequente. No caso da entropia, balanceia-se o número de registros em cada ramo (116).

Existem muitos algoritmos de classificação que utilizam a árvore de decisão. De acordo com o problema, um algoritmo pode ter melhor desempenho que outro. Dentre os algoritmos tem-se: ID3, CART, Assistant, C4.5, C5 e CHAID. Aqui será utilizado o algoritmo CART, pois é o que está implementado do *Scikit-Learn*. Este algoritmo utiliza a partição recursiva binária e suas principais características são a definição do conjunto de

regras para dividir cada nó da árvore, o fato de decidir quando a árvore está completa e associar cada nó terminal a uma classe ou a um valor preditivo no caso de regressão. Os procedimentos do CART são:

- Determinar o conjunto de regras para dividir de um nó em dois nós filhos. As perguntas do algoritmo têm como resposta "sim" ou "não".
- Descobrir a melhor divisão através do critério Gini.
- Repetir o processo de divisão até que este seja impossível ou interrompido.
- Empregar o processo de pós-podagem para determinar a árvore com o menor custo.

#### 4.3.4.5 *Gradient Boosting*

O Gradient Boosting (GB) é um método *ensemble* que produz um modelo de previsão usando uma coleção de modelos de árvore de decisão (117). O GB usa árvores de decisão de tamanho fixo como modelos de árvores de decisão. O GB considera modelos aditivos da seguinte forma:

$$F(x) = \sum_{m=1}^{N} \gamma_m h_m(x) \tag{4.12}$$

onde $h_m(x)$ são os modelos de árvore de decisão, $\gamma_m$ é o comprimento do passo de cada árvore e $N$ é o número de árvores. GB constrói o modelo aditivo da seguinte forma:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \tag{4.13}$$

onde o modelo inicial $F_0$ é problema específico, geralmente escolhe a média dos valores alvo.

Em cada estágio, a árvore de decisão $h_m(x)$ é escolhida para minimizar a função de perda $L$ dado o modelo corrente $F_{m-1}$ e sua função de ajuste $F_{m-1}(x_i)$ onde

$$F_m(x) = F_{m-1}(x) + \arg\min_{h} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x)) \tag{4.14}$$

O GB resolve este problema de minimização numericamente via método de gradiente descendente: a direção de descida mais acentuada é o negativo do gradiente da função de perda avaliada no modelo atual $F_{m-1}$ que pode ser calculado para qualquer função de perda diferenciável (118):

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_F L(y_i, F_{m-1}(x_i)) \tag{4.15}$$

onde o comprimento do passo $\gamma_m$ é escolhido usando a busca linear

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}). \tag{4.16}$$

Uma estratégia simples de regularização que melhora a qualidade de ajuste de cada modelo de árvore de decisão foi proposta por (117), que escala a contribuição de cada árvore de decisão por um fator $\nu$:

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x) \tag{4.17}$$

O parâmetro $\nu$ também é chamado de taxa de aprendizado, pois dimensiona o comprimento do passo do procedimento de descida do gradiente. Este parâmetro interage fortemente com o número de nós e a profundidade de árvores de decisão para ajustar. O tamanho de cada árvore pode ser controlado pelo número de nós de folha.

#### 4.3.4.6 *Máquina de Aprendizado Extremo*

A Máquina de Aprendizado Extremo (Extreme Learning Machine, ELM) (119) é uma rede neural artificial que possui uma única camada oculta. O ELM estabelece um equilíbrio entre velocidade e desempenho de generalização e atrai cada vez mais atenção por conta dos seus vários aspectos. Comparado com a Rede Neural Artificial (ANN), a Máquina de Vetor de Suporte (SVM) e outros modelos tradicionais de previsão, o modelo ELM mantém as vantagens de aprendizagem rápida, boa capacidade de generalização e conveniência em termos de modelagem (120).

Nos ELMs existem três níveis de aleatoriedade (121): (1) os parâmetros da camada oculta são gerados aleatoriamente, (2) nem todos os nós de entrada precisam estar conectados a um nó oculto em particular, e (3) um nó oculto em si pode ser uma sub-rede formada por vários nós, resultando no aprendizado de características locais. A função de saída do ELM usada neste trabalho é dada por

$$\hat{y}(\mathbf{x}) = \sum_{i=1}^{L} \beta_i G\left(\alpha, \mathbf{w}_i, b_i, \mathbf{c}, \mathbf{x}\right) = \sum_{i=1}^{L} \beta_i G\left(\alpha \mathrm{MLP}(\mathbf{w}_i, b_i, \mathbf{x}) + (1-\alpha)\mathrm{Rbf}(\mathbf{x}, \mathbf{c})\right) \tag{4.18}$$

onde $\hat{y}$ é a previsão do ELM associada ao vetor de entrada $\mathbf{x}$, $\mathbf{w}_i$ é o vetor de ponderação do $i$-ésimo nó oculto, $b_i$ são os bias dos neurônios na camada oculta, $\beta_i$ são os pesos de saída, $\mathbf{c}$ é o vetor de centros. MLP e Rbf são as funções de ativação de entrada, respectivamente, enquanto $\alpha$ é um multiplicador dos termos MLP$(\cdot)$ e Rbf$(\cdot)$. $G(\cdot)$ é a função de ativação de saída não linear e $L$ é o número de neurônios na camada oculta. A função de ativação de saída $G\left(\alpha, \mathbf{w}_i, b_i, \mathbf{c}, \mathbf{x}\right)$ com os pesos dos nós ocultos $(\mathbf{w}, b)$ são apresentados na Tabela 19.

Os parâmetros $(\mathbf{w}, b)$ são gerados randomicamente (normalmente distribuído com média zero e desvio padrão igual a um), e os pesos $\beta_i$ da camada de saída são determinados analiticamente. MLP e Rbf são as funções de ativação perceptron multicamadas e função de base radial, respectivamente, escritas como

$$\mathrm{MLP}(\mathbf{w}_i, b_i, \mathbf{x}) = \sum_{k=1}^{D} w_{ik} x_k + b_i \quad \text{e} \quad \mathrm{Rbf}(\mathbf{x}, \mathbf{c}) = \sum_{j=1}^{D} \frac{x_j - c_{ij}}{r_i} \tag{4.19}$$

Tabela 19 – Funções de ativação de saída usadas no ELM.

| # | Nome | Função de Ativação $G$ |
|---|---|---|
| 0 | Tribas | $G(x) = 1 - \lvert x \rvert$ se $-1 \geq x \geq 1$ caso contrário 0 |
| 1 | Identity | $G(x) = x$ |
| 2 | ReLU | $G(x) = \max(0, x_i; i = 1, \cdots, D)$ |
| 3 | Swish | $G(x) = \dfrac{x}{1 + exp(-x)}$ |
| 4 | Inverse Tribas | $G(x) = \lvert x \rvert$ se $-1 \geq x \geq 1$ caso contrário 0 |
| 5 | HardLim | $G(x) = 1$ se $x \geq 0$ caso contrário 0 |
| 6 | SoftLim | $G(x) = x$ se $0 \geq x \geq 1$ senão 0 se $x < 0$ caso contrário 1 |
| 7 | Gaussian | $G(x) = exp(-x^2)$ |
| 8 | Multiquadric | $G(x) = \sqrt{x^2 + b^2}$ |
| 9 | Inverse Multiquadric | $G(x) = \dfrac{1}{(x^2 + b^2)^{1/2}}$ |

Fonte: Elaborada pelo autor (2020).

onde $D$ é o número de características de saída, os centros $c_{ij}$ são tirados uniformemente do hiper-retângulo delimitador das variáveis de entrada e $r = \max(||\mathbf{x} - \mathbf{c}||)/\sqrt{2D}$.

O vetor de pesos de saída $[\beta_1, ..., \beta_L]$ pode ser encontrado pela minimização do erro de aproximação (122)

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^L} \|\mathbf{H}\boldsymbol{\beta} - \mathbf{y}\|$$

onde $\mathbf{y}$ é o vetor de dados de saída, $\mathbf{H}$ é a matriz de saída da camada oculta

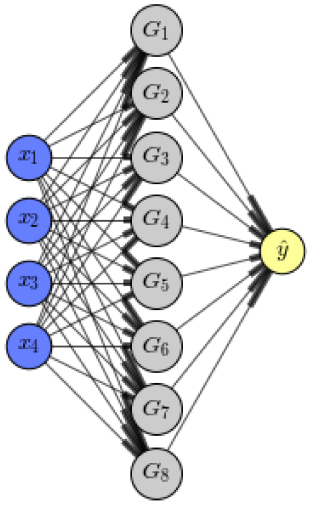
$$\mathbf{H} = \begin{bmatrix} G_1(\alpha, \mathbf{w}_1, b_1, \mathbf{c}, \mathbf{x}_1) & \cdots & G_L(\alpha, \mathbf{w}_L, b_L, \mathbf{c}, \mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ G_1(\alpha, \mathbf{w}_1, b_1, \mathbf{c}, \mathbf{x}_N) & \cdots & G_L(\alpha, \mathbf{w}_L, b_L, \mathbf{c}, \mathbf{x}_N) \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

é o vetor de dados de saída com $N$ pontos de dados. A solução ótima é dada por

$$\boldsymbol{\beta} = (\mathbf{H^T H})^{-1} \mathbf{H^T y} = \mathbf{H}^{\dagger} \mathbf{y}$$

onde $\mathbf{H}^{\dagger}$ é o pseudo inverso de $\mathbf{H}$. A Figura 24 apresenta um exemplo de 4-8-1 ELM com quatro entradas, uma camada oculta (8 neurônios) e uma saída.

Figura 24 - Conectividades para uma Máquina de Aprendizado Extremo 4-8-1.




Fonte: Elaborada pelo autor (2020).

### 4.3.5 **Validação Cruzada**

Os classificadores geralmente buscam aprender com o passado para conseguirem prever o futuro. O procedimento de aprendizagem é muito importante e a escolha do conjunto de treinamento, do conjunto de teste e dos parâmetros adequados influenciam diretamente no desempenho dos métodos de classificação. A Validação Cruzada aparece com uma alternativa bastante utilizada no contexto citado acima.

#### 4.3.5.1 *Técnicas de Validação Cruzada*

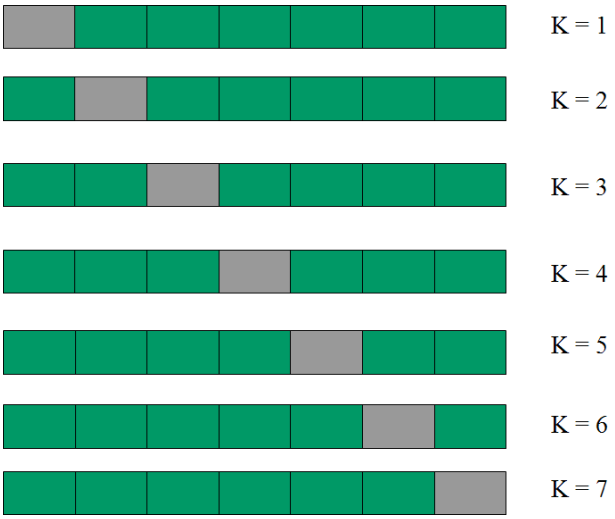
Para avaliar os classificadores descritos anteriormente, foram empregadas as estratégias de validação cruzada K-Fold (KF) (123).

No K-Fold (KF) a base de dados disponível contém $N$ amostras e é dividida em $K$ subconjuntos, onde $K > 1$. Depois da partição da base de dados, os K-1 subconjuntos gerados são usados para treinamento e o conjunto restante é usado para teste; dessa maneira, ao final do procedimento, é medido o erro de validação. Esse processo é repetido $K$ vezes usando um conjunto de teste distinto em cada iteração. O intuito desse método é treinar da melhor forma possível o classificador para que ele possa generalizar sobre as futuras entradas. A Figura 25 mostra um esquema exemplificando o K-Fold.

### 4.3.6 **Evolução Diferencial**

Evolução Diferencial (Differential Evolution, DE) é um método de otimização simples e eficiente que foi proposto por Rainer Storn e Kenneth Price em 1995 (107). É um método estocástico de busca que tinha como o objetivo de, inicialmente, resolver um problema de ajuste polinomial de Chebychev.

Figura 25 - K-Fold - $K = 7$. Conjunto de treinamento - 6 amostras (quadros verdes) e conjunto de teste - 1 amostra (quadro cinza)




Fonte: Elaborada pelo autor (2020).

O DE é um método que tem se mostrado eficiente numa grande classe de problemas, segundo (124). Apresenta-se eficaz para funções objetivo que não são diferenciáveis ou convexas e possui capacidade de encontrar a solução ótima com populações pequenas (125). Este método pode ser retratado como uma manipulação de indivíduos que representam as soluções candidatas. Com o passar das gerações, essas soluções candidatas sofrem transformações de mutação e cruzamento, onde são produzidas novas soluções candidatas, e posteriormente é realizada a seleção e o ciclo se repete.

Dada uma população de indivíduos, a mesma é sujeita a três operações: mutação, cruzamento e seleção. Essas operações são realizadas até que um critério de parada seja atingido, que pode ser um número definido de iterações ou um erro mínimo, por exemplo. Os operadores do DE se fundamentam no princípio da evolução natural que possuem os seguintes objetivos: preservar a multiplicidade da população, evitar convergências precipitadas e obter a melhor solução para o problema.

#### 4.3.6.1 *Mutação*

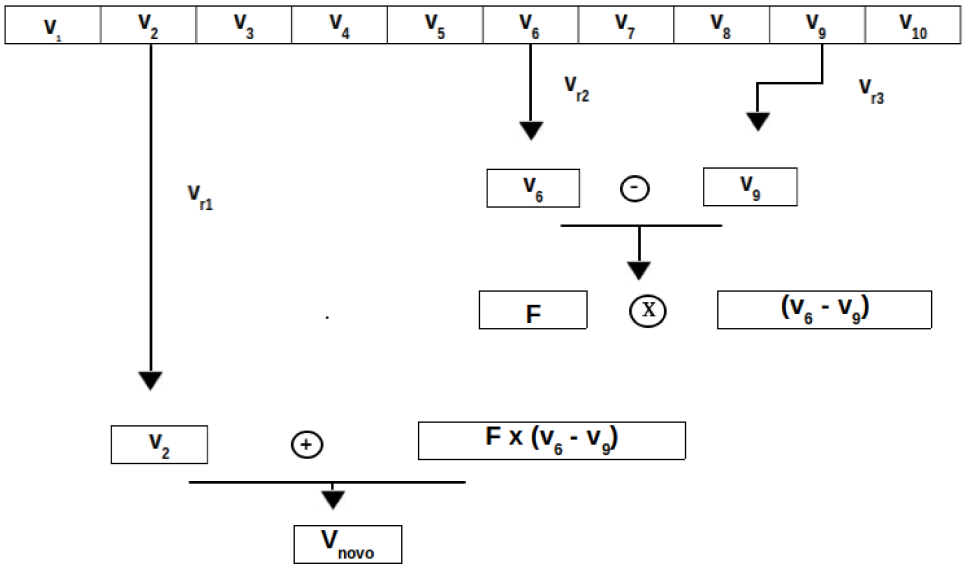
No operador mutação cada indivíduo é modificado a partir da soma da diferença vetorial ponderada entre dois indivíduos escolhidos aleatoriamente da população a um terceiro indivíduo. Os vetores novos são então gerados. O operador de mutação é definido da seguinte forma:

$$v_{novo} = v_{base} + F(v_1 - v_2) \tag{4.20}$$

onde $v_{novo}$ é o novo indivíduo gerado e F determina a ponderação da diferença entre $v_1$ e $v_2$. O vetor $v_{base}$ é o vetor base, que indica onde a perturbação é realizada. Esses três vetores são escolhidos de maneira aleatória.

Para garantir as diferenças entre os indivíduos selecionados aleatoriamente, a população deverá ser maior ou igual a 4 indivíduos. Ademais, o fator de mutação $F$, que controla a amplitude da diferença vetorial, recomenda-se estar no intervalo [0.5, 1] (107). A Figura 26 ilustra a operação de mutação.

Figura 26 - A operação de mutação




Fonte: Elaborada pelo autor (2020).

#### 4.3.6.2 *Cruzamento*

Com o objetivo de aumentar a diversidade da população, Storn e Price (107) apresentaram o operador de cruzamento. Os membros da população e os vetores mutantes trocam atributos para formar o vetor modificado. O vetor experimental $u_i^j$ é formado pela Equação 4.21

$$u_i^j = \begin{cases} v_i^j + F(v_{r1}^j - v_{r2}^j) \text{ se } r_i \leq \text{CR} \\ v_i^j, \text{caso contrário} \end{cases} \tag{4.21}$$

onde $r_i$ é um número gerado randomicamente e CR é um valor definido dentro de um intervalo e é informado pelo usuário, $v_i^j$ são as componentes do vetor alvo que pertence a população, e competirá com o novo vetor gerado.

#### 4.3.6.3 *Seleção*

O operador seleção é responsável por selecionar os melhores indivíduos. Este operador tem como objetivo escolher os indivíduos com melhores características, que serão mantidos para a próxima geração. Se a aptidão definida a partir do cálculo da

função objetivo do indivíduo $i$ da população corrente ($\vec{x}_{i,G+1}$) é maior do que a aptidão do indivíduo $i$ da população de cruzamento ($\vec{u}_{i,G+1}$), esse indivíduo passa para próxima geração com os melhores entre as duas populações, como mostra a Equação 4.22

$$\vec{x}_{i,G+1} = \begin{cases} \vec{u}_{i,G+1} \text{ se, } f(\vec{u}_{i,G+1}) \leq f(\vec{x}_{i,G}) \\ \vec{x}_{i,G}, \text{ caso contrário} \end{cases} \tag{4.22}$$

### 4.3.7 **Métricas para a seleção de modelos**

Realizou-se a avaliação dos métodos de classificação juntamente com os métodos de validação cruzada citados a partir das métricas acurácia (AC), recall (RECALL), F1 e Kappa. (126).

#### 4.3.7.1 *Acurácia*

A Acurácia mede a proporção das amostras que foram classificadas corretamente, como pode ser visto na Eq. (4.23),

$$\text{AC} = \frac{1}{N} \sum_{i=1}^{N} I(f(x_i) = y_i) \tag{4.23}$$

onde $f(x_i)$ é a classe predita pelo algoritmo de classificação e $y_i$ é a classe original da amostra. Considera se $I(verdadeiro) = 1$ e $I(falso) = 0$.

#### 4.3.7.2 *Recall*

O RECALL pode ser referido como a taxa de verdadeiros positivos ou sensibilidade, como definido na Eq. (4.24)

$$\text{RECALL}(k) = \frac{TP}{TP + FN} \tag{4.24}$$

onde TP é taxa de verdadeiro positivo e FN é a taxa de falso negativo. Os verdadeiros positivos são os testes que estão passando porque a aplicação está se comportando conforme o esperado e os falsos negativos são testes que estão falhando, porém, devido a inconsistência no próprio teste e não na aplicação testada. O RECALL para um problema com mais de duas classes é dado pela média do RECALL calculado para cada classe.

#### 4.3.7.3 *F1*

A medida de desempenho F1 pode ser definida como valor positivo preditivo

$$\text{F1} = \frac{2TP}{(2TP + FP + FN)} \tag{4.25}$$

onde TP é taxa de verdadeiro positivo, FN é a taxa de falso negativo e FP é a taxa de falsos positivos. Os falsos positivos são os testes que estão passando, porém, que deviam estar falhando. O F1 para um problema com mais de duas classes é dado pela média do F1 calculado para cada classe.

#### 4.3.7.4 *Kappa*

O Teste de Kappa é uma medida de concordância inter-classificador e mede o grau de concordância além do que seria esperado pelo acaso. Utiliza-se a medida Kappa para descrever se há ou não concordância entre dois ou mais avaliadores. Esta medida é baseada no número de respostas concordantes, isto é, no número de vezes em que o resultado é o mesmo entre os avaliadores. O valor máximo é 1, que indica concordância total. Pode-se obter valores próximos de 0 e até mesmo negativos, que representam nenhuma concordância. O coeficiente Kappa é calculado a partir da Eq. (4.26):

$$Kappa = \frac{P_o - P_E}{1 - P_E} \tag{4.26}$$

onde

$$P_o = \frac{n^\circ \quad de \quad concordâncias}{n^\circ \quad de \quad concordâncias \quad + \quad n^\circ \quad de \quad discordâncias} \tag{4.27}$$

e

$$P_E = \sum_{i=1}^{N} (p_{i1} \times p_{i2}) \tag{4.28}$$

sendo que N é o número de categorias, $i$ é o índice da categoria, $p_{i1}$ é a proporção de ocorrência da categoria $i$ para o avaliador 1 e $p_{i2}$ é a proporção de ocorrência da categoria $i$ para o avaliador 2. Para avaliar o nível de concordância, (127) sugerem a interpretação mostrada na Tabela 20.

Tabela 20 – Valor Kappa e Nível de Concordância

| Estatística Kappa | Nível de Concordância |
|---|---|
| $< 0.0$ | Nenhuma |
| $0.00 - 0.20$ | Pobre |
| $0.21 - 0.40$ | Leve |
| $0.41 - 0.60$ | Moderada |
| $0.61 - 0.80$ | Substancial |
| $0.81 - 1.00$ | Quase Perfeita |

Fonte: Elaborada pelo autor (2020).

### 4.3.8 **Teste de Wilcoxon**

Para comparar as médias de dois grupos pareados é recomendado o teste t de Student. No entanto, muitas vezes a variável de estudo não tem distribuição normal e que se deve usar este teste. Uma alternativa prática para resolver esta situação é o teste de Wilcoxon (128), no qual a exigência é de que a variável de estudo seja quantitativa ou qualitativa do tipo ordinal.

O objetivo é estudar as diferenças observadas em cada sujeito entre os dois momentos da pesquisa. O princípio deste teste consiste em avaliar se ocorreram modificações nos

dados entre os dois momentos da avaliação. Quando as modificações ou diferenças são muito pequenas elas podem ser devidas ao acaso, porém, quando são expressivas, é pouco provável que se devam ao acaso, sendo fruto de um fator causal. No teste de Wilcoxon as diferenças observadas entre os dois momentos são transformadas em ranks, que passam a ser os objetos da nossa avaliação.

Os seguintes passos são seguidos para a realização do teste de Wilcoxon:

1. Calcula-se para cada elemento do grupo de estudo a diferença $d$ entre a suas duas medidas, seja a primeira menos a segunda ou vice-versa
2. Ordena-se e atribui-se ranks aos valores absolutos das diferenças $d$ que sejam diferentes de zero.
3. Somam-se os ranks decorrentes das diferenças positivas, que recebe o nome de T+
4. Somam-se os ranks decorrentes das diferenças negativas, que recebe o nome de T
5. Seleciona-se entre T+ e T- o de menor valor, que será chamado de estatística T
6. Chamamos de $n$ o número de diferenças $d$ que receberam ranks.

As hipóteses a serem consideradas são descritas abaixo:

- H0: não há diferença entre os valores observados nos dois instantes
- H1: há diferença entre os valores observados nos dois instantes

O nível de significância habitualmente adotado nas pesquisas da área biomédica é 5% ($\alpha$=0,05). Para amostras com N $\leq$ 25, compara-se a estatística T com o valor T crítico, disponível em uma tabela específica de distribuição de T para o teste de Wilcoxon, considerando $\alpha$ e $n$ (número de casos em que d $\neq$ 0). Sempre que T $\leq T_{critico}$ a H0 é rejeitada.

Quando N > 25, a estatística T se ajusta à distribuição normal. Portanto, calcula-se o $z$-escore $Z_T$ de T e determina-se a probabilidade de ocorrência pela tabela da curva normal reduzida.

$$Z_T = \frac{T - \mu_T}{\sigma_T} \tag{4.29}$$

Calculando a média de T

$$\mu_T = \frac{n(n+1)}{4} \tag{4.30}$$

Calculando o desvio padrão de T

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{4.31}$$

Quando houver empates deve-se utilizar a fórmula abaixo:

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1) - C/2}{24}} \tag{4.32}$$

onde C é o total de pontos descontados em decorrência da repetição de valores iguais (empates) e n é o número de diferenças que receberam ranks (129). O Teste de Wilcoxon será utilizado para verificar se os resultados encontrados pelos classificadores são significativamente diferentes.

### 4.3.9 **Análise de Componentes Principais**

O procedimento descrito a seguir foi baseado em (130) e em (131). A Análise de Componentes Principais tem como objetivo tomar $p$ variáveis $X_1, X_2, ..., X_p$ e encontrar combinações lineares destas para gerar índices $Z_1, Z_2, ..., Z_p$ não correlacionados na sua ordem de importância, que represente a variação nos dados. Ser não correlacionados indica que os índices estão medindo dimensões diferentes dos dados. Sendo a ordem de tal forma que $Var(Z_1) \geq Var(Z_2) \geq ... \geq Var(Z_p)$, onde $Var(Z_i)$ indica a variância de $Z_i$. Os índices $Z$ são denominados como as componentes principais.

O procedimento se inicia com uma base de dados com $p$ variáveis e $n$ amostras. A combinação linear das variáveis originais $X_1, X_2, ..., X_p$ denomina a primeira componente principal.

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

Dessa forma, a primeira componente principal é escolhida de maneira que $Var(Z_1)$ seja a maior possível sujeito a restrição

$$a_{11}^2 + a_{12}^2 + \ldots + a_{1p}^2 = 1$$

A segunda componente principal é escolhida de forma análoga, acrescida da condição que a covariância entre $Z_1$ e $Z_2$ seja zero.

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

sujeito a

$$\begin{cases} a_{21}^2 + a_{22}^2 + \ldots + a_{2p}^2 = 1 \\ Cov(Z_1, Z_2) = 0 \end{cases}$$

As demais componentes principais são definidas do mesmo modo. Para $p$ variáveis originais, podem existir no máximo $p$ componentes principais.

Um dos procedimentos que faz parte é obter autovalores da matriz de covariância amostral. A matriz de covariância tem a seguinte forma

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{21} & c_{22} & \dots & c_{2p} \\ \vdots & \vdots & & \vdots \\ c_{p1} & c_{p2} & \dots & c_{pp} \end{bmatrix}$$

onde o elemento da diagonal, $c_{ii}$, é a variância de $X_i$ e $c_{ij}$, elemento que não faz parte da diagonal, é a covariância entre as variáveis originais $X_iX_j$.

Os autovalores da matriz $C$ são as variâncias das componentes principais. Existindo $p$ autovalores, sendo que estes podem ter o valor zero e não podem ser negativos, uma vez que se trata de matriz de covariância. Parte do pressuposto que os autovalores estão ordenados, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, portanto $\lambda_i$ equivale a i-ésima componente principal

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_pX_p$$

Então, $Var(Z_i) = \lambda_i$ e as constantes $a_{i1}, a_{i2}, ..., a_{ip}$ são elementos do equivalente autovetor, de maneira que a restrição seja satisfeita

$$a_{i1}^2 + a_{i2}^2 + \ldots + a_{ip}^2 = 1$$

Os autovalores possuem uma particularidade importante, o fato que a soma deles é igual ao traço da matriz de covariância $C$.

$$\lambda_1 + \lambda_2 + \ldots + \lambda_p = c_{11} + c_{22} + \ldots + c_{pp}$$

Isto deve-se ao fato de $c_{ii}$ ser a variância de $X_i$ e $\lambda_i$ a variância de $Z_i$, implicando que a a soma das variâncias das componentes principais é igual a soma das variâncias das variáveis originais. Pode-se, então, dizer que as componentes principais contam com toda a variação nos dados originais.

Com intuito de prevenir que variáveis tenham uma influência errônea nas componentes principais, é usual realizar um pré-processamento nos dados, para que as variáveis $X_1, X_2, ..., X_p$ tenham médias zero e variâncias um no começo da aplicação do procedimento. Isso é realizado afim de garantir que as primeiras componentes principais armazenam maior porcentagem de informação das variáveis originais. A matriz $C$ adota a seguinte forma

$$C = \begin{bmatrix} 1 & c_{12} & \dots & c_{1p} \\ c_{21} & 1 & \dots & c_{2p} \\ \vdots & \vdots & & \vdots \\ c_{p1} & c_{p2} & \dots & 1 \end{bmatrix}$$

onde $c_{ij}$ é a correlação entre as variáveis $X_i$ e $X_j$. Portanto, a Análise de Componentes Principais é realizada sobre a matriz de correlação. A soma dos autovalores, ou seja, dos elementos da diagonal, é igual a $p$.

Com base nessa descrição do procedimento, as etapas da Análise de Componentes Principais podem ser relatadas:

1. Realização do Pré-processamento, para que as varáveis $X_1, X_2, ..., X_p$ tenham médias zeros e variâncias um.

2. Cálculo da matriz de covariâncias $C$ ou matriz de correlação, caso a Etapa 1 tenha sido realizada.

3. Cálculo para encontrar os autovalores $\lambda_1, \lambda_2, ..., \lambda_p$ e os respectivos autovetores $a_1, a_2, ..., a_p$. Os coeficientes da i-ésima componente principal são os termos de $a_i$, e $\lambda_i$ sua variância.

4. Exclusão das componentes que expressam uma pequena proporção nos dados.

### 4.3.10 **Análise Filogenética**

Para a Análise Filogenética utilizou o método Neighbor-Joining (132), análise de silhueta e para visualização utilizou-se o pacote ETE 3 (133). A Filogenia é uma hipótese acerca das relações de parentesco entre os seres vivos. As filogenias são representadas a partir de árvores filogenéticas. Pode-se dizer que uma árvore filogenética seria uma espécie de árvore genealógica dos tipos de seres vivos (134). Ela pode incluir vários níveis de agrupamentos dos seres vivos. A filogenia aqui será empregada para representar os processos associados à diagênese, um estudo dos constituintes presentes nas bases dados e o tempo de deposição. Duas abordagens serão adotadas: a primeira analisando amostras e a segunda os constituintes. Cada amostra foi retirada de profundidades diferentes, cada uma tem sua composição e tempo de deposição. Através da filogenia pretende-se chegar a relações sobre a história deposicional da rocha e de suas similaridades. Em relação aos constituintes serão identificados os eventos que ocorreram na eodiagênese, mesodiagênese e telodiagênese.

Para realizar essa tarefa utilizou-se pacotes de análise filogenética. Uma das alternativas é empregar o pacote *Scikit-Bio* (software de código aberto, disponível em <http://scikit-bio.org/>) que permite a leitura, escrita, simulação, processamento e manipulação de árvores filogenéticas. Trabalhos publicados recentemente mostraram que o uso dessa abordagem em Linguística Computacional (135, 136) gerou bons resultados para determinar a diversificação das línguas indo-europeias, então pretende-se avaliar a aplicabilidade da técnica para classificar os dados petrográficos.

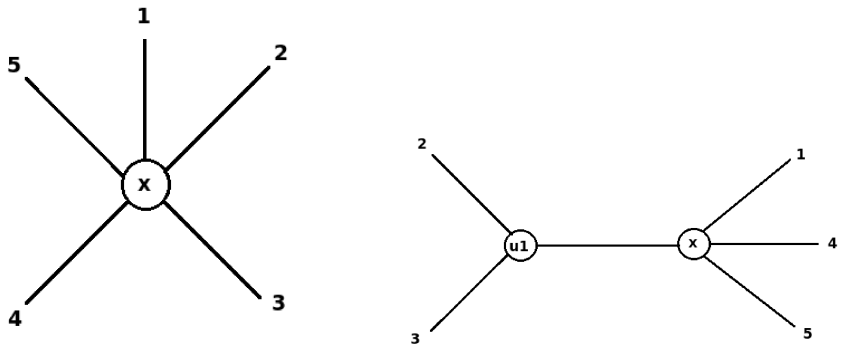
#### 4.3.10.1 *Neighbor-Joining*

Neighbor-Joining (132) é um método para geração de árvores filogenéticas através de uma matriz de distâncias empregando o princípio da evolução mínima e calculando os comprimentos de cada ramo dessa árvore. Este método busca construir uma árvore que torne a soma das distâncias de todos os ramos a menor possível, seguindo critérios de Evolução Mínima Equilibrada (BME - Balanced Minimum Evolution) (137). Para cada topologia, o método BME define o comprimento dos ramos para que seja uma soma de distâncias presente na matriz de distâncias, sendo estas distâncias dependentes da topologia utilizada.

A topologia de BME ótima será aquela que minimize a distância de todos os ramos. O algoritmo de Neighbor-Joining tenta juntar pares de nós cuja distância seja a menor possível para o cálculo do comprimento da árvore. Este processo não garante que ao final a árvore obtida terá a melhor topologia BME possível.

Os nós vizinhos são definidos com um par que possui uma ligação entre eles. Na Figura 27, pode-se observar como a árvore se inicia e como fica após o primeiro par ser agrupado.

Figura 27 - Possíveis estados da árvore

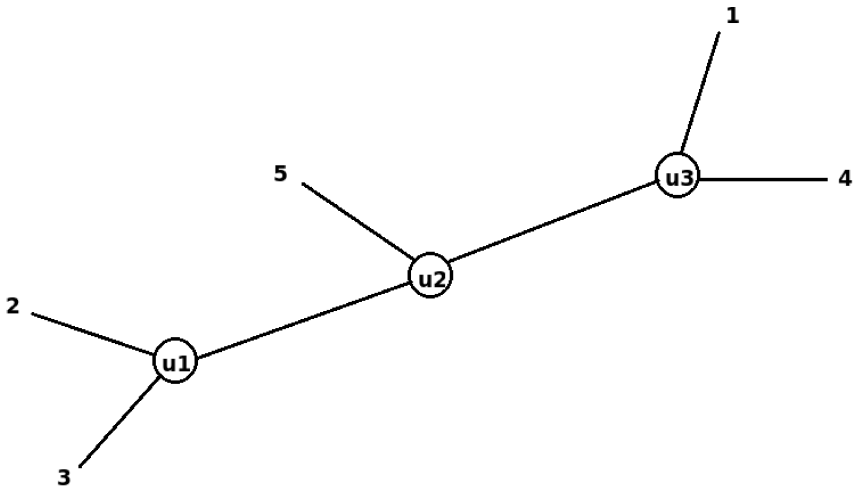



(a) Estado inicial da árvore. (b) Árvore após a união dos nós 2 e 3.

Fonte: Elaborada pelo autor (2020).

A árvore de evolução gerada nem sempre é a árvore de evolução mínima, uma vez que, minimizar o comprimento da árvore a cada passo do algoritmo não acarreta minimizar o comprimento da árvore gerada no final do processo. Isto ocorre porque a soma das distâncias mínimas calculadas a cada iteração pode diferenciar-se de acordo com a escolha realizada nas ocorrências de empate. Quando uma distância mínima é detectada em vários pares de nós acontece um empate, possibilitando que a escolha recaia sobre qualquer um deles.

Figura 28 - Árvore gerada pelo algoritmo Neighbor-Joining




Fonte: Elaborada pelo autor (2020).

Iniciando com uma árvore em forma de estrela, de forma recursiva o par de nós vizinhos com menor distância (soma mínima dos comprimentos destes ramos) é determinado. Para tal uma matriz de distâncias é utilizada, na qual cada elemento deverá reapresentar a relação de distância entre todos os nós.

Este algoritmo pode ser descrito em alguns passos que se repetirão $N - 2$ vezes (onde $N$ é o número de elementos a calcular no momento).

1. Realiza o cálculo da soma de um dado nó $i$ para todos os restantes nós.
2. Procura-se qual par de nós $(i, j)$ que terá a menor distância.
3. Cria a união deste nós e calcula a distância dos seus ramos (uma para cada elemento do par de nós de menor distância).
4. Após a criação da união, este é então colocado na árvore no lugar do ar de elementos (Figura 28(b)).
5. Recalcula a matriz de distâncias.

O procedimento é repetido até que apenas exista uma matriz de distâncias com dois elementos, complementando a árvore com o último valor presente na matriz. A Figura 28 ilustra uma árvore obtida após a aplicação do método Neighbor-Joining.

# 5 EXPERIMENTOS COMPUTACIONAIS

Os experimentos se dividem em dois grandes grupos: supervisionado e não supervisionado. Na parte supervisionada a aprendizagem é realizada a partir das classes definidas previamente pelo analista. O sistema determina a descrição para cada classe, isto é, o conjunto de propriedades comuns nos exemplos que lhe são fornecidos. Dessa forma, é possível formular a regra de classificação que pode ser utilizada para prever a classe de um objeto que não tenha sido considerado durante a aprendizagem. A parte não supervisionada é efetuada com base em observação e descoberta. As classes não são definidas antes, então o sistema necessita observar as amostras e reconhecer os padrões por si próprio. Gerando grupos com amostras semelhantes e propondo o conjunto de classes. Os experimentos serão descritos a seguir separadamente.

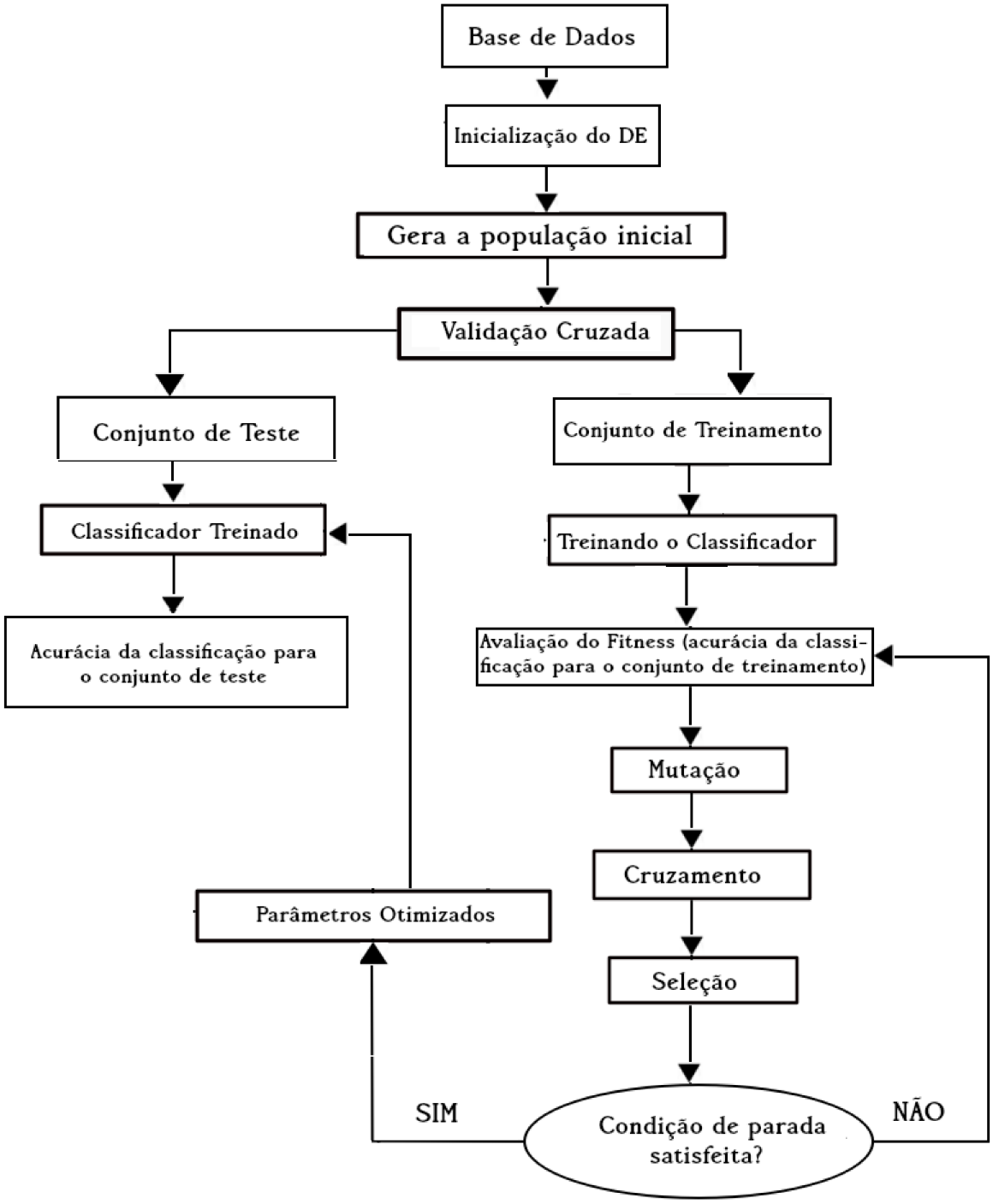
## 5.1 ESTRATÉGIA DE APRENDIZADO SUPERVISIONADO

A estratégia de aprendizado supervisionado envolve duas abordagens: híbrida e booststrap e análise de agrupamento. Essas abordagens serão descritas a seguir.

### 5.1.1 **Abordagem Híbrida**

A abordagem híbrida consiste no algoritmo de Evolução Diferencial (DE) + Classificador. O DE foi empregado para encontrar os melhores parâmetros possíveis para os classificadores. A Figura 29 mostra o esquema representando o procedimento utilizado para uma iteração.

Figura 29 - Esquema ilustrando o procedimento DE + Classificador




Fonte: Elaborada pelo autor (2020).

### 5.1.2 **Bootstrap e Análise de Agrupamento**

O Bootstrap foi aplicado no método de agrupamento K-Means para identificar as coordenadas dos centroides ótimas, a partir da maximização da silhueta, que posteriormente foi utilizado como parâmetro para os dados originais. Para empregar o Bootstrap deve-se conhecer as classes previamente, fazendo com que nesse caso seja um experimento

supervisionado.

## 5.2 ESTRATÉGIA DE APRENDIZADO NÃO SUPERVISIONADO

O método de agrupamento utilizado em todos testes foi o K-Means. O parâmetro de entrada é o número de agrupamentos que foi selecionado através de uma busca exaustiva no intervalo de 1 a 10, onde o objetivo era maximizar o valor do coeficiente de silhueta.

### 5.2.1 **Análise de Agrupamento e Análise de Componentes Principais**

A Análise de Componentes Principais foi utilizada com dois objetivos: (I) analisar quais características petrofísicas possuem maior importância sobre as componentes e (II) auxiliar na visualização dos resultados gerados pelo método de agrupamentos. O método de agrupamento foi utilizado para identificar as classes litológicas a partir de dados petrofísicos.

### 5.2.2 **Características Polinomiais e Análise de Agrupamento**

O número de constituintes dos de cada amostra foi aumentado através das características polinomiais. A partir desse procedimento avaliou-se os agrupamentos encontrados com o cálculo do coeficiente de silhueta. O intuito é verificar se os grupos encontrados possuem a distância intragrupo menor e a intergrupo maior, ou seja, se os grupos formados estão melhor definidos.

## 5.3 RESULTADOS E DISCUSSÕES

Para avaliar a metodologia proposta, experimentos extensivos foram realizados para comparar o desempenho da estrutura com bases de dados petrográficas e petrofísicas. Nesta seção, será descrito os resultados de acordo com o experimentos que foram aplicados em cada base de dados.

### 5.3.1 **Dados Petrográficos - Tibagi**

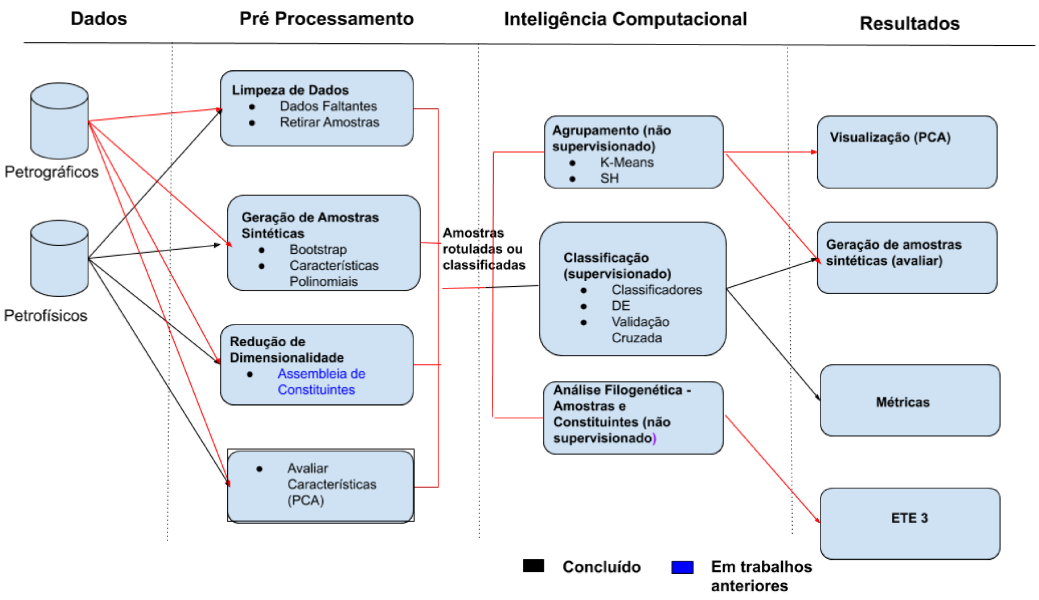
Os procedimentos da metodologia descritos nas Seções 5.1 e 5.2 que foram aplicados nas bases de dados Tibagi, Tibagi - Constituintes Diagenéticos e Tibagi - sem I-1 e I-2 estão ilustrados na Figura 30 mostra, através das linhas vermelhas.

Para a base Tibagi a Análise Filogenética foi aplicada nas amostras pois o objetivo é verificar se amostras pertencentes a mesma petrofácies passaram por processos diagenéticos semelhantes. A base de dados Tibagi - sem I-1 e I-2 é formada pelas amostras da Tibagi com exceção das amostras pertencentes às petrofácies I-1 e I-2. Essas petrofácies são de unidades litoestraigráficas diferentes das demais o que pode interferir no resultado. Essas duas pertencem ao Grupo Itararé e as demais ao Grupo Paraná o que por si só é

um classificador litológico (e possivelmente diagenético) a priori. Então, realizou-se esse experimento para verificar essa questão.

Na base de dados Tibagi - Constituintes Diagenéticos utilizou-se apenas os constituintes diagenéticos uma vez que essas características são aquelas analisadas para identificar os eventos diagenéticos. Nesse caso, a Análise Filogenética foi empregada nos constituintes com o objetivo de identificar esses eventos.

Figura 30 - Fluxograma ilustrando, através da linhas vermelhas, a metodologia aplicada a base de dados Tibagi.




Fonte: Elaborada pelo autor (2020).

A Análise de Componentes Principais foi aplicada para avaliar a influência das propriedades sobre as componentes principais. Na Tabela 21 encontram-se as componentes obtidas. Para a Componente 1 as características Bioclasto, Caolinita, Porosidade Intergranular, Opaco e Turmalina influenciam mais, na Componente 2 são as características Bioclasto, Porosidade Intergranular, Muscovita, Opaco, Rutilo e Pseudo Matriz. Dentre essas características a Caolinita, Porosidade Intergranular e Pseudo Matriz são referentes aos processos diagenéticos identificados nas lâmina que auxiliam na identificação de petrofácies sedimentares.

A Tabela 22 apresenta os resultados da análise de variância intergrupos para cada característica. Pode-se observar que $H_0$ não é rejeitada para todos as características, o que indica que a média de algumas são iguais para os grupos encontrados. O $H_{cal}$ é o valor calculado pelo método de *Kruskal-Wallis*, nos caso em que p-valor $< 5\%$, $H_{cal}$ é maior que o valor tabelado na Distribuição Qui-Quadrado.

Tabela 21 – Componentes Principais em relação as propriedades petrográficas. Os valores com * indicam os valores mais significativos obtidos para cada componente sugerindo qual propriedade petrográfica tem mais atuação sobre as mesmas - Tibagi

| Características | Componente 1 | Componente 2 |
|---|---|---|
| Bioclasto | -8.26452212E-01* | -1.18447342E-01* |
| Cresc. Sec. Qtz | 8.97042479E-03 | -4.50971694E-02 |
| Caolinita | 2.11148283E-01* | 7.38066229E-02 |
| Ilita/Smectita | 1.57921625E-02 | 3.72060078E-02 |
| Pirita | -2.72258747E-03 | -2.71797277E-03 |
| Siderita | 7.68825042E-03 | -3.35381007E-03 |
| Cim. Carbonático | 2.33490846E-03 | 1.44745471E-03 |
| Cim. Silicoso | 4.25796345E-04 | 1.07299266E-02 |
| Cim. Ferruginoso | 1.51636925E-03 | -2.59604297E-04 |
| Por. Intergranular | 2.90236152E-01* | -1.12512686E-01* |
| Por. Intragranular | -9.46541085E-03 | -1.10813632E-02 |
| Quartzo | -2.11772824E-03 | 4.90316885E-04 |
| Feldspato | -2.20040638E-04 | 8.78870451E-03 |
| Muscovita | -4.21824150E-02 | -3.86005159E-01* |
| Opaco | 4.03675901E-01* | -1.60569242E-01* |
| Turmalina | 1.17732599E-01* | -1.62624959E-02 |
| Zircão | 2.53961254E-02 | -9.16899778E-03 |
| Rutilo | -4.91245812E-02 | 8.73204991E-01* |
| Glauconita | 1.56159425E-03 | 1.70142842E-04 |
| Clorita | 5.85455514E-03 | -6.48219073E-04 |
| Pseudo Matriz | -7.67553822E-02 | -1.62759849E-01* |
| Litoclasto | -8.76537740E-03 | 5.26967179E-04 |

Fonte: Elaborada pelo autor (2020).

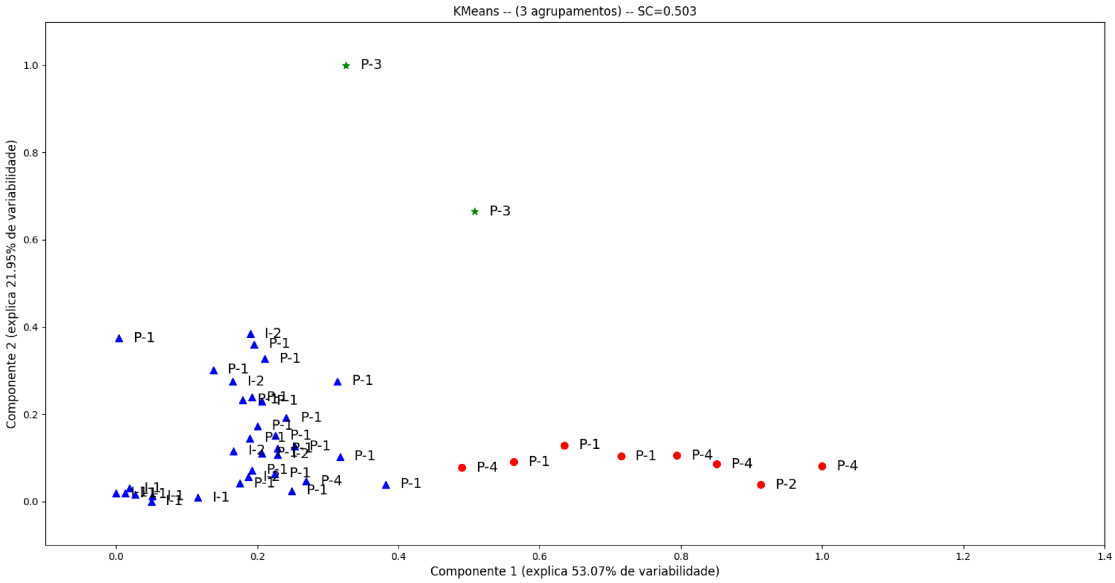
A Figura 31 exibe o resultado do K-Means para o Tibagi. A primeira componente principal expressa 53.07% de variabilidade dos dados, a segunda expressa 21.95% da variabilidade. Três grupos foram identificados, onde as amostras pertencentes a P-3 foram agrupadas em um grupo nos demais o método encontrou semelhança nas amostras de petrofácies diferentes.

Tabela 22 – Análise de variância intergrupos para cada característica - Tibagi. Os valores com * indicam os p-valores que rejeitaram a $H_0$. Os grupos foram determinados pelo procedimento computacional. O nível de significância é de 0.05.

| Características | $H_{cal}$ | p-valor |
|---|---|---|
| Bioclasto | 3.502895 | 0.173523 |
| Cresc. Sec. Qtz | 12.678612 | 0.001766* |
| Caolinita | 1.828926 | 0.400732 |
| Ilita/Smectita | 6.137484 | 0.046480* |
| Pirita | 2.003017 | 0.367325 |
| Siderita | 1.670127 | 0.433847 |
| Cim. Carbonático | 16.591191 | 0.000250* |
| Cim. Silicoso | 2.253789 | 0.324038 |
| Cim. Ferruginoso | 1.091542 | 0.579395 |
| Por. Intergranular | 9.370579 | 0.009230* |
| Por. Intragranular | 1.056016 | 0.589779 |
| Quartzo | 4.314257 | 0.115657 |
| Feldspato | 4.839918 | 0.088925 |
| Muscovita | 14.675509 | 0.000651* |
| Opaco | 4.016212 | 0.134243 |
| Turmalina | 0.316157 | 0.853783 |
| Zircão | 23.500511 | 0.000008* |
| Rutilo | 7.149994 | 0.028016* |
| Glauconita | 2.908219 | 0.233608 |
| Clorita | 5.125000 | 0.077112 |
| Pseudo Matriz | 6.190128 | 0.045272* |
| Litoclasto | 6.485507 | 0.039056* |

Fonte: Elaborada pelo autor (2020).

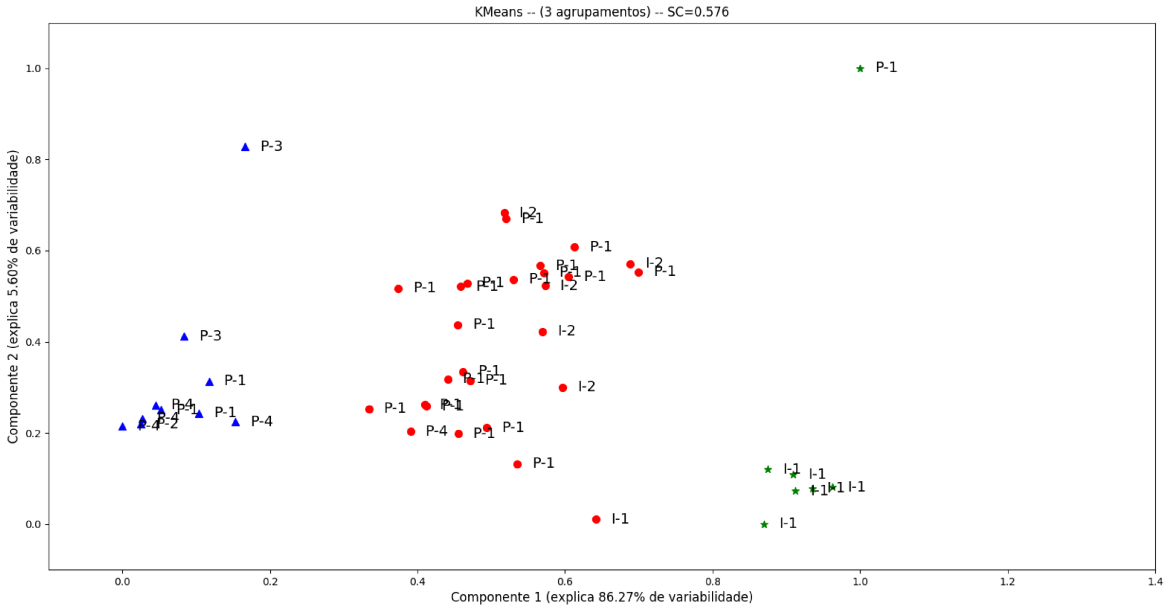
Figura 31 - Resultado K-Means - Tibagi. SC = 0.503




Fonte: Elaborada pelo autor (2020).

A Figura 32 exibe o resultado do K-Means aplicado nos dados após o pré-processamento o acréscimo das características polinomiais. O grau do polinômio foi 4 e os dados ficaram com 14950 características. A primeira componente principal expressa 86.27% de variabilidade dos dados, a segunda expressa 5.60% da variabilidade. Pode-se observar que houve um aumento no valor do coeficiente de silhueta em relação aos dados originais. Três grupos foram encontrados pelo método e em todos há amostras pertencentes a diferentes petrofácies.

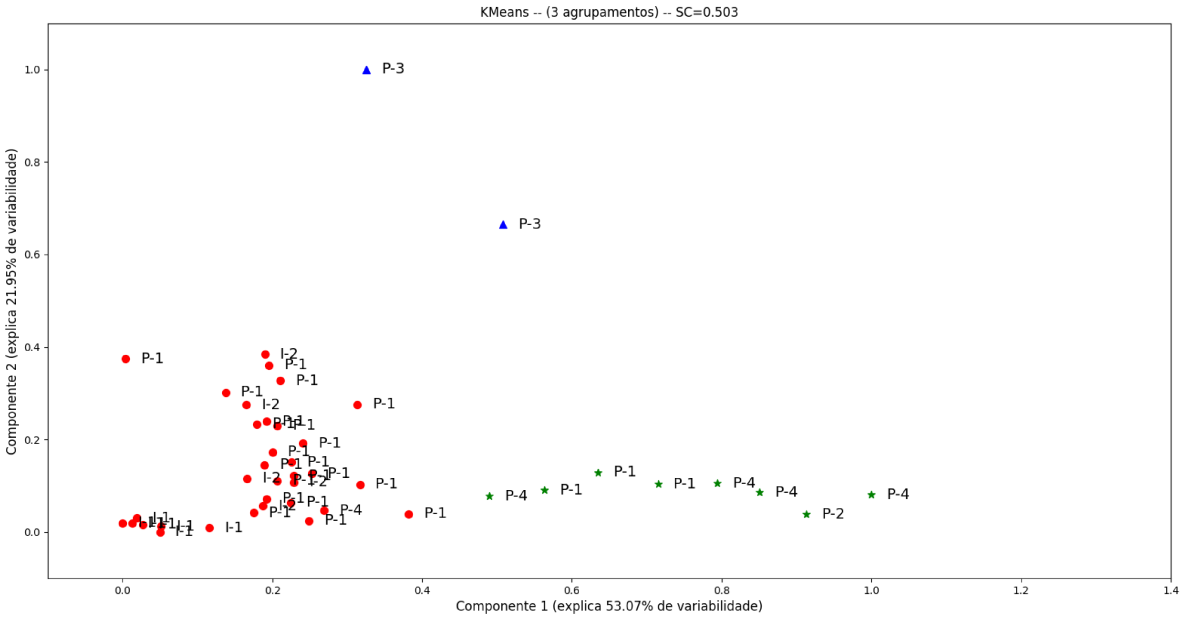
Figura 32 - Resultado K-Means (Características Polinomiais) - Tibagi. SC = 0.576




Fonte: Elaborada pelo autor (2020).

O K-Means aplicado nos dados originais após o Bootstrap ter sido empregado para encontrar os parâmetros. A Figura 33 exibe o resultado. Nota-se que foi semelhante ao gerado pelo K-Means quando utilizado nos dados originais, uma vez que os parâmetros sugeridos pelo Bootstrap foram similares.

Figura 33 - Resultado K-Means (Bootstrap) - Tibagi. SC = 0.503



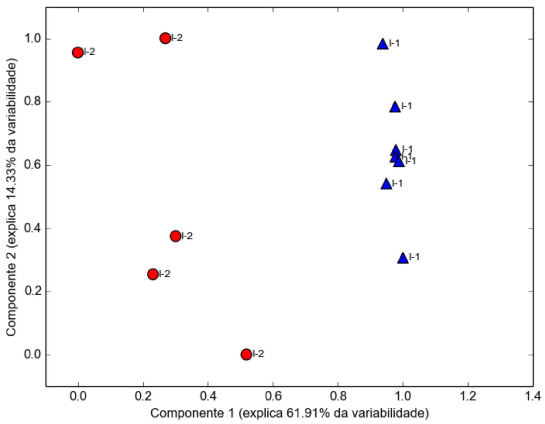

Fonte: Elaborada pelo autor (2020).

Com intuito de comparar o resultado obtido através da Análise Filogenética, dois procedimentos (Análise Intra-Poço e Inter-Poço) que constam em (138) serão expostos. Análise Intra-Poço tem o intuito de verificar como os métodos procedem diante de cada poço, de forma individual e a Análise Inter-Poço de analisar como ocorre a divisão das lâminas em grupos quando as informações de todos os poços estão armazenadas na mesma base de dados.

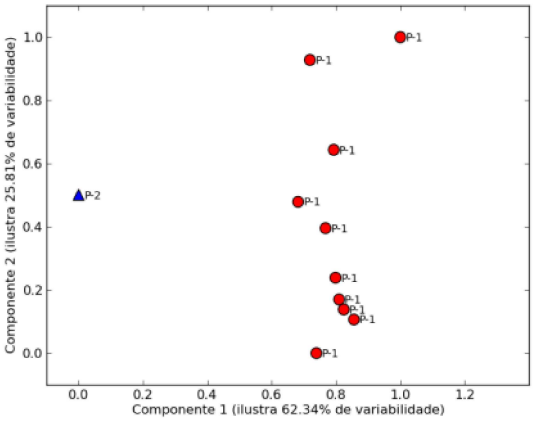
O resultado da Análise Intra-Poço pode ser visto na Tabela 23 e na Figura 34. O número de agrupamentos e o valor do coeficiente de silhueta encontra-se na Tabela 23. A Figura 34 ilustra os agrupamentos encontrados para os poços PPG1, PPG2, PPG3, PPG4 e PPG5, respectivamente. Para o poço PPG1 o método gerou 2 agrupamentos, coincidindo com o método manual nas classificação das classes I-1 e I-2, agrupando 100% das amostras de forma correta. Nos poços PPG2 e PPG3 o K-Means gerou 2 agrupamentos, concordando com a classificação de (139). No PPG4, o método encontrou 2 agrupamentos, diferindo da classificação manual, uma vez que ela propôs a existência de uma única petrofácies (P-4). Duas amostras segundo o K-Means (PPG4-1 e PPG4-5) não pertencem a P-4, nesse caso 60% das amostras foram classificadas corretamente. No último poço PPG5, as amostras foram 100% agrupadas de forma correta.

Os resultados da Análise Inter-Poço são apresentados na Tabela 24 e na Figura 35. A Tabela 24 contém as amostras e petrofácies encontradas por cada grupo, onde TS são as amostras e PF as respectivas petrofácies para cada grupo. As técnicas relacionaram, para cada grupo, amostras associadas a uma petrofácies. A Figura 35 mostra a distribuição dos

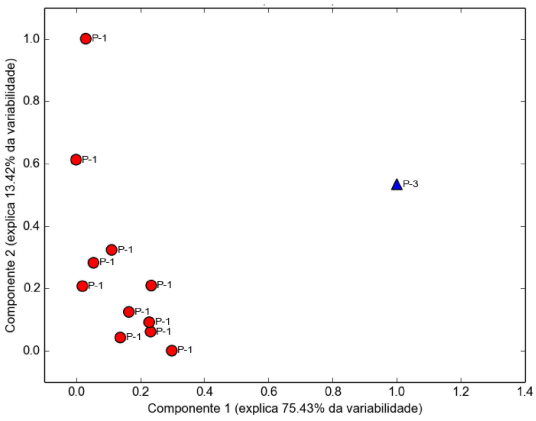
Figura 34 - Visualização da análise Intra-Poço dos poços PPG1, PPG2, PPG3, PPG4 e PPG5, respectivamente.



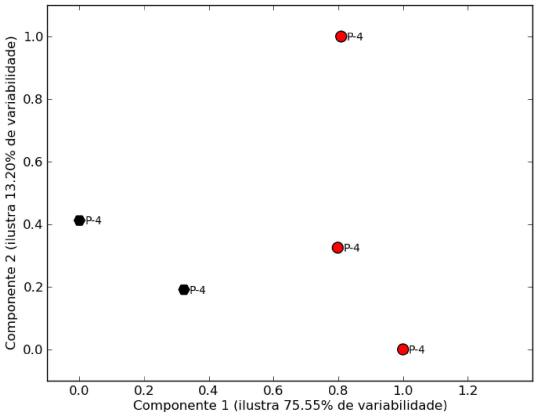

(a) PPG1 - $SC = 0.456$



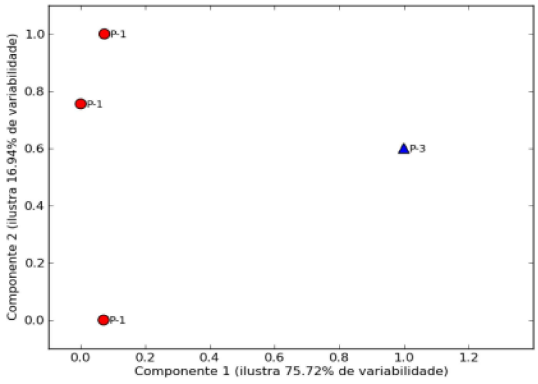

(b) PPG2 - $SC = 0.586$




(c) PPG3 - $SC = 0.657$




(d) PPG4 - $SC = 0.426$




(e) PPG5 - $SC = 0.395$

Fonte: Elaborada pelo autor (2020).

Tabela 23 – Critério de validação encontrado pelo K-Means em cada poço.

| Poços | Nº de CLUSTERS | SC |
|---|---|---|
| PPG1 | 2 | 0.455913 |
| PPG2 | 2 | 0.585717 |
| PPG3 | 2 | 0.656816 |
| PPG4 | 2 | 0.426477 |
| PPG5 | 2 | 0.395879 |

Fonte: Elaborada pelo autor (2020).

grupos encontrados na Análise Inter-Poço e sua proximidade. A primeira figura mostra como as petrofácies foram distribuídas pelo método convencional. A segunda representa a forma em que as petrofácies foram separadas pelo K-Means.

O grupo GPPG0-1 aparece isolado, contendo somente amostras da petrofácies I-1. De acordo com a classificação manual e a análise Intra-Poço a petrofácies P-2 foi atribuída ao grupo PPG2-1 e P-1 ao GPPG5-1, mas segundo a Análise Inter-Poço esses dois grupos possuem atributos semelhantes. Como ocorreu na Análise Intra-Poço, as petrofácies P-4 foram atribuídas aos grupos GPPG4-0 e GPPG4-1. Os resultados indicam que o grupo GPPG1-1 compartilha recursos com os grupos GPPG2-0 e GPPG3-0, bem como com os grupos GPPG3-1 e GPPG5-0, indicando que em cada conjunto as amostras compartilham os atributos similares.

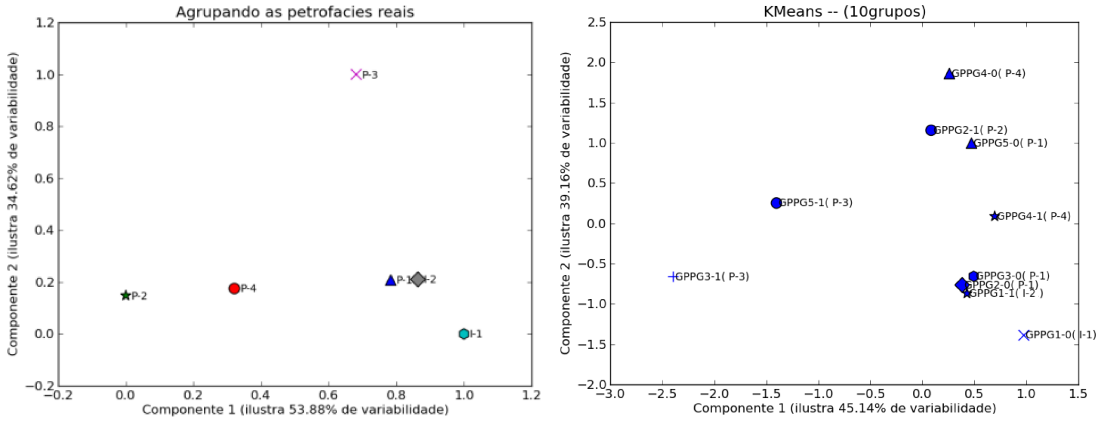
Tabela 24 – Grupos obtidos pelo K-Means (Análise Inter-Poço).

| GRUPOS | | K-MEANS | | (139) | |
|---|---|---|---|---|---|
| | | TS | PF | TS | PF |
| GPPG1-0 | PPG1 | 2 a 8 | I-1 | 2 a 8 | I-1 |
| GPPG1-1 | PPG1 | 1, 9 a 12 | I-2 | 1, 9 a 12 | I-2 |
| GPPG2-0 | PPG2 | 1 | P-2 | 1 | P-2 |
| GPPG2-1 | PPG2 | 2 a 11 | P-1 | 2 a 11 | P-1 |
| GPPG3-0 | PPG3 | 1 a 10, 12 | P-1 | 1 a 10, 12 | P-1 |
| GPPG3-1 | PPG3 | 11 | P-3 | 11 | P-3 |
| GPPG4-0 | PPG4 | 2 a 4 | P-4 | 1 a 5 | P-4 |
| GPPG4-1 | PPG4 | 1, 5 | P-4 | – | – |
| GPPG5-0 | PPG5 | 1, 2 e 4 | P-1 | 1, 2 e 4 | P-1 |
| GPPG5-1 | PPG5 | 3 | P-3 | 3 | P-3 |

Fonte: Elaborada pelo autor (2020).

A Figura 36 mostra o resultado da Análise Filogenética nas amostras, onde pode-se observar que as amostras pertencentes à petrofácies I-1 são separadas em um subgrupo, o que indica que tiveram processo de formação semelhante. Amostras pertencentes à P-3 são colocadas juntas em um subgrupo, na análise Intra-Poço as duas amostras foram separadas, o que pode ser uma indicação de que essas amostras podem pertencer a P-1 e

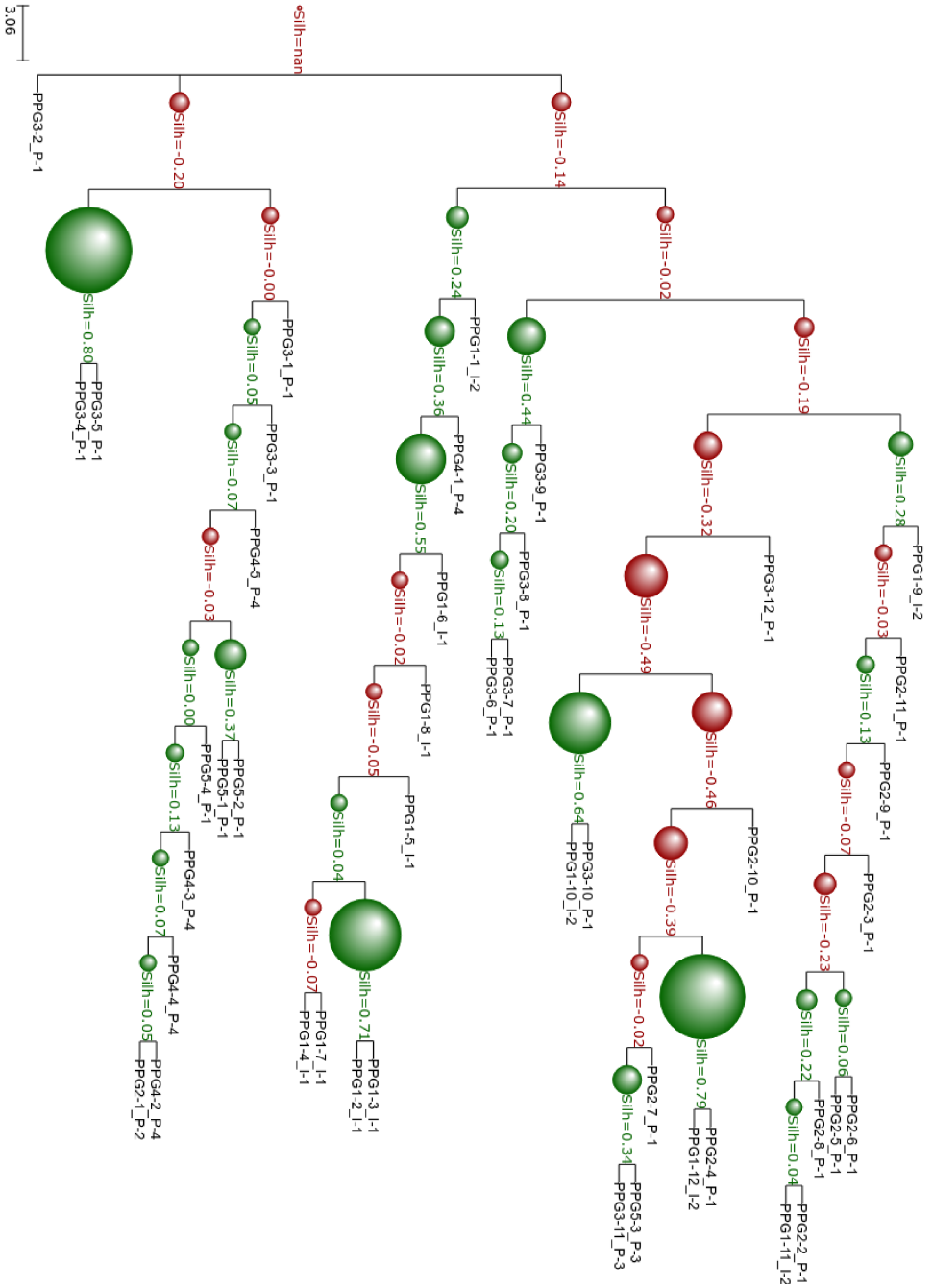
Figura 35 - Visualização da distribuição dos grupos encontrados na Análise Inter-Poço -Tibagi




Fonte: Elaborada pelo autor (2020).

que o processo diagenético sofrido por estas amostras foram semelhantes. As amostras pertencentes a I-2 estão entre as amostras de petrofácies P-1, na análise Inter-Poço pode-se observar que o(s) grupo(s) da amostra das petrofácies I-2 estão próximos do(s) grupo(s) de P-1, a mesma suposição feita anteriormente pode ser empregada. Em relação a P-4, o mesmo que acontece na análise Intra-Poço e Inter-Poço aparece aqui, duas amostras de P-4 se separam das restantes, isso deve ao fato do valor da Ilita/Smectita serem mais baixos, comparados às outras amostras da mesma petrofácies, o que faz com que elas se assemelhem a P-1. A amostra de P-2 possui o valor de Ilita/Smectita alto, sendo semelhante as três amostras de P-4 que possuem esse mesmo comportamento.

Figura 36 - Visualização do resultado da Análise Filogenética Amostras - Tibagi.




Fonte: Elaborada pelo autor (2020).

### 5.3.2 **Dados Petrográficos - Tibagi (Constituintes Diagenéticos)**

A Análise de Componentes Principais foi aplicada para avaliar a influência das propriedades sobre as componentes principais. Na Tabela 25 encontram-se as componentes obtidas. Para as duas componentes as características Pseudo Matriz, Ilita/Smectita e Cimento Carbonático influenciam mais. A análise dessas características auxiliam na identificação de petrofácies sedimentares indicando processos diagenéticos que ocorreram

nas amostras.

Tabela 25 – Componentes Principais em relação as propriedades petrográficas. Os valores com * indicam os valores mais significativos obtidos para cada componente sugerindo qual propriedade petrográfica tem mais atuação sobre as mesmas - Tibagi (Constituintes Diagenéticos)

| Características | Componente 1 | Componente 2 |
|---|---|---|
| Glauconita | 0.00802184 | 0.00693684 |
| Clorita | -0.00183977 | 0.00179907 |
| Pseudo Matriz | -0.3883322* | 0.38587951* |
| Cres. Sec. Quartzo | 0.02895468 | -0.02242587 |
| Caolinita | -0.09269756 | -0.41542663 |
| Ilita/Smectita | -0.47530446* | 0.55824798* |
| Pirita | -0.13666041 | 0.18301389 |
| Siderita | -0.05377517 | 0.0425419 |
| Cim. Carbonatico | 0.76816832* | 0.51146258* |
| Cim. Silicoso | -0.00153176 | 0.00244933 |
| Feldspato | -0.0062245 | 0.00769282 |
| Por. intergranular | -0.04514711 | -0.26279268 |
| Por. intragranular | 0.00742082 | -0.01730876 |

Fonte: Elaborada pelo autor (2020).

A Tabela 26 apresenta os resultados da análise de variância intergrupos para cada característica. Pode-se observar que $H_0$ não é rejeitada para todas as características, o que indica que a média de algumas são iguais para os grupos encontrados. O $H_{cal}$ é o valor calculado pelo método de *Kruskal-Wallis*, nos caso em que p-valor $< 5\%$, $H_{cal}$ é maior que o valor tabelado na Distribuição Qui-Quadrado.

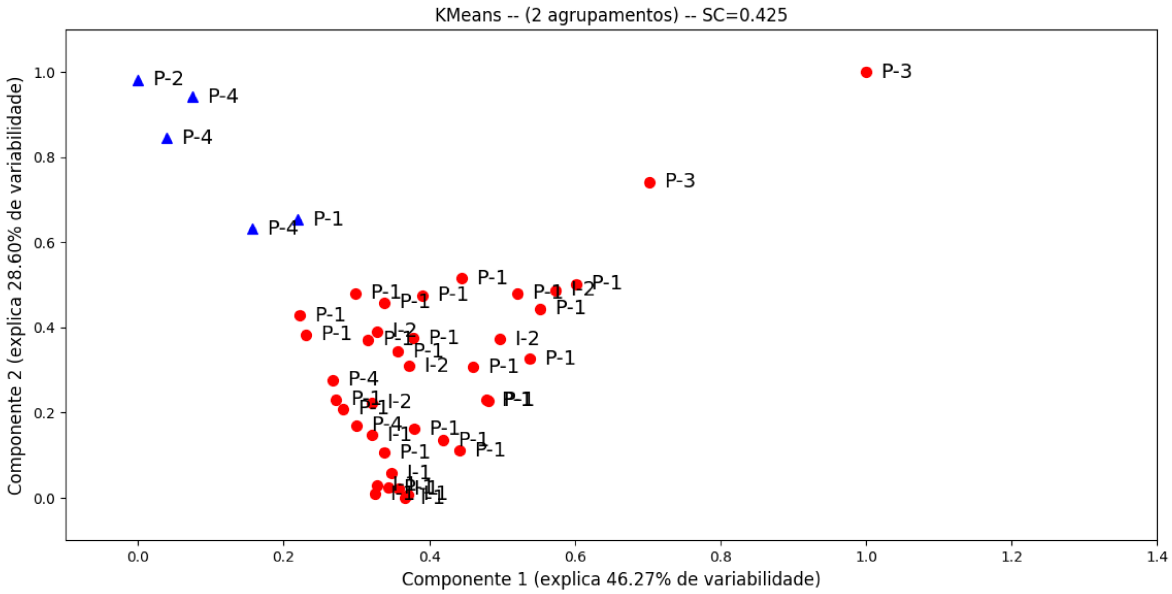
Tabela 26 – Análise de variância intergrupos para cada característica - Tibagi (Constituintes Diagenéticos). Os valores com * indicam os p-valores que rejeitaram a $H_0$. Os grupos foram determinados pelo procedimento computacional. O nível de significância é de 0.05.

| Características | $H_{cal}$ | p-valor |
|---|---|---|
| Cim. Carbonático | 2.718945 | 0.099164* |
| Clorita | 3.214073 | 0.073008 |
| Feldspato | 10.069543 | 0.001507* |
| Glauconita | 0.372076 | 0.541875 |
| Ilita/Smectita | 12.551746 | 0.000396* |
| Caolinita | 2.776667 | 0.095647 |
| Por. intragranular | 2.224996 | 0.135793 |
| Por. intergranular | 0.111208 | 0.738773 |
| Pseudo Matriz | 12.429474 | 0.000423* |
| Pirita | 3.817682 | 0.050714 |
| Cres. Sec. Quartzo | 0.420909 | 0.516484 |
| Siderita | 2.037710 | 0.153441 |
| Cim. Silicoso | 7.800000 | 0.005225* |

Fonte: Elaborada pelo autor (2020).

A Figura 37 exibe o resultado do K-Means para o Tibagi. A primeira componente principal expressa 46.27% de variabilidade dos dados, a segunda expressa 28.60% da variabilidade. O método identificou dois grupos somente, o que indica que utilizando apenas os constituintes diagenéticos o método observou maior semelhança entre as amostras.

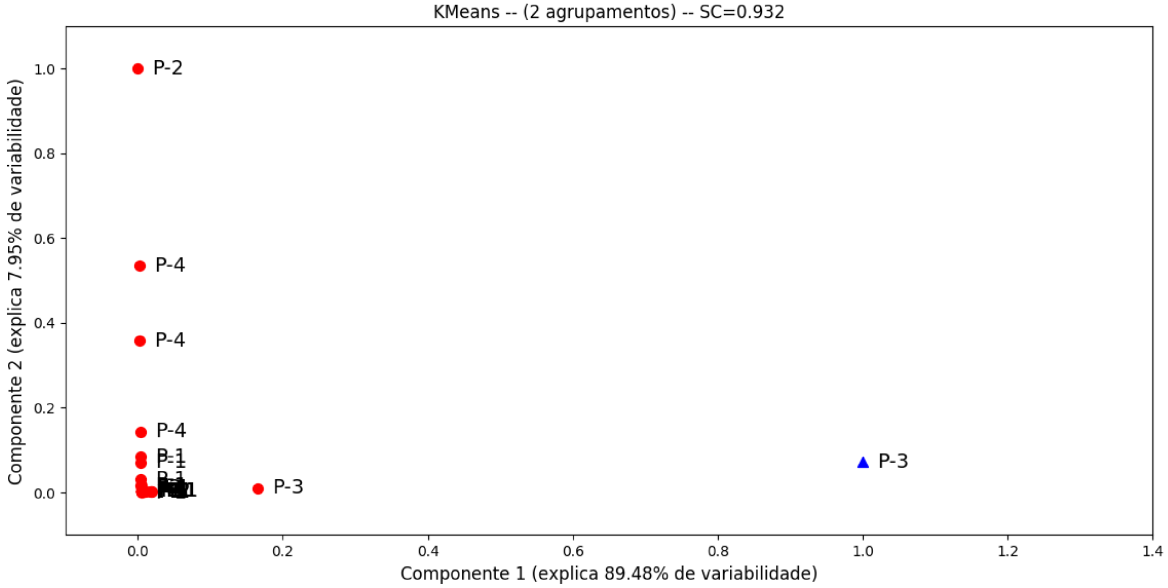
Figura 37 - Resultado K-Means - Tibagi (Constituintes Diagenéticos). SC = 0.425




Fonte: Elaborada pelo autor (2020).

A Figura 38 exibe o resultado do K-Means aplicado nos dados após o pré-processamento o acréscimo das características polinomiais. O grau do polinômio foi 4 e os dados ficaram com 2380 características. A primeira componente principal expressa 89.48% de variabilidade dos dados, a segunda expressa 7.95% da variabilidade. Pode-se observar que houve um aumento no valor do coeficiente de silhueta em relação aos dados originais. Apesar de ter identificado somente dois grupos, a visualização mostrou que as amostras estão mais espaçadas o que pode auxiliar o geólogo na tomada de decisões.

Figura 38 - Resultado K-Means (Características Polinomiais) - Tibagi (Constituintes Diagenéticos). SC = 0.932

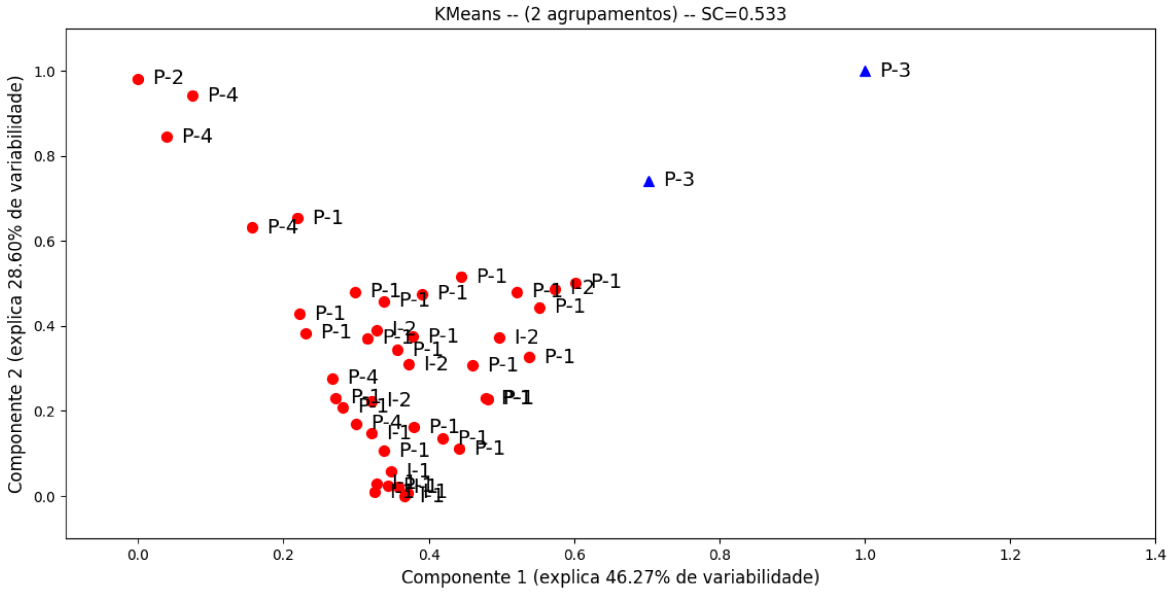



Fonte: Elaborada pelo autor (2020).

O K-Means aplicado nos dados originais após o Bootstrap ter sido empregado para encontrar os parâmetros. A Figura 39 exibe o resultado. Nota-se que foi semelhante ao gerado pelo K-Means quando utilizado nos dados originais, uma vez que os parâmetros sugeridos pelo Bootstrap foram similares. Dois grupos foram identificados, porém as amostras referente a P-3 estão em um único grupo e as amostras pertencentes a P-4 aparecem mais afastadas apesar de estar em um grupo com amostras de outra petrofácies.

Nessa análise foram utilizados apenas constituintes diagenéticos uma vez que os eventos são marcados por eles. Segundo (139) na eodiagênese ocorreu a precipitação de siderita em forma de nódulos e a precipitação de pirita; na mesodiagênese houve o crescimento secundário de quartzo, ilitização de micas e caolinitas e cimentação por calcita; na telodiagênese ocorreu a geração de porosidade secundária e precipitação de caolinita. Como apresentado na Figura 40, o método propõe que na eodiagênese ocorreu a precipitação de siderita em forma de nódulos, a precipitação de pirita e ilitização de micas e caolinitas; na mesodiagênese houve o crescimento secundário de quartzo e cimentação por

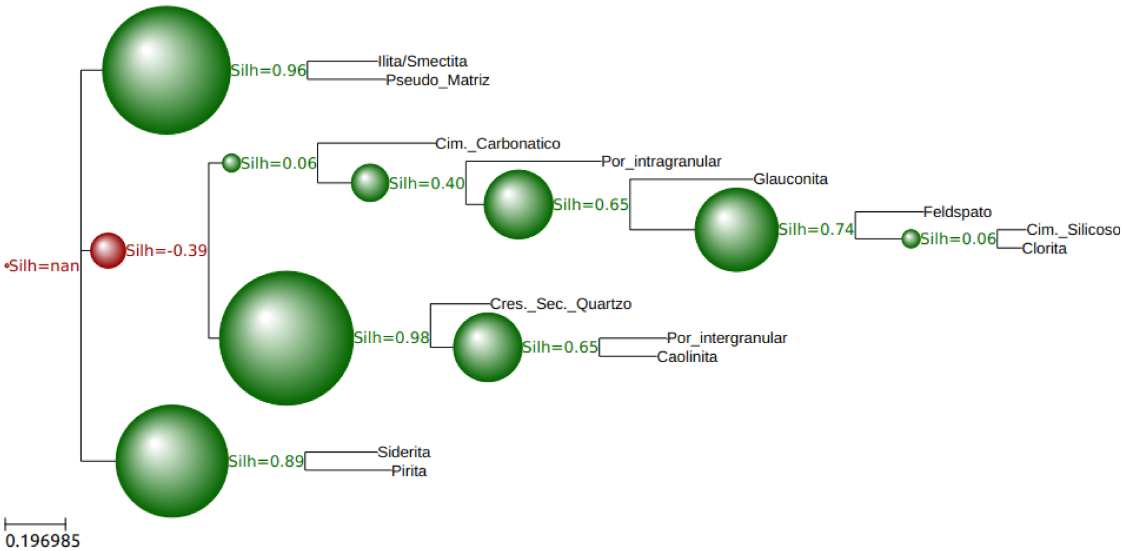
Figura 39 - Resultado K-Means (Bootstrap) - Tibagi (Constituintes Diagenéticos). SC = 0.533




Fonte: Elaborada pelo autor (2020).

calcita; sugere uma mesodiagênese tardia, onde ocorreu a geração de porosidade secundária e a precipitação de caolinita.

Figura 40 - Visualização do resultado da Análise Filogenética Constituintes - Tibagi.




Fonte: Elaborada pelo autor (2020).

### 5.3.3 **Dados Petrográficos - Tibagi (Sem Petrofácies I-1 e I-2)**

A Análise de Componentes Principais foi aplicada para avaliar a influência das propriedades sobre as componentes principais. Na Tabela 27 encontram-se as componentes obtidas. Para a Componente 1 as características Bioclasto, Caolinita, Porosidade Intergranular, Opaco e Turmalina influenciam mais, na Componente 2 são as características Bioclasto, Porosidade Intergranular, Muscovita, Opaco, Rutilo e Pseudo Matriz. Dentre essas características a Caolinita, Porosidade Intergranular e Pseudo Matriz são referentes aos processos diagenéticos identificados nas lâmina que auxiliam na identificação de petrofácies sedimentares.

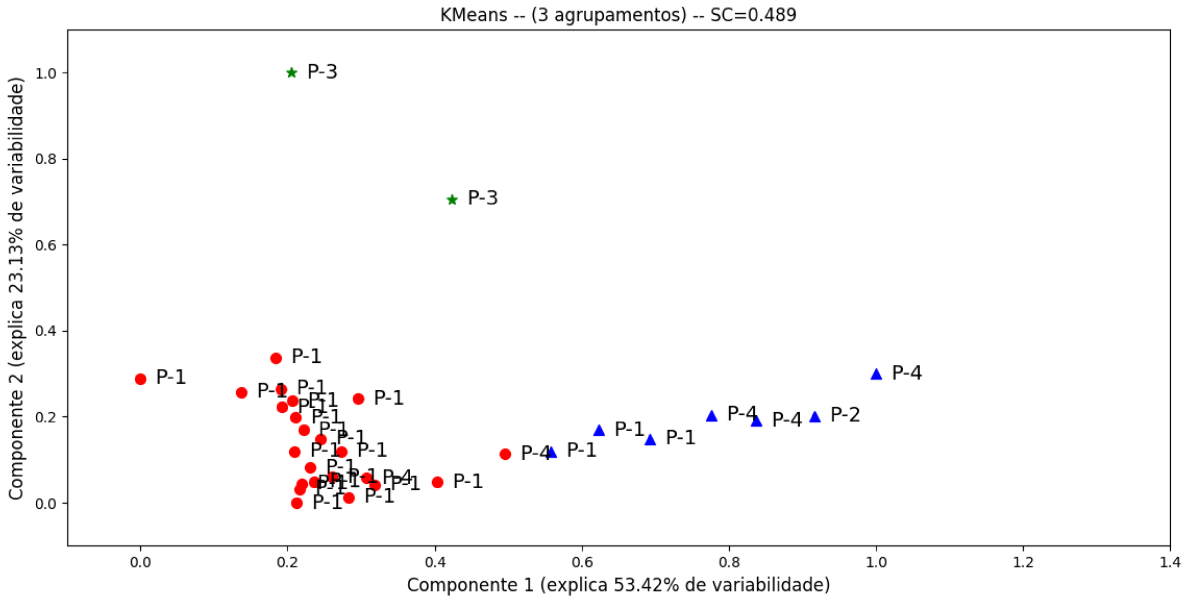
Tabela 27 – Componentes Principais em relação as propriedades petrográficas. Os valores com * indicam os valores mais significativos obtidos para cada componente sugerindo qual propriedade petrográfica tem mais atuação sobre as mesmas - Tibagi (Sem Petrofácies I-1 e I-2)

| Características | Componente 1 | Componente 2 |
|---|---|---|
| Bioclasto | -7.72906550e-01* | -3.56667806e-01* |
| Cresc. Sec. Qtz | 1.54503174e-02 | -2.07044794e-02 |
| Caolinita | 1.45141349e-01 | -1.43613633e-01 |
| Ilita/Smectita | -3.34189938e-03 | -1.26473513e-02 |
| Pirita | -4.71358729e-03 | -4.28115866e-03 |
| Siderita | 6.09032024e-03 | -4.45394969e-03 |
| Cim. Carbonático | 1.41724701e-03 | 6.13274798e-03 |
| Cim. Silicoso | -2.69679361e-03 | 1.00283490e-02 |
| Cim. Ferruginoso | 1.47861627e-03 | -9.69644739e-04 |
| Por. Intergranular | 3.93806914e-01* | 2.86123148e-02 |
| Por. Intragranular | -0.00000000e+00 | 0.00000000e+00 |
| Quartzo | -3.70923420e-03 | -5.14714384e-03 |
| Feldspato | -1.60188994e-02 | 2.29917112e-02 |
| Muscovita | 4.13803493e-04 | -1.42478590e-01 |
| Opaco | 3.99870819e-01* | -1.61626443e-01 |
| Turmalina | 1.28332282e-01 | -1.21081473e-02 |
| Zircão | 1.16346124e-02 | -6.27086321e-02 |
| Rutilo | -2.21889464e-01* | 8.92736209e-01* |
| Glauconita | 1.62557314e-03 | -2.86128817e-04 |
| Clorita | 6.51274582e-03 | -1.85814635e-04 |
| Pseudo Matriz | 7.73061039e-03 | -5.03921337e-02 |
| Litoclasto | -5.78017844e-03 | -1.96236676e-03 |

Fonte: Elaborada pelo autor (2020).

A Figura 41 exibe o resultado do K-Means para o Tibagi (Sem Petrofácies I-1 e I-2). A primeira componente principal expressa 53.42% de variabilidade dos dados, a segunda expressa 23.13% da variabilidade. Três grupos foram identificados pelo método. Um deles contém somente amostras de P-3, os demais grupos possui amostras de diferentes petrofácies. No grupo marcado de vermelho a maioria das amostras de P-1 ficaram concentradas.

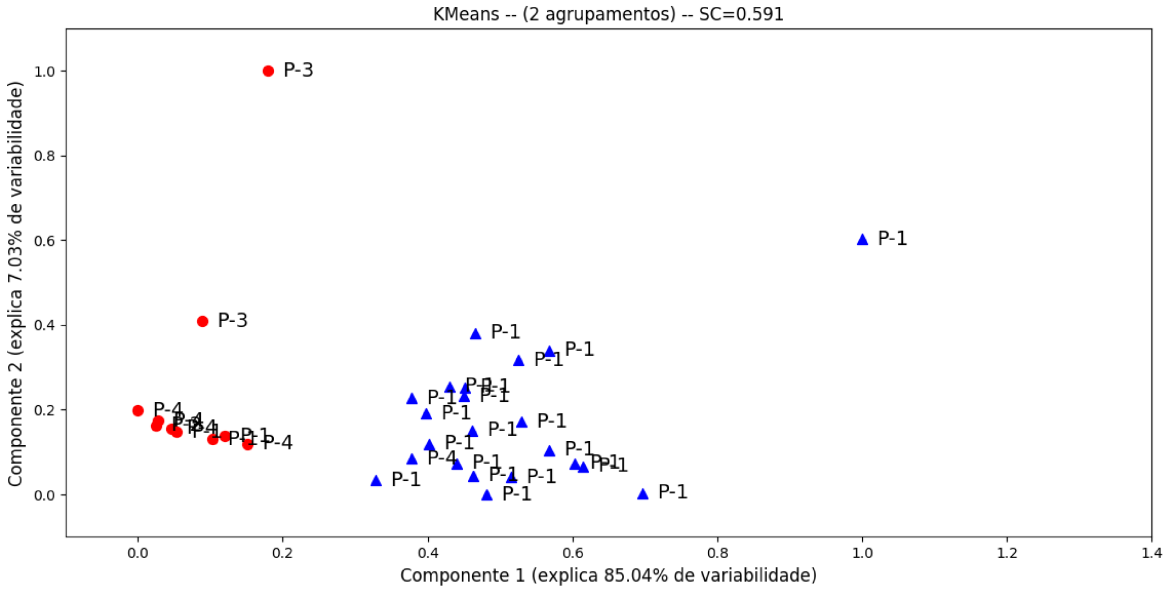
Figura 41 - Resultado K-Means - Tibagi (Sem Petrofácies I-1 e I-2). SC = 0.489




Fonte: Elaborada pelo autor (2020).

A Figura 42 exibe o resultado do K-Means aplicado nos dados após o pré-processamento o acréscimo das características polinomiais. O grau do polinômio foi 4 e os dados ficaram com 14950 características. A primeira componente principal expressa 85.04% de variabilidade dos dados, a segunda expressa 7.03% da variabilidade. Pode-se observar que houve um aumento no valor do coeficiente de silhueta em relação aos dados originais. O método separou as amostras em dois grupos somente, mas as amostras ficaram mais espaçadas o que pode auxiliar na tomada de decisão do especialista.

Figura 42 - Resultado K-Means (Características Polinomiais) - Tibagi (Sem Petrofácies I-1 e I-2). SC = 0.591



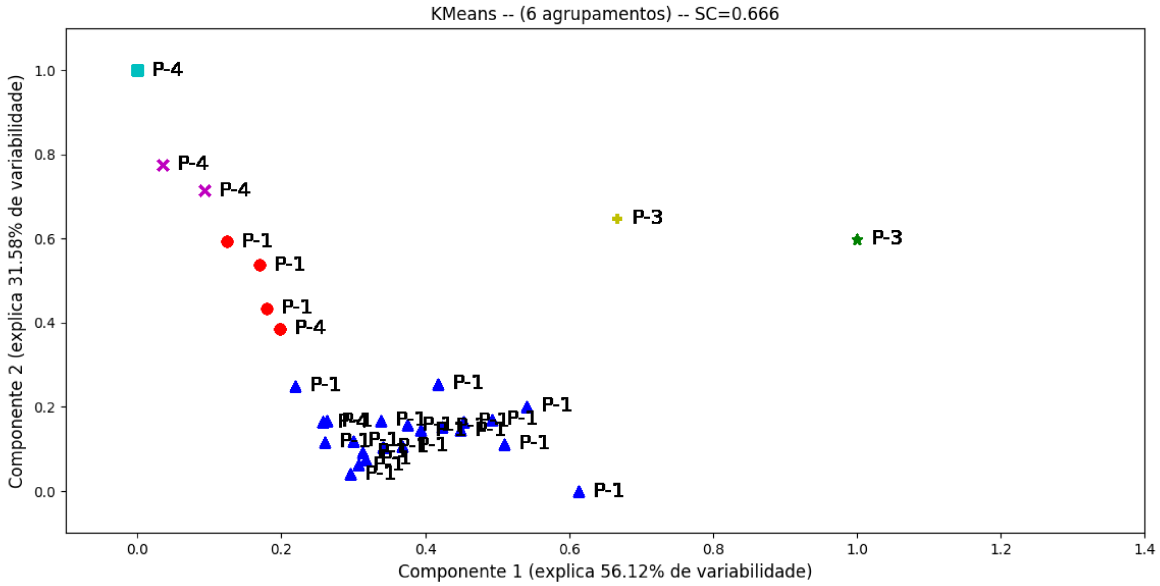

Fonte: Elaborada pelo autor (2020).

O K-Means aplicado nos dados originais após o Bootstrap ter sido empregado para encontrar os parâmetros. A Figura 43 exibe o resultado. Foram gerados seis grupos, as amostras pertencentes a P-3 foram agrupadas em um grupo, a maioria das amostras de P-1 ficaram em um mesmo grupo. As amostras pertencentes a P-4 foram separadas em quatro grupos, mostrando que o método não encontrou similaridade entre essas amostras.

O resultado da Análise Intra-Poço pode ser visto na Tabela 28 e na Figura 44. O número de agrupamentos e o valor do coeficiente de silhueta encontra-se na Tabela 28. A Figura 44 ilustra os agrupamentos encontrados para os poços PPG2, PPG3, PPG4 e PPG5, respectivamente. Nos poços PPG2 e PPG3 o K-Means gerou 2 agrupamentos, concordando com a classificação de (139). No PPG4, o método encontrou 2 agrupamentos, diferindo da classificação manual, uma vez que ela propôs a existência de uma única petrofácies (P-4). Duas amostras segundo o K-Means (PPG4-1 e PPG4-5) não pertencem a P-4, nesse caso 60% das amostras foram classificadas corretamente. No último poço PPG5, as amostras foram 100% agrupadas de forma correta. Nota-se que o resultado é o mesmo encontrado no Tibagi, uma vez que excluiu-se o poço PPG-1 que continha as petrofácies I-1 e I-2.

Os resultados da Análise Inter-Poço são apresentados na Tabela 29 e na Figura 45. A Tabela 29 contém as amostras e petrofácies encontradas por cada grupo, onde TS são as amostras e PF as respectivas petrofácies para cada grupo. As técnicas relacionaram, para cada grupo, amostras associadas a uma petrofácies. A Figura 45 mostra a distribuição dos grupos encontrados na Análise Inter-Poço e sua proximidade. A primeira figura mostra

Figura 43 - Resultado K-Means (Bootstrap) - Tibagi (Sem Petrofácies I-1 e I-2). SC = 0.666




Fonte: Elaborada pelo autor (2020).

Tabela 28 – Critério de validação encontrado pelo K-Means em cada poço - Sem Petrofácies I-1 e I-2.

| Poços | Nº de CLUSTERS | SC |
|---|---|---|
| PPG2 | 2 | 0.585717 |
| PPG3 | 2 | 0.656816 |
| PPG4 | 2 | 0.426477 |
| PPG5 | 2 | 0.395879 |

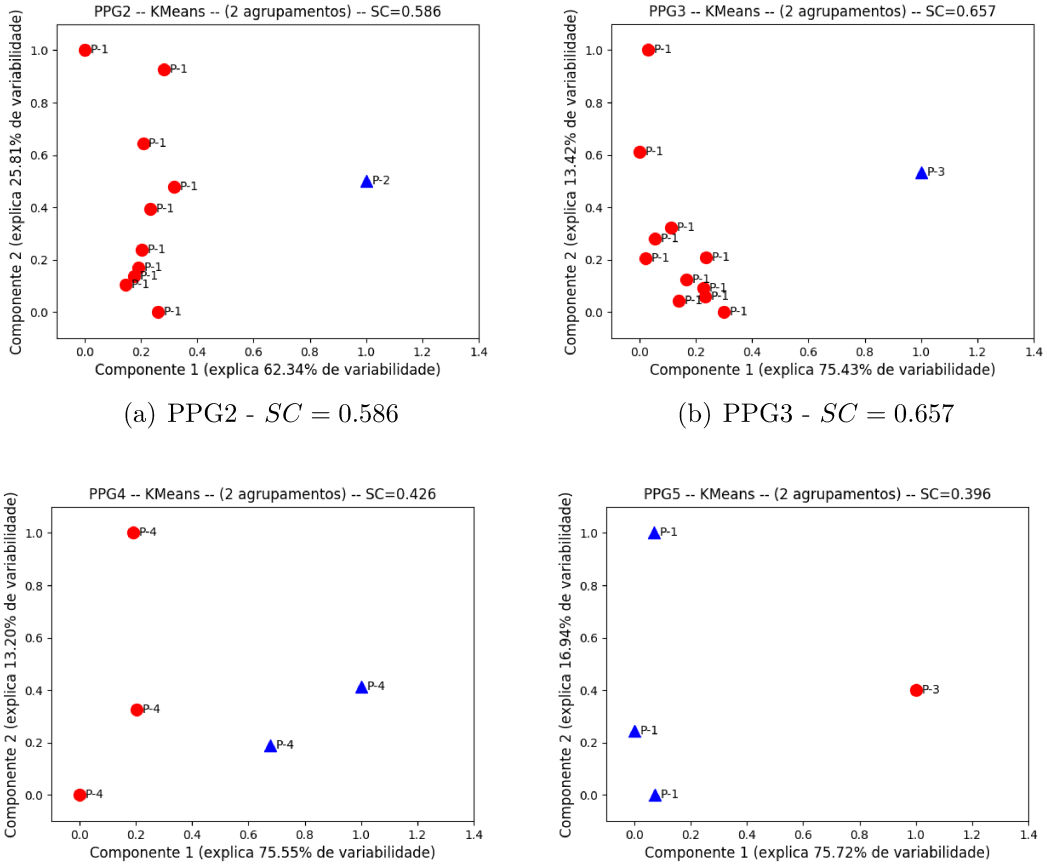
Fonte: Elaborada pelo autor (2020).

como as petrofácies foram distribuídas pelo método convencional. A segunda representa a forma em que as petrofácies foram separadas pelo K-Means.

De acordo com a classificação manual e a análise Intra-Poço a petrofácies P-2 foi atribuída ao grupo PPG2-1 e P-1 ao GPPG5-1, mas segundo a Análise Inter-Poço esses dois grupos possuem atributos semelhantes com o grupo GPPG4-1 (petrofácies P-4). Essa questão foi discutida anteriormente. Como ocorreu na Análise Intra-Poço, as petrofácies P-4 foram atribuídas aos grupos GPPG4-0 e GPPG4-1. Os resultados indicam que o grupo GPPG1-1 compartilha recursos com os grupos GPPG2-0 e GPPG3-0, bem como com os grupos GPPG3-1 e GPPG5-0, indicando que em cada conjunto as amostras compartilham os atributos similares.

A Figura 46 mostra o resultado da Análise Filogenética nas amostras, Amostras pertencentes à P-3 são colocadas juntas em um subgrupo, na análise Intra-Poço as duas

Figura 44 - Visualização da análise Intra-Poço dos poços PPG2, PPG3, PPG4 e PPG5, respectivamente.




(a) PPG2 - $SC = 0.586$ (b) PPG3 - $SC = 0.657$

(c) PPG4 - $SC = 0.426$ (d) PPG5 - $SC = 0.395$

Fonte: Elaborada pelo autor (2020).

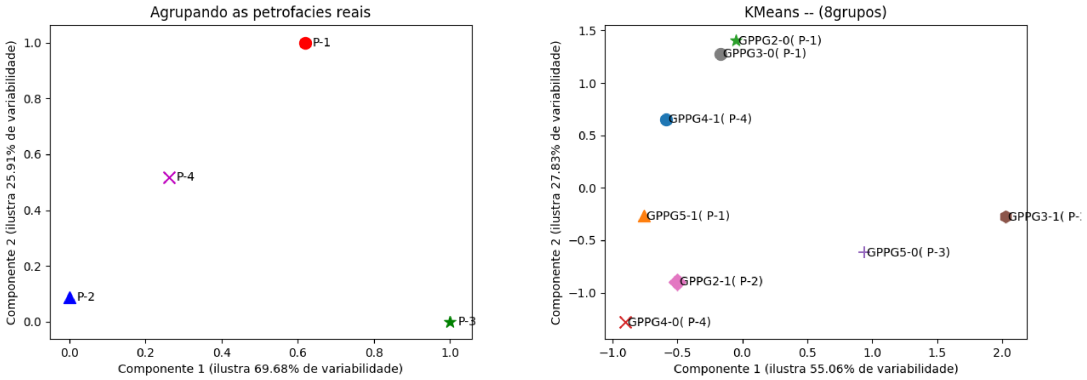
Tabela 29 – Grupos obtidos pelo K-Means (Análise Inter-Poço) - Sem Petrofácies I-1 e I-2.

| GRUPOS | | K-MEANS | | (139) | |
|---|---|---|---|---|---|
| | | TS | PF | TS | PF |
| GPPG2-0 | PPG2 | 2 a 11 | P-1 | 2 a 11 | P-1 |
| GPPG2-1 | PPG2 | 1 | P-2 | 1 | P-2 |
| GPPG3-0 | PPG3 | 1 a 10, 12 | P-1 | 1 a 10, 12 | P-1 |
| GPPG3-1 | PPG3 | 11 | P-3 | 11 | P-3 |
| GPPG4-0 | PPG4 | 2 a 4 | P-4 | 1 a 5 | P-4 |
| GPPG4-1 | PPG4 | 1, 5 | P-4 | – | – |
| GPPG5-0 | PPG5 | 3 | P-3 | 3 | P-3 |
| GPPG5-1 | PPG5 | 1, 2 e 4 | P-1 | 1, 2 e 4 | P-1 |

Fonte: Elaborada pelo autor (2020).

amostras foram separadas, o que pode ser uma indicação de que essas amostras podem pertencer a P-1 e que o processo diagenético sofrido por estas amostras foram semelhantes. Em relação a P-4, o mesmo que acontece na análise Intra-Poço e Inter-Poço aparece aqui,

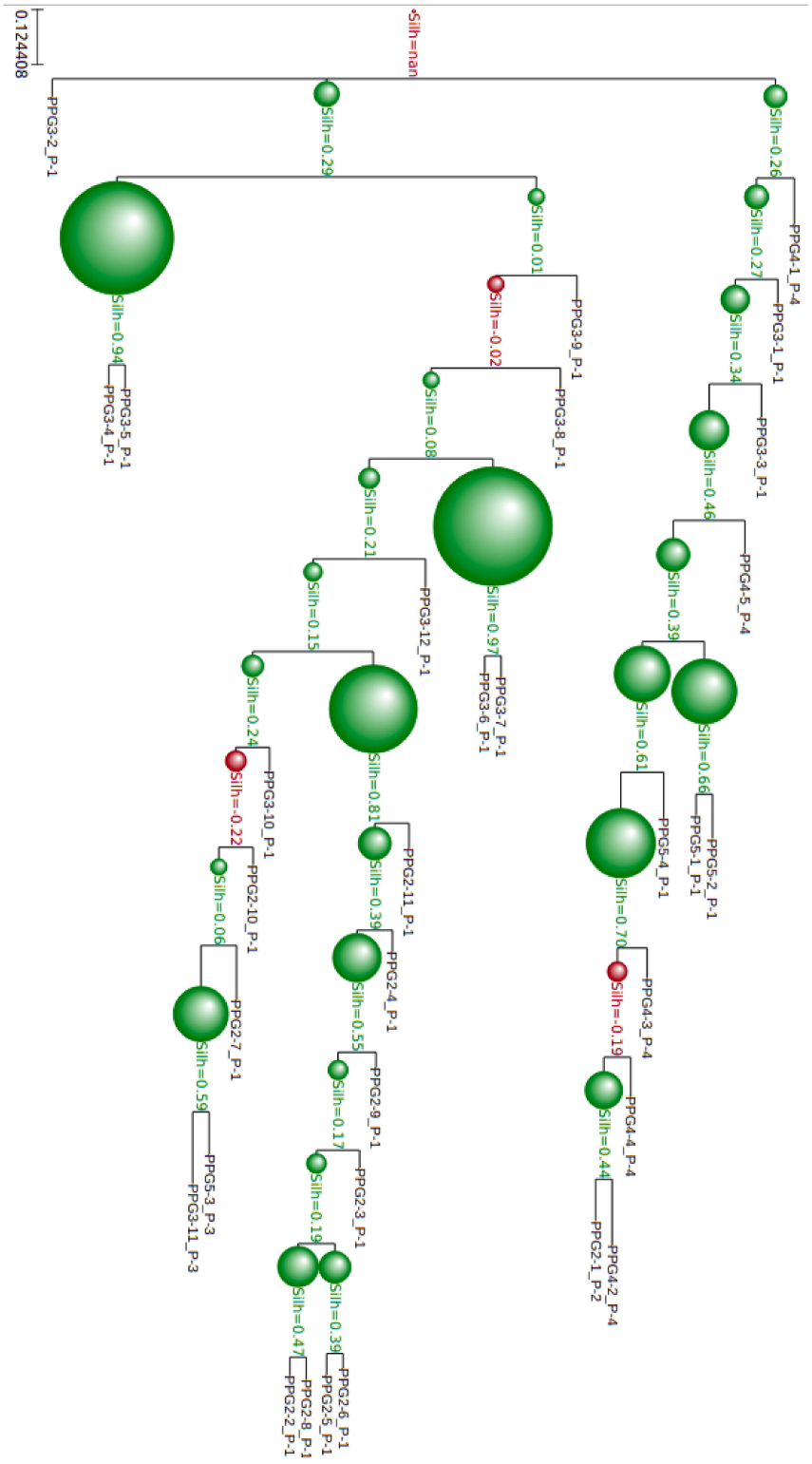
Figura 45 - Visualização da distribuição dos grupos encontrados na Análise Inter-Poço - Tibagi (Sem Petrofácies I-1 e I-2)




Fonte: Elaborada pelo autor (2020).

duas amostras de P-4 se separam das restantes, isso deve ao fato do valor da Ilita/Smectita serem mais baixos, comparados as outras amostras da mesma petrofácies, o que faz com que elas se assemelhem a P-1. A amostra de P-2 possui o valor de Ilita/Smectita alto, sendo semelhante as três amostras de P-4 que possuem esse mesmo comportamento.

Figura 46 - Visualização do resultado da Análise Filogenética Amostras - Tibagi (Sem Petrofácies I-1 e I-2).




Fonte: Elaborada pelo autor (2020).

### 5.3.4 **Dados Petrográficos - Tibagi (Constituintes Diagenéticos - Sem Petrofácies I-1 e I-2)**

A Análise de Componentes Principais foi aplicada para avaliar a influência das propriedades sobre as componentes principais. Na Tabela 30 encontram-se as componentes obtidas. Para as duas componentes as características Pseudo Matriz, Caolinita, Ilita/Smectita e Cimento Carbonático influenciam mais. A análise dessas características auxiliam na identificação de petrofácies sedimentares indicando processos diagenéticos que ocorreram nas amostras.

Tabela 30 – Componentes Principais em relação as propriedades petrográficas. Os valores com * indicam os valores mais significativos obtidos para cada componente sugerindo qual propriedade petrográfica tem mais atuação sobre as mesmas - Tibagi (Constituintes Diagenéticos - Sem Petrofácies I-1 e I-2)

| Características | Componente 1 | Componente 2 |
|---|---|---|
| Glauconita | 0.00839665 | 0.00700345 |
| Clorita | -0.00195939 | 0.00144592 |
| Pseudo Matriz | -0.40523328* | 0.54948775* |
| Cres. Sec. Quartzo | 0.0369755 | -0.04364423 |
| Caolinita | -0.05719655* | -0.40034217* |
| Ilita/Smectita | -0.51468561* | 0.39143103* |
| Pirita | -0.14793256 | 0.17979119 |
| Siderita | -0.05377399 | -0.00485922 |
| Cim. Carbonático | 0.73508847* | 0.57882584* |
| Cim. Silicicoso | -0.00167667 | 0.0025833 |
| Feldspato | -0.00673341 | 0.00852404 |
| Por. Intergranular | -0.03108194 | -0.12193514 |
| Por. Intragranular | 0.0051732 | -0.01672755 |

Fonte: Elaborada pelo autor (2020).

A Tabela 31 apresenta os resultados da análise de variância intergrupos para cada característica. Pode-se observar que $H_0$ não é rejeitada para todos as características, o que indica que a média de algumas são iguais para os grupos encontrados. O $H_{cal}$ é o valor calculado pelo método de *Kruskal-Wallis*, nos caso em que p-valor $< 5\%$, $H_{cal}$ é maior que o valor tabelado na Distribuição Qui-Quadrado.

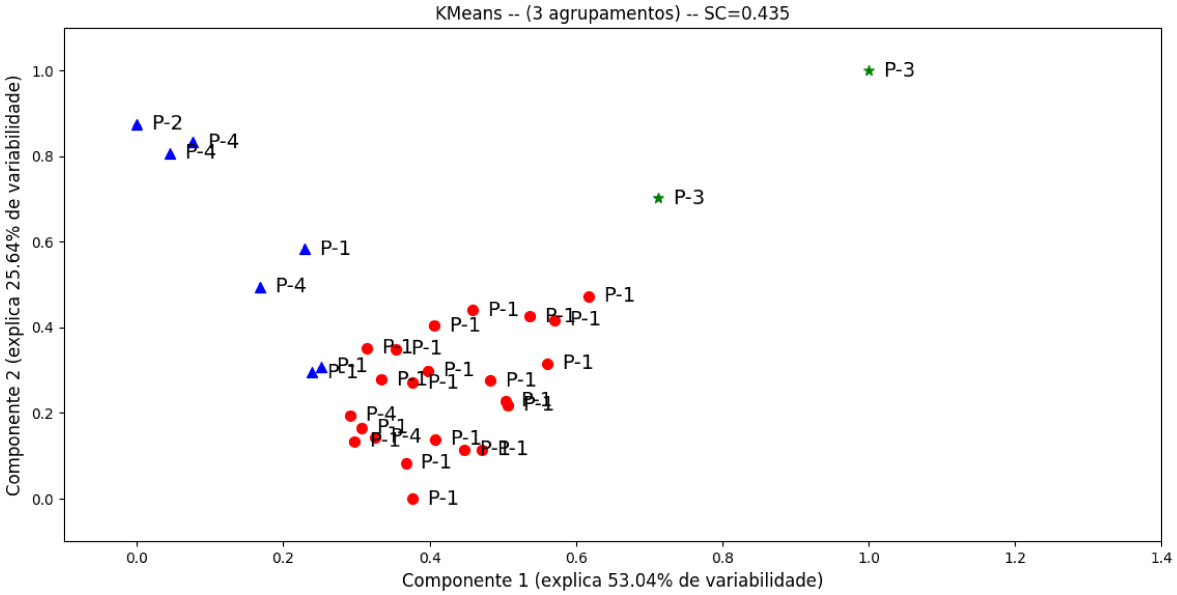
Tabela 31 – Análise de variância intergrupos para cada característica - Tibagi (Constituintes Diagenéticos - Sem Petrofácies I-1 e I-2). Os valores com * indicam os p-valores que rejeitaram a $H_0$. Os grupos foram determinados pelo procedimento computacional. O nível de significância é de 0.05.

| Características | $H_{cal}$ | p-valor |
|---|---|---|
| Cim. Carbonático | 11.701804 | 0.002877* |
| Clorita | 1.118614 | 0.571605 |
| Feldspato | 4.146096 | 0.125802 |
| Glauconita | 1.056536 | 0.589625 |
| Ilita/Smectita | 16.037452 | 0.000329* |
| Caolinita | 1.138278 | 0.566013 |
| Por. Intragranular | 2.762804 | 0.251226 |
| Por. Intergranular | 3.816098 | 0.148370 |
| Pseudo Matriz | 13.667047 | 0.001077* |
| Pirita | 1.857130 | 0.395120 |
| Cres. Sec. Quartzo | 0.204434 | 0.902834 |
| Siderita | 3.158005 | 0.206181 |
| Cim. Silicoso | 3.571429 | 0.167677 |

Fonte: Elaborada pelo autor (2020).

A Figura 47 exibe o resultado do K-Means para o Tibagi (Constituintes Diagenéticos - Sem Petrofácies I-1 e I-2). A primeira componente principal expressa 53.04% de variabilidade dos dados, a segunda expressa 25.64% da variabilidade. Três grupos foram encontrados, em um deles as amostras pertencentes a P-3 foram agrupadas. No que está marcado de vermelho há a maior concentração de amostras de P-1.

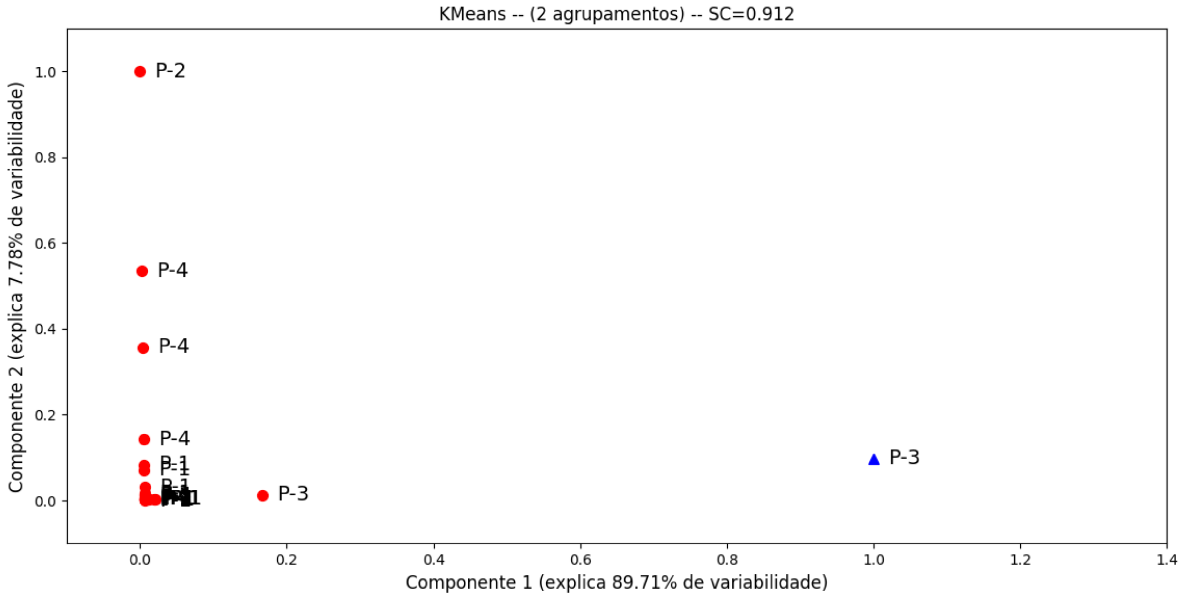
Figura 47 - Resultado K-Means - Tibagi (Constituintes Diagenéticos - Sem Petrofácies I-1 e I-2). SC = 0.435




Fonte: Elaborada pelo autor (2020).

A Figura 48 exibe o resultado do K-Means aplicado nos dados após o pré-processamento o acréscimo das características polinomiais. O grau do polinômio foi 4 e os dados ficaram com 2380 características. A primeira componente principal expressa 89.71% de variabilidade dos dados, a segunda expressa 7.78% da variabilidade. Pode-se observar que houve um aumento no valor do coeficiente de silhueta em relação aos dados originais. Apesar de somente dois grupos terem sido gerados, as amostras estão mais afastadas podendo auxiliar o geólogo na tomada de decisão.

Figura 48 - Resultado K-Means (Características Polinomiais) - Tibagi (Constituintes Diagenéticos - Sem Petrofácies I-1 e I-2). SC = 0.912

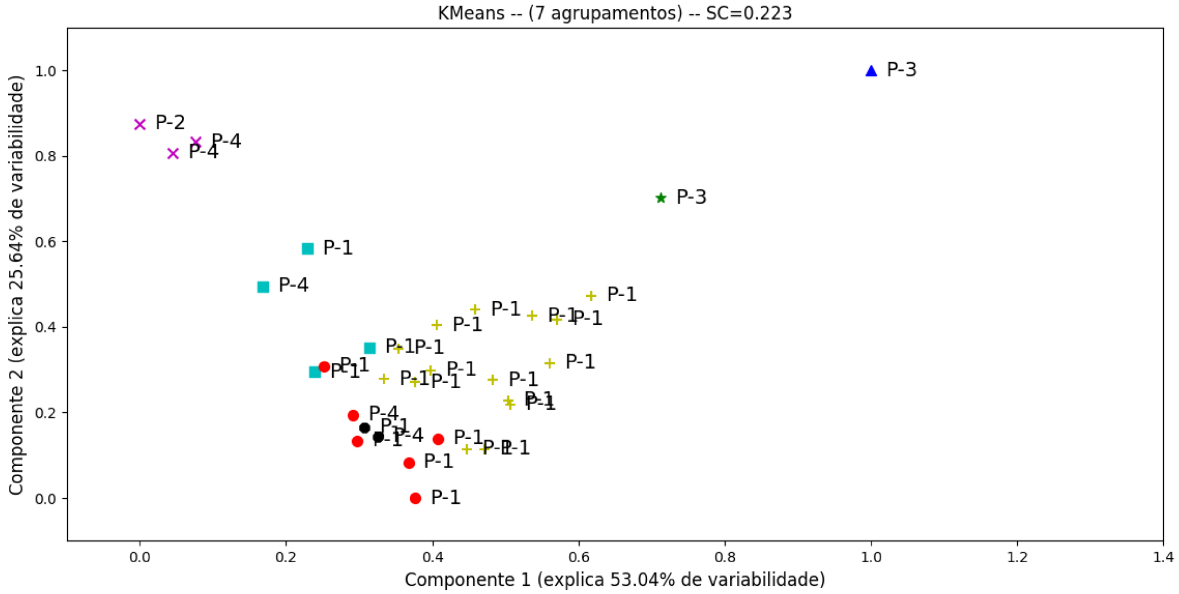



Fonte: Elaborada pelo autor (2020).

O K-Means aplicado nos dados originais após o Bootstrap ter sido empregado para encontrar os parâmetros. A Figura 49 exibe o resultado. O valor do coeficiente de silhueta foi menor e que o número de grupos encontrados aumentou, no entanto amostras de petrofácies distintas estão sendo agrupadas em um mesmo grupo.

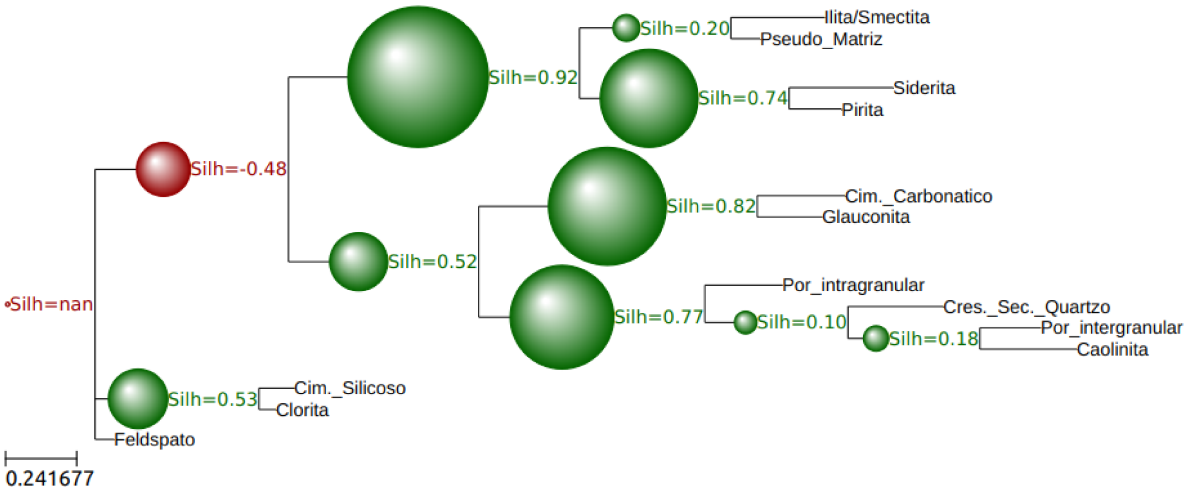
Nessa análise foram utilizados apenas constituintes diagenéticos uma vez que os eventos são marcados por eles. Segundo (139) na eodiagênese ocorreu a precipitação de siderita em forma de nódulos e a precipitação de pirita; na mesodiagênese houve o crescimento secundário de quartzo, ilitização de micas e caolinitas e cimentação por calcita; na telodiagênese ocorreu a geração de porosidade secundária e precipitação de caolinita. Como apresentado na Figura 50, o método propõe que na eodiagênese ocorreu a precipitação de siderita em forma de nódulos, a precipitação de pirita e ilitização de micas e caolinitas; na mesodiagênese houve o crescimento secundário de quartzo e cimentação por calcita; sugere o acontecimeneto de uma mesodiagênese tardia onde há a geração de porosidade secundária e na telodiagênese a precipitação de caolinita.

Figura 49 - Resultado K-Means (Bootstrap) - Tibagi (Constituintes Diagenéticos - Sem Petrofácies I-1 e I-2). SC = 0.223




Fonte: Elaborada pelo autor (2020).

Figura 50 - Visualização do resultado da Análise Filogenética Constituintes - Tibagi (Sem Petrofácies I-1 e I-2).



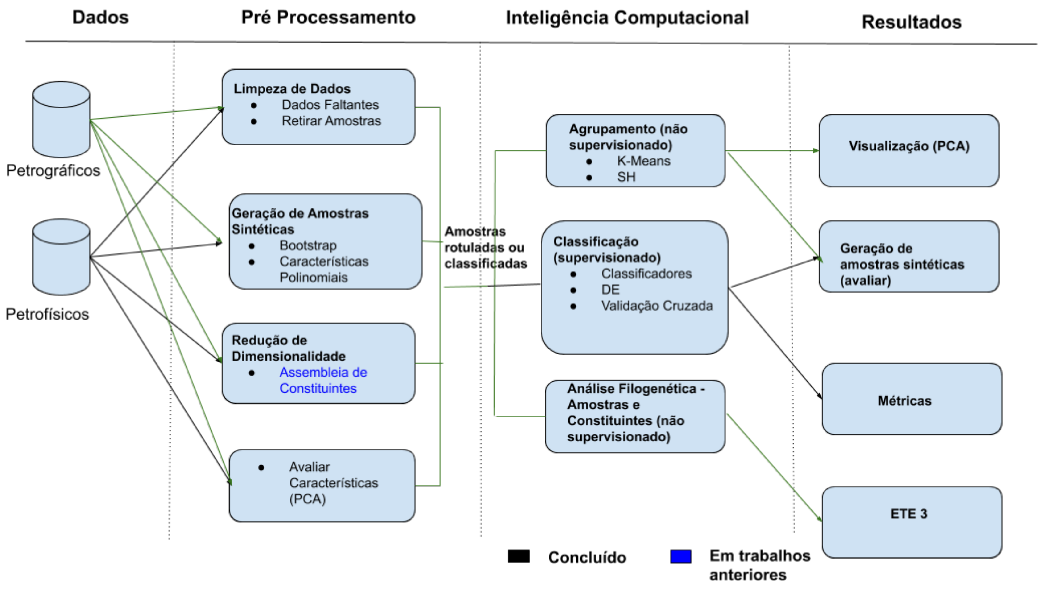

Fonte: Elaborada pelo autor (2020).

### 5.3.5 **Dados Petrográficos - Paleosul**

A Figura 51 mostra, através das linhas verdes, quais procedimentos da metodologia descritos nas Seções 5.1 e 5.2 foram aplicados as bases de dados Paleosul e Paleosul-Constituintes Diagenéticos.

Para Paleosul a Análise Filogenética foi aplicada da mesma maneira descrita para Tibagi. Os constituintes diagenéticos também foram utilizados para identificar os eventos ocorridos.

Figura 51 - Fluxograma ilustrando, através das linhas verdes, a metodologia aplicada a base de dados Paleosul.




Fonte: Elaborada pelo autor (2020).

Na Tabela 32 encontram-se as componentes obtidas. Para a Componente 1 as características Argilomineral não-identificado, Minerais Pesados, Mica, Porosidade Intergranular e Pirita influenciam mais, na Componente 2 são as características Ooide Berthierina, Minerais Pesados, Mica e Pirita.

A Tabela 33 apresenta os resultados da análise de variância intergrupos para cada característica. Pode-se observar que $H_0$ não é rejeitada para todas as características, o que indica que a média de algumas são iguais para os grupos encontrados. O $H_{cal}$ é o valor calculado pelo método de *Kruskal-Wallis*, nos caso em que p-valor $< 5\%$, $H_{cal}$ é maior que o valor tabelado na Distribuição Qui-Quadrado.

A Figura 52 exibe o resultado do K-Means para o Paleosul. A primeira componente principal expressa 36.69% de variabilidade dos dados, a segunda expressa 23.74% da variabilidade.

Tabela 32 – Componentes Principais em relação as propriedades petrográficas. Os valores com * indicam os valores mais significativos obtidos para cada componente sugerindo qual propriedade petrográfica tem mais atuação sobre as mesmas - Paleosul

| Características | Componente 1 | Componente 2 |
|---|---|---|
| Argilomineral não-identificado | -5.15542883E-01* | -4.21611036E-02 |
| Bioclasto | -1.77491999E-02 | -6.63092584E-04 |
| Por bioturbação | -5.07662884E-02 | 1.37252497E-02 |
| Substituição de grão Berthierina (F, M) | -6.07407454E-03 | 5.24363393E-04 |
| Ooide Berthierina | 3.10108343E-02 | 1.40464522E-01* |
| Carbonato | -3.72400519E-03 | -1.94984644E-03 |
| K-Feldspato Detrítico | 6.10389937E-03 | -8.32036904E-02 |
| Plagioclásio Detrítica | -1.35881165E-02 | -1.62002194E-02 |
| Albitização do Feldspato | 6.54813970E-03 | 1.22891866E-02 |
| Glauconita | 6.43204881E-02 | 4.03020322E-02 |
| Ooide Goetita | 7.65903538E-02 | -3.97180117E-02 |
| Minerais Pesados | 2.89043159E-01* | 7.99944689E-01* |
| Caolinita | 2.27807904E-03 | 4.48413005E-02 |
| Mica | 7.20507626E-01* | -4.84904640E-01* |
| Moldic | 6.87158706E-02 | -2.17496246E-03 |
| Outros (Ti-óxidos, super cres. F) | -9.48135083E-03 | -1.12466640E-02 |
| Oversized | -9.47384074E-02 | -4.35890991E-02 |
| Por. Intragranular | -1.21575853E-01* | -2.46949033E-02 |
| Por. Intergranular | -1.25331478E-02 | 7.15489177E-04 |
| Pirita | -2.90164018E-01* | -2.96365360E-01* |
| Qz Monocristalino | -1.36441801E-02 | 8.63471135E-04 |
| Super Cresc. Qtz | -4.93835712E-02 | 9.05016948E-04 |
| Qz Policristalino | -4.24551025E-03 | 2.39796664E-03 |
| Siderita | -2.69163965E-02 | -1.46869139E-02 |
| Nódulos de Siderita | -3.27569794E-02 | 3.06240248E-03 |

Fonte: Elaborada pelo autor (2020).

A Figura 53 exibe o resultado do K-Means aplicado nos dados após o pré-processamento o acréscimo das características polinomiais. O grau do polinômio foi 4 e os dados ficaram com 23751 características. A primeira componente principal expressa 65.59% de variabilidade dos dados, a segunda expressa 20.04% da variabilidade. Pode-se observar que houve um aumento no valor do coeficiente de silhueta em relação aos dados originais.

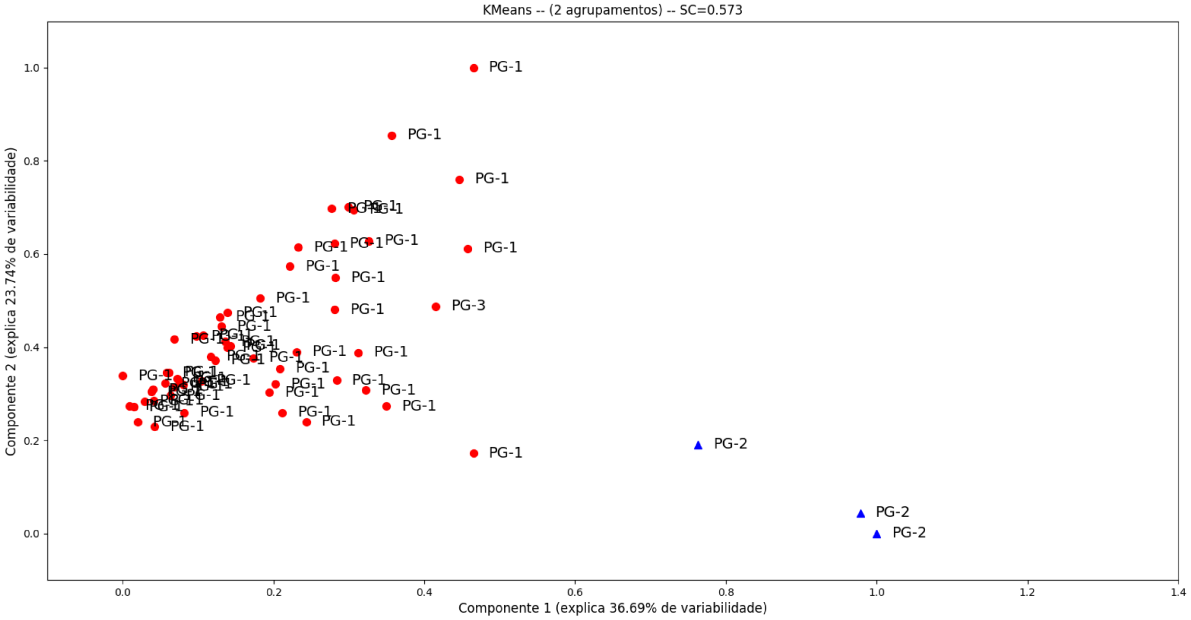
Tabela 33 – Análise de variância intergrupos para cada característica - Paleosul. Os valores com * indicam os p-valores que rejeitaram a $H_0$. Os grupos foram determinados pelo procedimento computacional. O nível de significância é de 0.05.

| Características | $H_{cal}$ | p-valor |
|---|---|---|
| Argilomineral não-identificado | 1.007524 | 0.315497 |
| Bioclasto | 0.535955 | 0.464114 |
| Por bioturbação | 0.923428 | 0.336576 |
| Substituição de grão Berthierina (F, M) | 1.533197 | 0.215633 |
| Ooide Berthierina | 11.481396 | 0.000703* |
| Carbonato | 1.773387 | 0.182963 |
| K-Feldspato Detrítico | 7.850530 | 0.005081* |
| Plagioclásio Detrítica | 3.141515 | 0.076323 |
| Albitização do Feldspato | 6.516361 | 0.010689* |
| Glauconita | 2.268122 | 0.132060 |
| Ooide Goetita | 0.052632 | 0.818546 |
| Minerais Pesados | 1.694480 | 0.193011 |
| Caolinita | 4.905281 | 0.026775* |
| Mica | 6.514356 | 0.010701* |
| Moldic | 3.423073 | 0.064291 |
| Outros (Ti-óxidos, super cres. F) | 5.145905 | 0.023301* |
| Oversized | 1.184193 | 0.276504 |
| Por. Intergranular | 1.577311 | 0.209149 |
| Por. Intergranular | 3.934599 | 0.047302* |
| Pirita | 3.946604 | 0.046966 |
| Qz Monocristalino | 7.833726 | 0.005128* |
| Super Cresc. Qtz | 7.189741 | 0.007332 |
| Qz Policristalino | 3.764928 | 0.052338 |
| Siderita | 0.543292 | 0.461071 |
| Nódulos de Siderita | 8.523515 | 0.003506* |

Fonte: Elaborada pelo autor (2020).

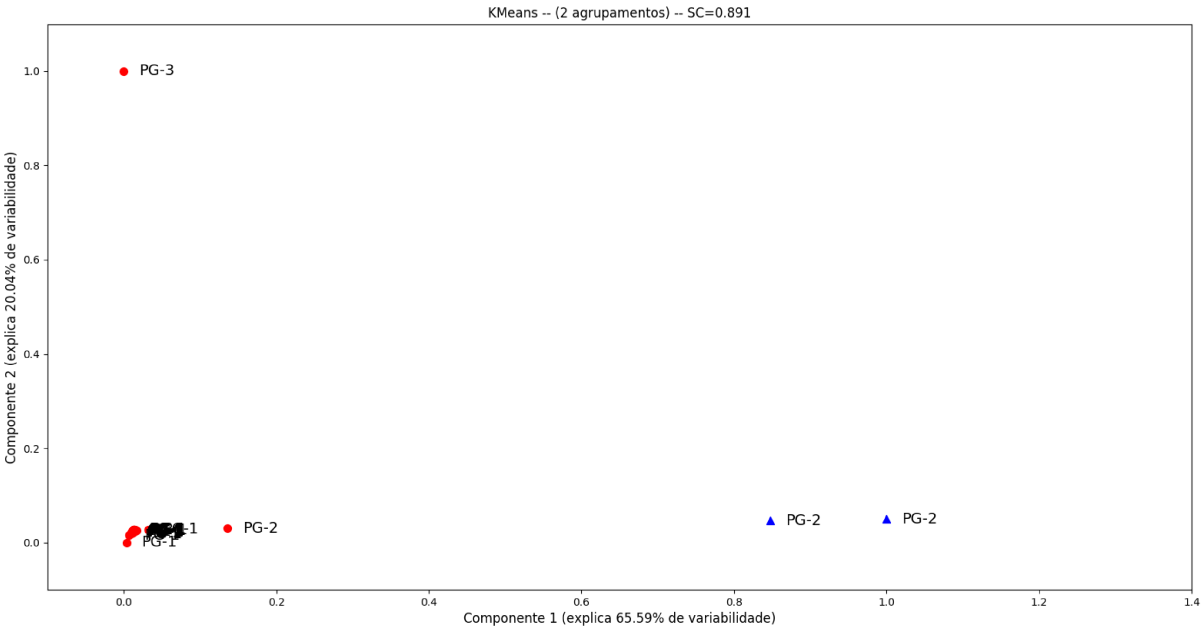
A Figura 54 exibe o resultado do K-Means aplicado nos dados originais após o Bootstrap ter sido empregado para encontrar os parâmetros. O resultado foi inferior ao gerado pelo K-Means utilizado nos dados originais.

Figura 52 - Resultado K-Means - Paleosul. SC = 0.573



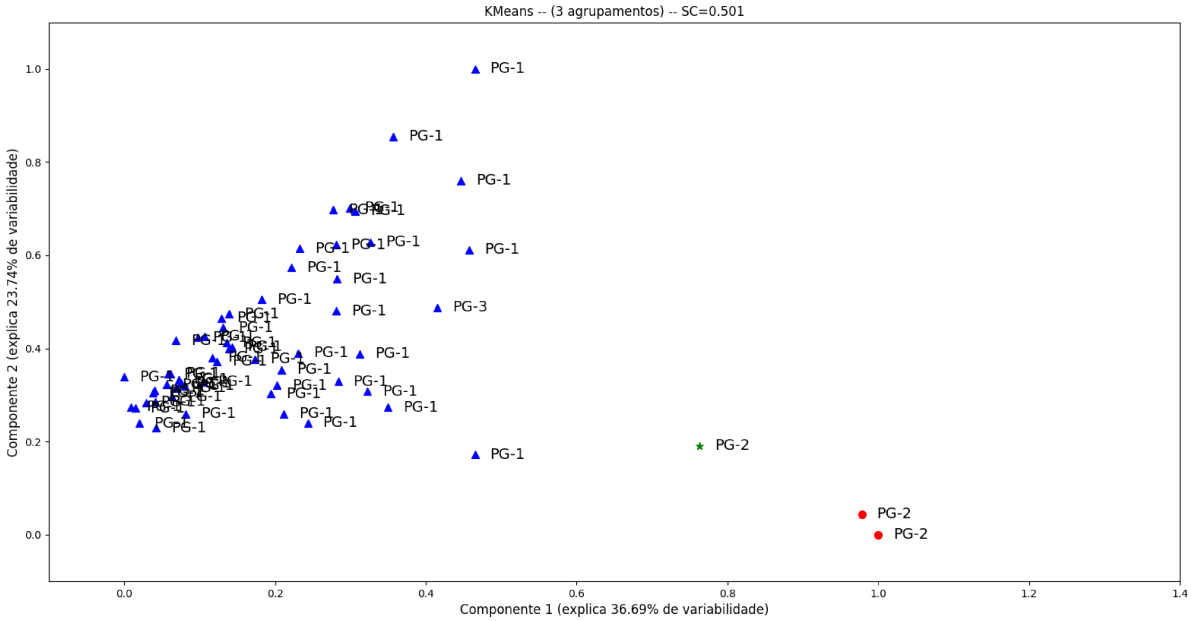

Fonte: Elaborada pelo autor (2020).

Figura 53 - Resultado K-Means (Características Polinomiais) - Paleosul. SC = 0.891




Fonte: Elaborada pelo autor (2020).

Figura 54 - Resultado K-Means (Bootstrap) - Paleosul. SC = 0.501




Fonte: Elaborada pelo autor (2020).

Com intuito de comparar o resultado obtido através da Análise Filogenética, dois procedimentos (Análise Intra-Poço e Inter-Poço), que já foram realizados anteriormente para a base da dados Tibagi, foram empregados para Paleosul.

O resultado da Análise Intra-Poço pode ser visto na Tabela 34 e na Figura 55. O número de agrupamentos e o valor do coeficiente de silhueta encontra-se na Tabela 34. A Figura 55 ilustra os agrupamentos encontrados para os poços RPL, RSP e RVR, respectivamente. Para o poço RPL o método gerou 2 agrupamentos, coincidindo com o método manual nas classificação das classes PG-1 e PG-2, agrupando 100% das amostras de forma correta. No poço RSP o K-Means gerou 3 agrupamentos, concordando com a classificação de (41). No RVR, o método encontrou 3 agrupamentos, diferindo da classificação manual, uma vez que ela propôs a existência de uma única petrofácies (PG-1).

Tabela 34 – Critério de validação encontrado pelo K-Means em cada poço - Paleosul.

| Poços | $N^o$ de CLUSTERS | SC |
|---|---|---|
| RPL | 2 | 0.510521 |
| RSP | 3 | 0.448842 |
| RVR | 3 | 0.291781 |

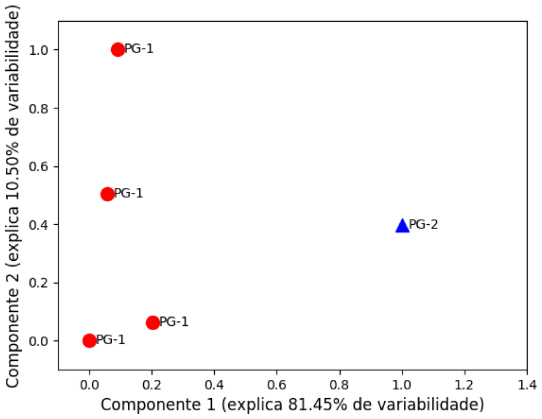
Fonte: Elaborada pelo autor (2020).

Os resultados da Análise Inter-Poço são apresentados na Tabela 35 e na Figura 56. A Tabela 35 contém as amostras e petrofácies encontradas por cada grupo, onde TS são as amostras e PF as respectivas petrofácies para cada grupo. As técnicas relacionaram, para cada grupo, amostras associadas a uma petrofácies. A Figura 56 mostra a distribuição dos grupos encontrados na Análise Inter-Poço e sua proximidade. A primeira figura mostra como as petrofácies foram distribuídas pelo método convencional. A segunda representa a forma em que as petrofácies foram separadas pelo K-Means.

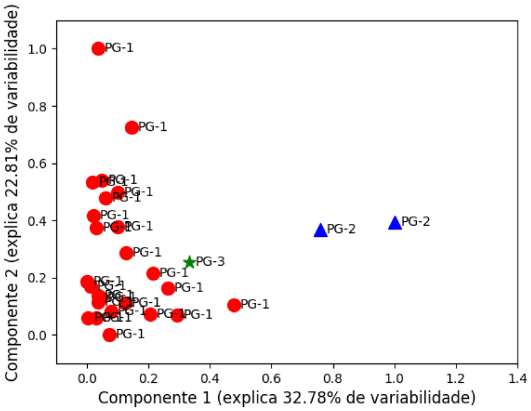
O grupo GRSP-2 aparece isolado, contendo somente amostras da petrofácies PG-3. Os resultados indicam que o grupo GRPL-0 compartilha recursos com os grupos GRSP-0, GRVR-0,GRVR-1 e GRVR-2, bem como com os grupos GRPL-1 e GRSP-1, indicando que em cada conjunto as amostras compartilham os atributos similares.

A Figura 57 mostra o resultado da Análise Filogenética, onde pode-se observar que as amostras pertencente à petrofácies PG-1 aparecem em dois subgrupos maiores e quatro subgrupos menores, o que indica que tiveram processo de formação semelhante. A amostras pertencente à PG-3 está um subgrupo isolada, o que coincide pois há somente uma amostra pertencente a essa petrofácies, esse subgrupo está próximo de amostras pertencentes a PG-1, o que pode ser uma indicação de que essas amostras passaram por processo diagenéticos semelhante ao de PG-3. As três amostras pertencentes a PG-2 estão em um mesmo subgrupo, concordando com os resultados obtidos na Análise Intra-Poço e Inter-Poço.

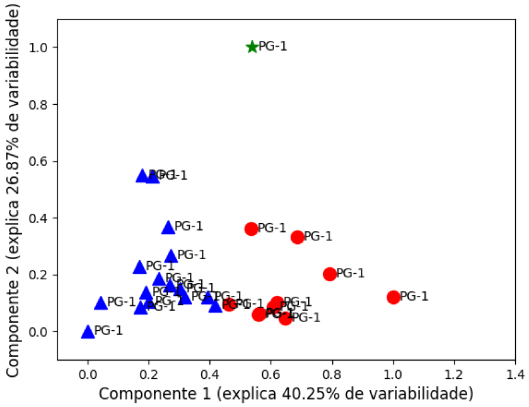
Figura 55 - Visualização da análise Intra-Poço dos poços RPL, RSP e RVR respectivamente.




(a) RPL - $SC = 0.510$




(b) RSP - $SC = 0.448$




(c) RVR - $SC = 0.291$

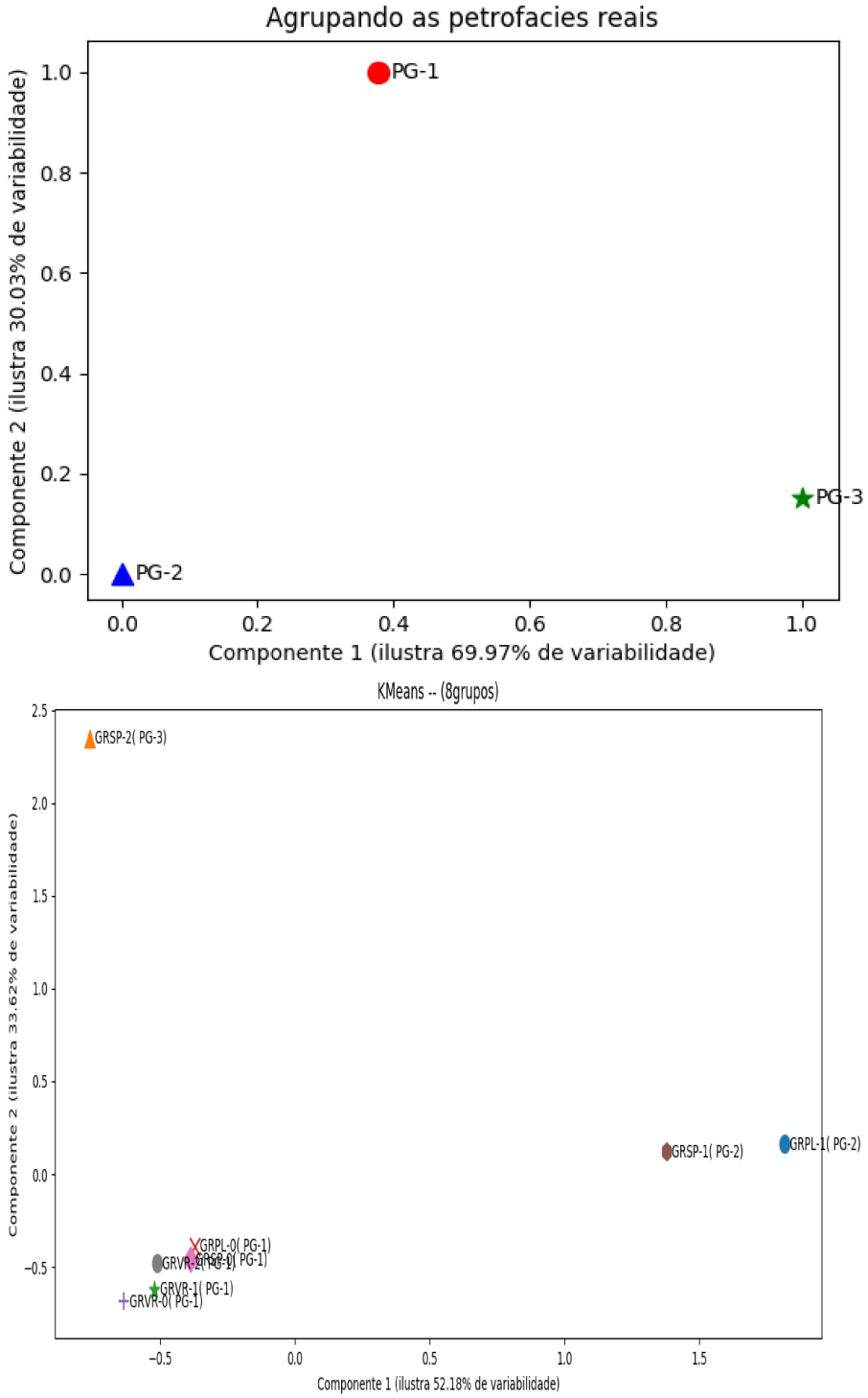
Fonte: Elaborada pelo autor (2020).

Tabela 35 – Grupos obtidos pelo K-Means (Análise Inter-Poço) - Paleosul.

| GRUPOS | | K-MEANS | | (41) | |
|---|---|---|---|---|---|
| | | TS | PF | TS | PF |
| GRLP-0 | RPL | 2 a 5 | PG-1 | 2 a 5 | PG-1 |
| GRPL-1 | RPL | 1 | PG-2 | 1 | PG-2 |
| GRSP-0 | RSP | 2 a 12, 14 a 17 e 19 a 28 | PG-1 | 2 a 12, 14 a 17 e 19 a 28 | PG-1 |
| GRSP-1 | RSP | 13 e 18 | PG-2 | 13 e 18 | PG-2 |
| GRSP-2 | RSP | 1 | PG-3 | 1 | PG-3 |
| GRVR-0 | RVR | 3 a 5,9,12,13,16,22,24,25 | PG-1 | 1 a 27 | PG-1 |
| GRVR-1 | RVR | 1,6 a 8,10,11,14,15,17 a 21,23,26,27 | PG-1 | – | – |
| GRVR-2 | RVR | 2 | PG-1 | – | – |

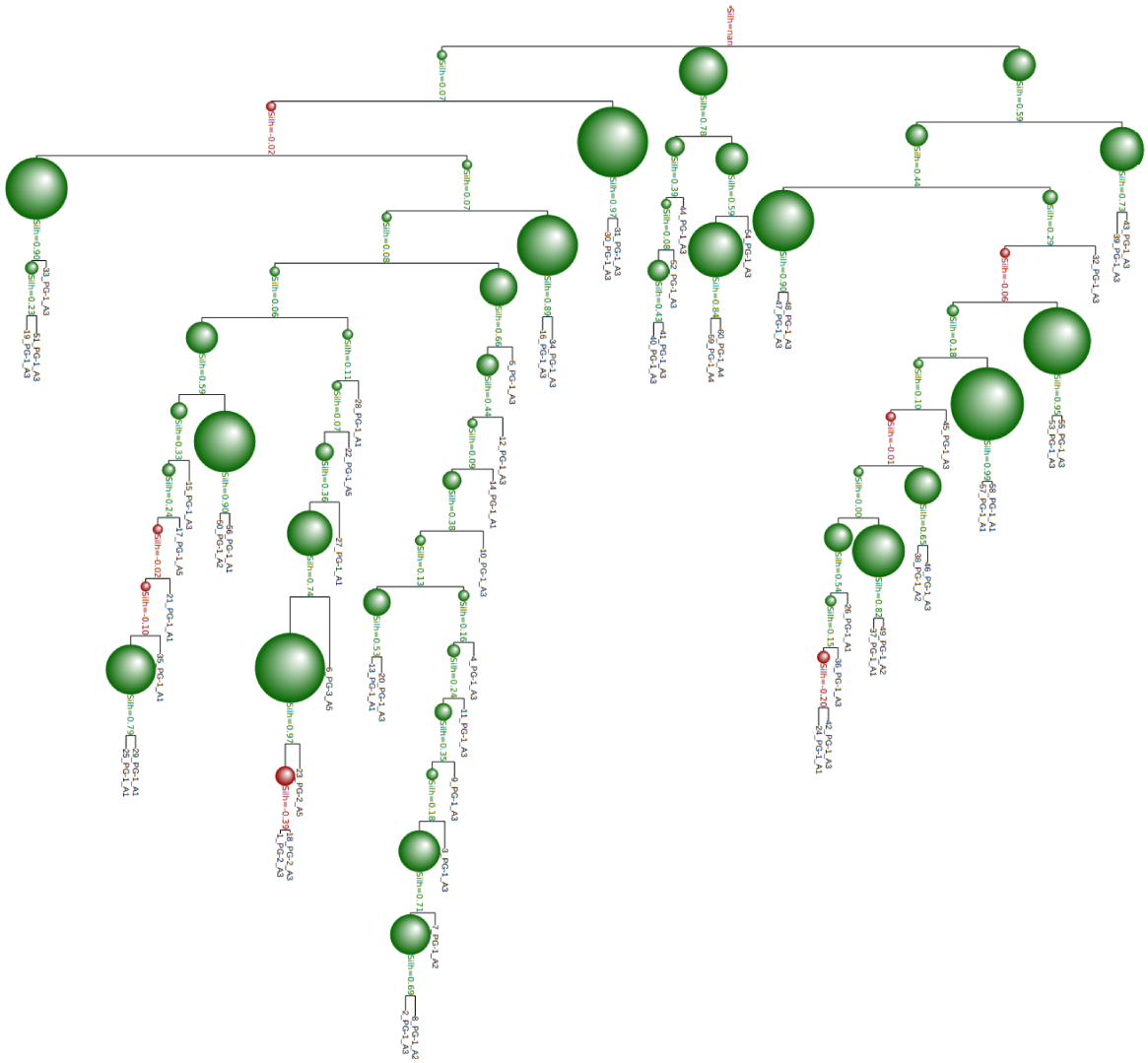
Fonte: Elaborada pelo autor (2020).

Figura 56 - Visualização da distribuição dos grupos encontrados na Análise Inter-Poço e sua proximidade - Paleosul




Fonte: Elaborada pelo autor (2020).

Figura 57 - Visualização do resultado da Análise Filogenética nas Amostras - Paleosul.



Fonte: Elaborada pelo autor (2020).

### 5.3.6 **Dados Petrográficos - Paleosul (Constituintes Diagenéticos)**

Na Tabela 36 encontram-se as componentes obtidas. Para a Componente 1 o Cimento Carbonático apresenta maior influência, enquanto os constituintes Siderita e Crescimento Secundário de Quartzo exercem maior influência na Componente 2.

Tabela 36 – Componentes Principais em relação as propriedades petrográficas. Os valores com * indicam os valores mais significativos obtidos para cada componente sugerindo qual propriedade petrográfica tem mais atuação sobre as mesmas - Paleosul (Constituintes Diagenéticos)

| Características | Componente 1 | Componente 2 |
|---|---|---|
| Glauconita | 1.54209574e-02 | 8.33375735e-04 |
| Siderita | -9.13364621e-02 | 8.04136045e-01* |
| Pirita | 1.51809091e-02 | 4.82276888e-03 |
| Caolinita | 6.90988771e-02 | -3.32308582e-01 |
| Cresc. Sec. Quartzo | 7.88655646e-02 | -4.09767628e-01* |
| Cim. Carbonático | 9.89720271e-01* | 1.23690770e-01 |
| Por. Intergranular | -2.54275884e-02 | -2.43356777e-01 |
| Por. Intragranular | -1.47495148e-03 | -2.19279913e-02 |

Fonte: Elaborada pelo autor (2020).

A Tabela 37 apresenta os resultados da análise de variância intergrupos para cada característica. Pode-se observar que a hipótese nula $H_0$ não é rejeitada para todas as características, o que indica que a média de algumas são diferentes para os grupos encontrados. O $H_{cal}$ é o valor calculado pelo método de *Kruskal-Wallis*, nos caso em que p-valor $< 5\%$, $H_{cal}$ é maior que o valor tabelado na Distribuição Qui-Quadrado.

Tabela 37 – Análise de variância intergrupos para cada característica - Paleosul (Constituintes Diagenéticos). Os valores com * indicam os p-valores que rejeitaram a $H_0$. Os grupos foram determinados pelo procedimento computacional. O nível de significância é de 0.05.

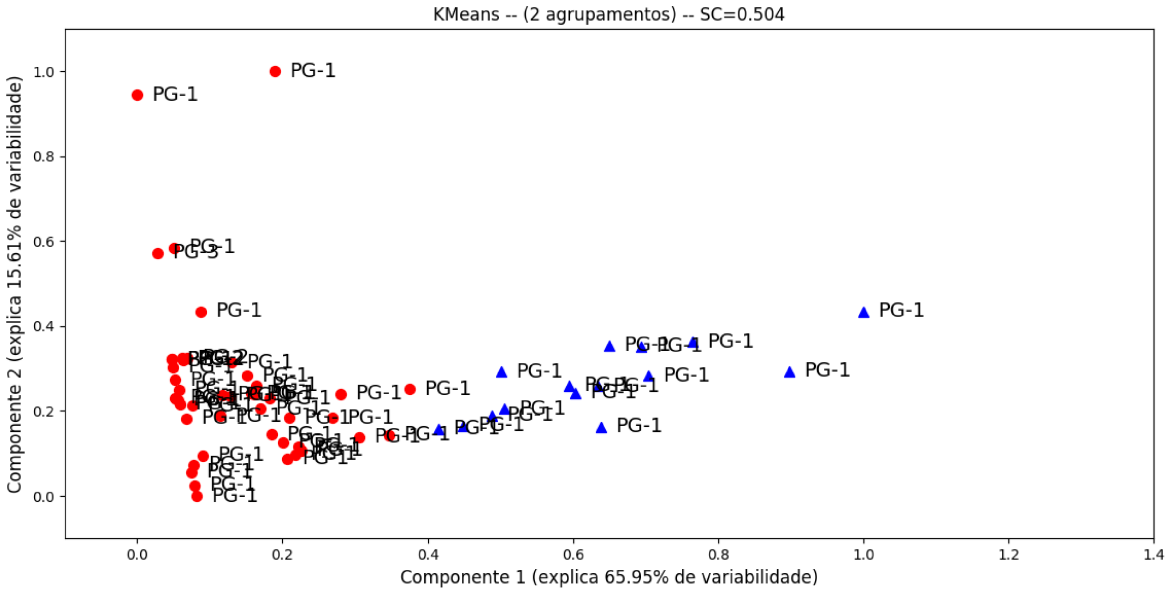
| Características | $H_{cal}$ | $p$-valor |
|---|---|---|
| Cim. Carbonático | 33.644505 | 6.616153e-09* |
| Glauconita | 2.104031 | 1.469114e-01 |
| Caolinita | 4.409500 | 3.573931e-02* |
| Por. Intragranular | 0.103093 | 7.481492e-01 |
| Por. Intergranular | 0.285393 | 5.931883e-01 |
| Pirita | 1.969455 | 1.605057e-01 |
| Cres. Sec. Quartzo | 1.962362 | 1.612609e-01 |
| Siderita | 1.238707 | 2.657204e-01 |

Fonte: Elaborada pelo autor (2020).

A Figura 58 exibe o resultado do K-Means para o Paleosul (Constituintes Diagenéticos). A primeira componente principal expressa 65.95% de variabilidade dos dados e a

segunda expressa 15.61% da variabilidade.

Figura 58 - Resultado K-Means - Paleosul (Constituintes Diagenéticos). SC = 0.504

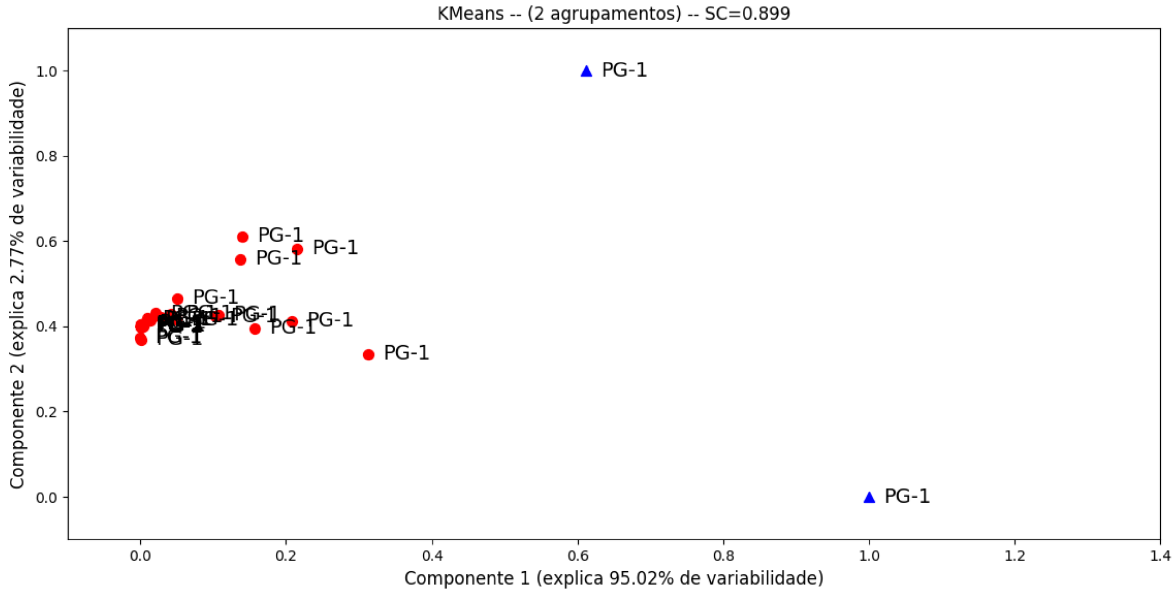



Fonte: Elaborada pelo autor (2020).

A Figura 59 exibe o resultado do K-Means aplicado nos dados após o pré-processamento o acréscimo das características polinomiais. O grau do polinômio foi 4 resultando em 495 características. A primeira componente principal expressa 95.02% de variabilidade dos dados enquanto a segunda expressa 2.77% da variabilidade. Pode-se observar que houve um aumento no valor do coeficiente de silhueta em relação aos dados originais.

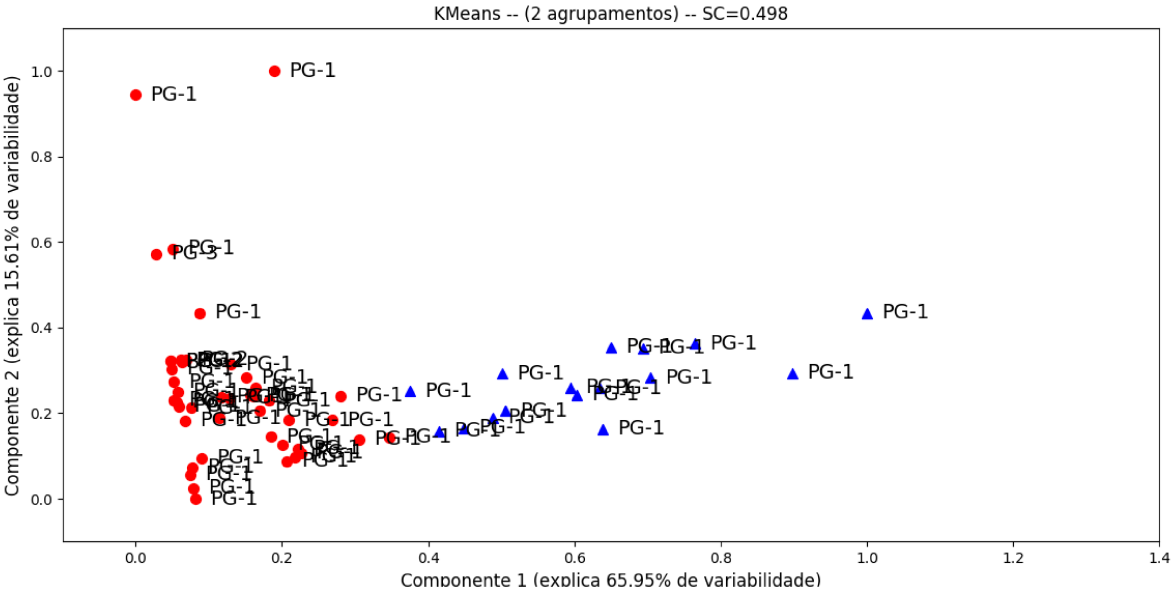
A Figura 60 exibe o resultado do K-Means aplicado nos dados originais após o Bootstrap ter sido empregado para encontrar os parâmetros.

Figura 59 - Resultado K-Means (Características Polinomiais) - Paleosul (Constituintes Diagenéticos). SC = 0.899




Fonte: Elaborada pelo autor (2020).

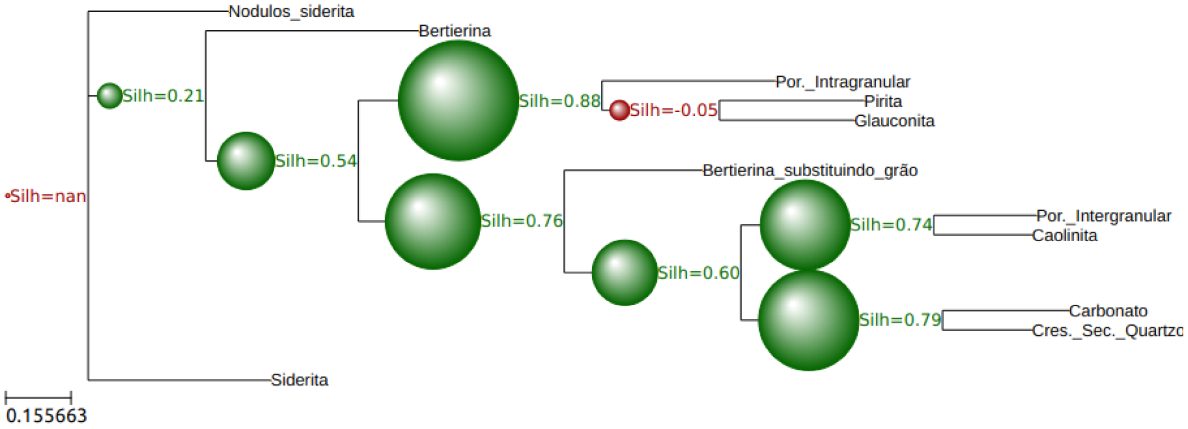
Figura 60 - Resultado K-Means (Bootstrap) - Paleosul (Constituintes Diagenéticos). SC = 0.533




Fonte: Elaborada pelo autor (2020).

Segundo (41) os eventos ocorreram da seguinte forma: na eodiagênese ocorreu a precipitação de bertierina na forma de cutículas e na forma de óoides, a precipitação de siderita na forma de nódulos; na eodiagênese secundária houve a precipitação de pirita e dolomita; na mesodiagênese houve o crescimento secundário de quartzo e cimentação por calcita; na telodiagênese ocorreu a geração de porosidade secundária e a precipitação de caolinita. O modelo proposto (Figura 61) sugere que na eodiagênese houve a precipitação de bertierina na forma de cutículas e na forma de óoides, a precipitação de siderita na forma de nódulos; na mesodiagênese ocorreu a precipitação de pirita, a geração de porosidade secundária e a precipitação de caolinita; na telodiagênese houve o crescimento secundário de quartzo e cimentação por calcita.

Figura 61 - Visualização do resultado da Análise Filogenética Constituintes - Paleosul.



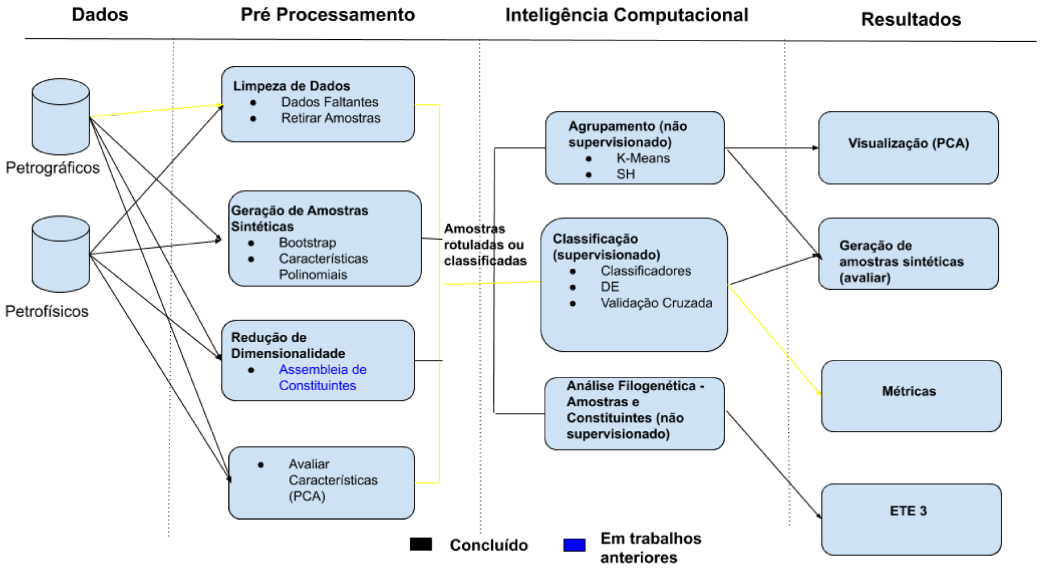

Fonte: Elaborada pelo autor (2020).

### 5.3.7 **Dados Petrográficos - La Ciotat-1**

A Figura 62 mostra, através das linhas amarelas, quais procedimentos da metodologia descritos na Seção 5.1 foram aplicados na na base de dados La Ciotat-1.

As configurações de parâmetros usadas no processo evolucionário para a seleção de modelos dos classificadores ANN, DT, ELM, GB, KNN e SVM são exibidas na Tabela 38 e os parâmetros dos métodos na Tabela 39. $CR$ foi definido para 0.7 e $F$ foi escolhido aleatoriamente no intervalo [0.5, 1]. Uma técnica chamada Dither, proposta por (140), seleciona aleatoriamente o parâmetro F do intervalo [0.5, 1.0] para cada geração ou para cada vetor de diferença que melhora significativamente o comportamento de convergência, especialmente em funções objetivas ruidosas. Um total de 50 indivíduos evoluíram com menos de 50 gerações para cada iteração. Para os parâmetros que são inteiros ou *strings*, o inteiro mais próximo foi usado para definir os parâmetros a serem usados no classificador. No caso da *string*, cada inteiro representa uma *string*, com por exemplo a função de

Figura 62 - Fluxograma ilustrando, através das linhas amarelas, a metodologia aplicada a base de dados La Ciotat-1.




Fonte: Elaborada pelo autor (2020).

ativação do ELM (Tabela 19. A função objetivo (a ser maximizada) é o F1 dada pela Eq. (4.25).

Tabela 38 – Configurações de parâmetros DE usadas na otimização de hiperparâmetros do classificadores.

| Parâmetros | Nome | Valor/Variação |
|---|---|---|
| $CR$ | Fator de Amplificação | 0.7 |
| $F$ | Taxa de Mutação | [0.5, 1] (escolhido aleatoriamente) |
| $NP$ | Tamanho da população | 50 |
| $J_{max}$ | Nº de Gerações | 50 |
| $\theta_L$ | Limites Inferiores | Tabela 40 |
| $\theta_U$ | Limites Superiores | Tabela 40 |
| Função Fitness | F1 | Eq. (4.25) |

Fonte: Elaborada pelo autor (2020).

A Tabela 41 apresenta a média e o desvio padrão da Acurácia, F1, Kappa, Recall e $R^2$ para os métodos ANN, DT, ELM, GB, KNN e SVM. Em (141) o coeficiente de correlação ($R^2$) foi utilizado para avaliar o desempenho de classificadores. Pode-se observar que o ANN apresentou resultado superior quando comparado com métodos comumente aplicados na literatura. Observando a acurácia, ANN classificou em média 87.8% das amostras corretamente e segundo o teste Kappa, os resultados concordam substancialmente

Tabela 39 – Configuração dos classificadores.

| Método | Variável de Decisão | Descrição |
|---|---|---|
| ANN | $\theta_{i,1}$ | Função de Ativação |
| | $\theta_{i,2}$ | Algoritmo de Treinamento |
| | $\theta_{i,3}$ | $L_2$ Coef. Regularização |
| | $\theta_{i,4}$ | N$^o$ de camadas |
| | $\theta_{i,5:9}$ | N$^o$ de neurônios, 1$^a$-5$^a$ camadas |
| DT | $\theta_{i,1}$ | Função para medir a qualidade de uma divisão |
| | $\theta_{i,2}$ | N$^o$ mínimo de amostras necessárias para dividir um nó interno |
| ELM | $\theta_{i,1}$ | N$^o$ de neurônios |
| | $\theta_{i,2}$ | Função de Ativação |
| | $\theta_{i,3}$ | $\alpha$ |
| GB | $\theta_{i,1}$ | Taxa de aprendizado |
| | $\theta_{i,2}$ | N$^o$ de Estimadores |
| | $\theta_{i,3}$ | Profundidade Máxima da Árvore |
| | $\theta_{i,4}$ | N$^o$ mínimo de amostras para cada divisão |
| | $\theta_{i,5}$ | N$^o$ mínimo de amostras para cada nó |
| | $\theta_{i,6}$ | Fração ponderada mínima da soma total de pesos necessários para estar em um nó folha |
| | $\theta_{i,7}$ | Fração de amostras a serem usadas para ajustar os aprendizes individuais |
| KNN | $\theta_{i,1}$ | Potência para a métrica de Minkowski |
| | $\theta_{i,2}$ | N$^o$ de Vizinhos mais próximos |
| | $\theta_{i,3}$ | Função de peso usada na previsão |
| SVM | $\theta_{i,1}$ | Kernel |
| | $\theta_{i,2}$ | C |
| | $\theta_{i,3}$ | $\gamma$ |

Fonte: Elaborada pelo autor (2020).

Tabela 40 – Limites Inferiores ($\theta_L$) e Superiores ($\theta_U$) dos métodos ANN, DT, ELM, GB, KNN e SVM utilizados no DE

| Método | $\theta_L$ | $\theta_U$ |
|---|---|---|
| ANN | $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9) =$ $(0.0, 0.0, 0.0, 1, 1, 1, 1, 1, 1)$ | $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9) =$ $(3.0, 2.0, 0.1, 5, 50, 50, 50, 50, 50)$ |
| DT | $(\theta_1, \theta_2) = (0, 2)$ | $(\theta_1, \theta_2) = (1, 20)$ |
| ELM | $(\theta_1, \theta_2, \theta_3) = (1, 0, 0)$ | $(\theta_1, \theta_2, \theta_3) = (300, 9, 1)$ |
| GB | $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7) =$ $(0.001, 300, 10, 5, 5, 0.0, 0)$ | $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7) =$ $(0.8, 900, 100, 50, 50, 0.5, 1)$ |
| KNN | $(\theta_1, \theta_2, \theta_3) = (1, 0, 1)$ | $(\theta_1, \theta_2, \theta_3) = (50, 1, 3)$ |
| SVM | $(\theta_1, \theta_2, \theta_3) = 1, 10^{-5}, 0)$ | $(\theta_1, \theta_2, \theta_3) = 10^5, 10^3, 2$ |

Fonte: Elaborada pelo autor (2020).

com a classificação realizada por (3). Por ter apresentado melhor desempenho o ANN foi escolhido como base para o teste de Wilcoxon. A diferença do resultado obtido pelo ANN e os encontrados pelos métodos GB, KNN e SVM não são considerados estatisticamente significantes segundo o teste de Wilcoxon. Para comparar os resultados com (141), o Coeficiente de Correlação de Pearson foi analisado. Embora adequada para tarefas de regressão, essa medida de desempenho foi mantida para comparar o desempenho dos classificadores. O método obteve $R^2 = 98.0\%$ (em média para 50 iterações), alcançando melhor desempenho do que a metodologia proposta por (141) ($R^2 = 85.62\%$).

A Figura 63 ilustra os perfis com a litologia classificada por (3) e pelos classificadores empregados. A litologia classificada pelo ANN está destacada de vermelho. Analisando as fácies em relação a profundidade pode-se observar que em muitos intervalos a classificação do ANN coincide com a de (3), por exemplo por volta dos 60 metros ambos apontaram que a identificação naquele local é C4.

A Tabela 42 exibe os melhores valores encontrados a partir do DE para os parâmetros do ANN. Os parâmetros são AF (Função de Ativação), HL (Camadas Ocultas) e Solver. Sob tal configuração, o ANN foi capaz de classificar 92.5% das amostras corretamente.

Tabela 41 – Média e Desvio Padrão da Acurácia, F1, Recall, Kappa e $R^2$ para validação cruzada 5-fold. Os melhores resultados estão em negrito enquanto * indica que a diferença observada não é estatisticamente significante com o respectivo melhor resultado. Um par de conjuntos de resultados são estatisticamente significantemente diferente quando o p-valor a partir do teste não-paramétrico de Wilcoxon é menor que 0.05. O * indica o melhor resultado obtido no conjunto de treinamento

| Classificador | Acurácia | F1 | Recall | Kappa | $R^2$ |
|---|---|---|---|---|---|
| ANN | **0.878±0.031** | **0.870±0.030** | **0.877±0.030** | **0.878±0.031** | **0.980±0.008** |
| DT | 0.853±0.044 | 0.845±0.047 | 0.853±0.044 | 0.822±0.053 | 0.976±0.012 |
| ELM | 0.807±0.037 | 0.795±0.038 | 0.807±0.037 | 0.765±0.045 | 0.914±0.049 |
| GB | 0.863±0.040* | 0.854±0.042 | 0.863±0.040 | 0.834±0.049 | 0.975±0.017 |
| KNN | 0.668±0.036* | 0.653±0.037 | 0.668±0.036 | 0.597±0.043 | 0.931±0.025 |
| SVM | 0.690±0.051* | 0.676±0.053 | 0.690±0.051 | 0.625±0.061 | 0.952±0.014 |
| Ref. (141) | – | – | – | – | 0.8562(*) |

Fonte: Elaborada pelo autor (2020).

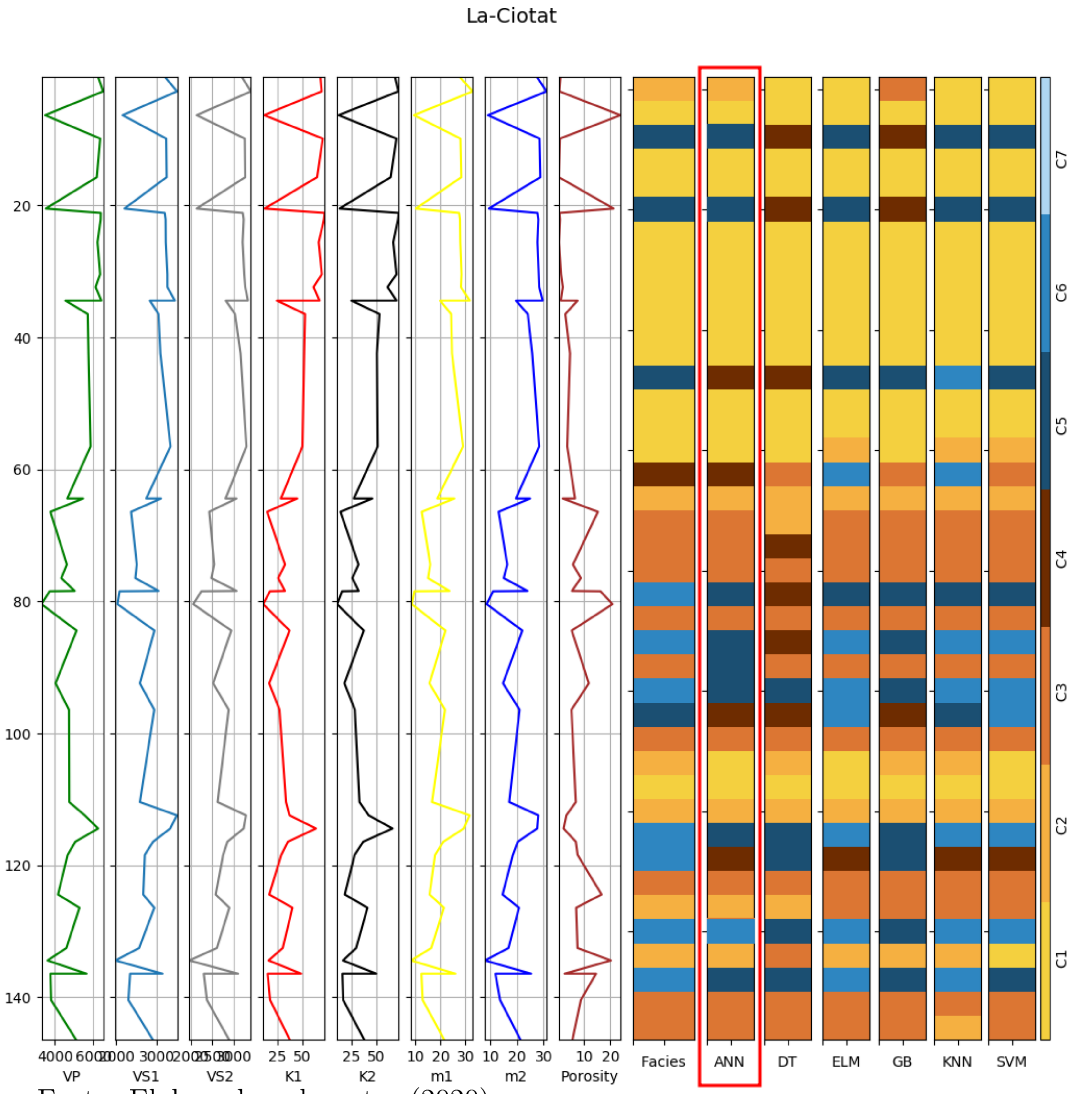
Tabela 42 – Melhor modelo do ANN (de acordo com F1) produzida pela Evolução Diferencial (para 50 iterações independentes).

| Parâmetros | Acurácia | F1 | Recall | Kappa | $R^2$ |
|---|---|---|---|---|---|
| AF = Identity, HL = (10, 1), Solver = L-BFGS, RC = 0.026 | 0.925 | 0.937 | 0.925 | 0.910 | 0.972 |

Fonte: Elaborada pelo autor (2020).

A Tabela 43 exibe a média e o desvio padrão da Acurácia, F1 e Recall para cada classe. Pode observar que as classes C1 e C7 produziram os melhores valores médios para

Figura 63 - Perfis e Fácies Preditas - La Ciotat-1




Fonte: Elaborada pelo autor (2020).

Acurácia, F1 e Recall.

A Figura 64 mostra a matriz de confusão das sete classes litológicas, onde as colunas representam a classificação obtida, as linhas representa a da referência e a diagonal principal representa a a proporção das classes que foram classificadas corretamente. Observa-se que classes litológicas são classificadas erroneamente para outras classes. No geral, as classes C1, C4 e C7 possuem os maiores valores de acurácia. Na classe C1, 7% das amostras foram classificadas como C2. Considerando a classe C2, 4% foram preditas como C1 e 5% como C3. Para as amostras pertencentes a C3 19% como C2 e% como C4. Para a classe C6 6% das amostras foram classificadas como C7 enquanto para a classe C7 6% foram consideradas como C6. Para C5, que possui apenas uma amostra, que foi mais classificada na classe C4 devido às semelhanças nas propriedades elásticas, mineralógicas

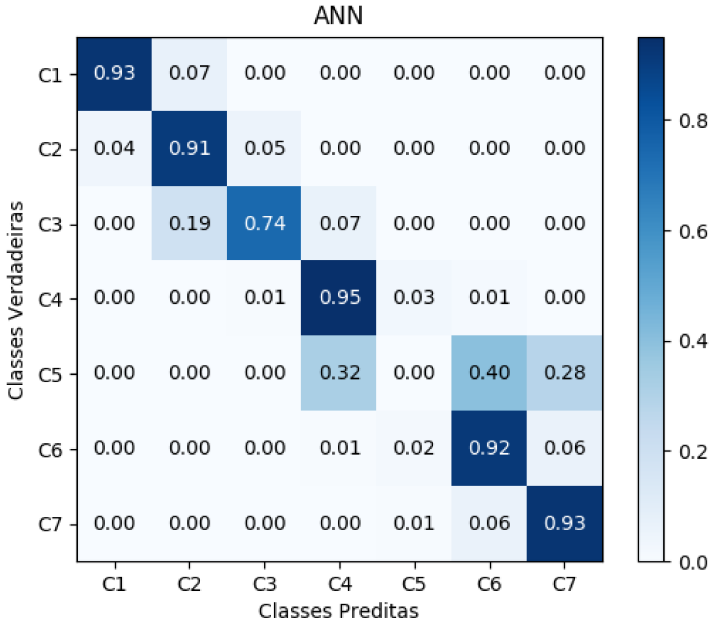
Tabela 43 – Média e Desvio Padrão da Acurácia, F1 e Recall, para cada classe, para validação cruzada 5-fold (ANN). Um total de 50 iterações independentes foram avaliadas.

| Classes | # Número de Amostras | Acurácia | F1 | Recall |
|---|---|---|---|---|
| C1 | 5 | 0.941± 0.088 | 0.932± 0.082 | 0.932± 0.111 |
| C2 | 8 | 0.836± 0.071 | 0.867± 0.062 | 0.907± 0.094 |
| C3 | 6 | 0.920± 0.101 | 0.809± 0.081 | 0.740± 0.131 |
| C4 | 9 | 0.928± 0.082 | 0.935± 0.051 | 0.951± 0.078 |
| C5 | 1 | 0± 0 | 0± 0 | 0± 0 |
| C6 | 4 | 0.818± 0.179 | 0.856± 0.152 | 0.915± 0.157 |
| C7 | 7 | 0.931± 0.088 | 0.928± 0.065 | 0.931± 0.083 |

Fonte: Elaborada pelo autor (2020).

e petrográficas, como pode ser visto na Tabela 15. Embora o ANN tenha aprendido a amostra de treinamento do C5, ele não registrou como generalizar novas situações, o que representa um problema de overfitting.

Figura 64 - Matriz de Confusão do Conjunto de Teste (ANN). As entradas normalizadas foram medidas em 50 execuções independentes.




Fonte: Elaborada pelo autor (2020).

A construção de bancos de dados litológicos geralmente requer um processo subjetivo e manual para interpretar e classificar os dados descritivos. Embora detalhadas, algumas inconsistências podem estar presentes nas descrições litológicas, como (142): as informações registradas dependem da experiência, habilidade e conhecimento prévio do operador que registra os logs. Os bancos de dados de litologia frequentemente contêm dados coletados durante um período de muitos anos e gerados por diferentes equipamentos de perfuração e com diferentes objetivos e metas. Outra limitação de potencial é a grande variação

na composição de alguns materiais comuns (143). Como resultado, nos casos em que a litologia foi mapeada incorretamente nos dados de origem ou não foi registrada devido a limitações de escala, podendo levar a resultados insatisfatórios no desempenho dos classificadores.

Essa abordagem produziu um modelo com boa precisão de classificação (o melhor modelo apresentado na Tabela 42 obteve 92.5%), o que pode ajudar geólogos/petrólogos a determinar a heterogeneidade de um reservatório. Além disso, especialistas podem aplicar o modelo de classificação para analisar um banco de dados de registros de poços durante a fase exploratória, o que também possibilita uma melhoria da análise da bacia, proporcionando modelos geológicos mais realistas e, consequentemente, melhor eficiência no desenvolvimento dos campos. Os resultados obtidos com a metodologia proposta aplicada na base de dados La Ciotat-1 originaram a publicação do artigo (144).

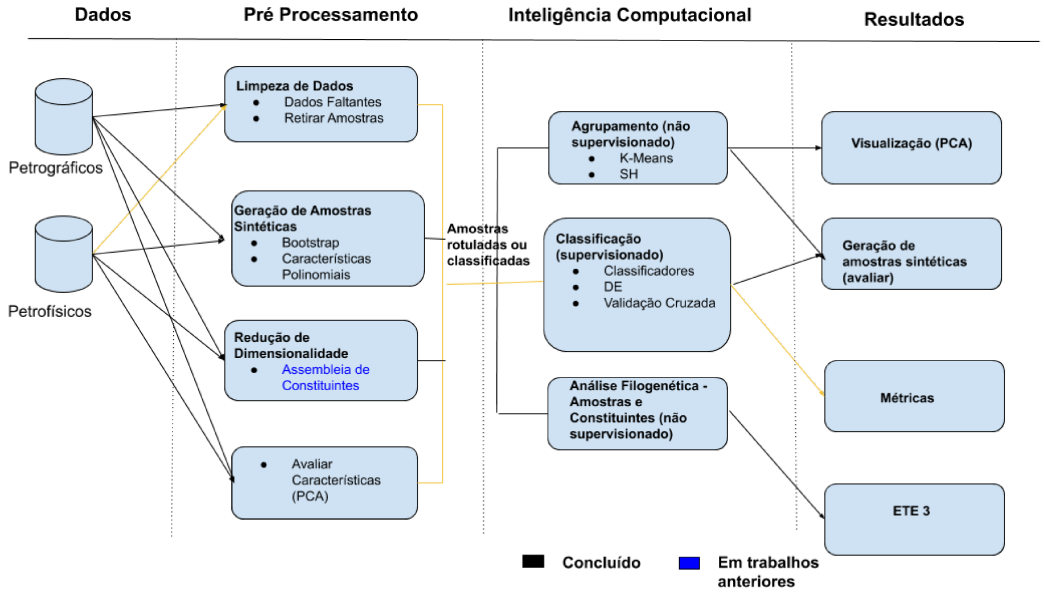
### 5.3.8 **Dados Petrofísicos - Daniudui (DGF) e Hangjinqi (HGF)**

A Figura 65 mostra, através das linhas amarelas, quais procedimentos da metodologia descritos na Seção 5.1. A Tabela 44 apresenta a média e o desvio padrão da Acurácia, F1, Kappa e Recall para os métodos ANN, DT, ELM, GB, KNN e SVM. Nesse caso não avaliou-se $R^2$, pois não teria como compará-la com trabalhos anteriores. Observando a acurácia, GB classificou em média 81.9% das amostras corretamente nos dados DGF e 85.0% nos dados HGF. Por ter apresentado melhor desempenho o GB foi escolhido como base para o teste de Wilcoxon. Para DGF, a diferença do resultado obtido pelo GB e os encontrados pelos métodos ANN, DT e KNN não são considerados estatisticamente significantes segundo o teste de Wilcoxon e para HGF, os classificadores DT e KNN.

As Figuras 66 e 67 ilustram os perfis com a litologia, dos campos DGF e HGF respectivamente, definida em (2) e pelos classificadores empregados. A litologia classificada pelo GB está destacada de vermelho. Analisando as fácies em relação a profundidade pode-se observar que em muitos intervalos a classificação do GB coincide com a de (3), por exemplo, pode-se destacar para DGF por volta dos 2500 metros ambos apontaram que a identificação naquele local é CR e para HGF por volta de 2100 metros ambos identificaram como CS. Os parâmetros do melhor modelo encontraram a seleção do modelo evolutivo (com melhor valor de F1) e as respectivas métricas de desempenho são mostradas na Tabela 45. Sob tal configuração, o GB foi capaz de classificar 83.1% das amostras corretamente para o DGF e 86.4% para o HGF. A Tabela 46 exibe a média e desvio padrão da Acurácia, F1 e Recall para cada classe. Observa-se que, para o DGF a classe que obteve maior acurácia foi a C, 95.3% e para o HGF, a classe que obteve a maior precisão foi a M, 91.1%.

A Figura 68 exibe a matriz de confusão das classes litológicas. No caso dos dados referente a DGF, as classes C (95%), CR (94%) e M (91%) têm os valores maiores de acurácias. Na classe C, 4% das amostras foram classificadas erroneamente como M. Para a

Figura 65 - Fluxograma ilustrando, através das linhas amarelas, a metodologia aplicada a base de dados DGF e HGF.




Fonte: Elaborada pelo autor (2020).

classe CR, 4% como M e 1% como arenito (CS). Na classe CS, 33% como arenito (FS, MS, PS), 2% como S, 3% como M e 1% como C. Considerando a classe FS, 14% das amostras foram classificadas erroneamente como arenito (CS, MS), 4% como S, 1% como CR. Na classe MS, 19% foram classificadas incorretamente como arenito (CS, FS, PS) e 2% como M. Para amostras de PS, 15% como arenito (CS, FS, MS) e 1% como M, enquanto para a classe S foram consideradas 13% como arenito (FS, MS, PS) e 6% como M.

No geral, para o conjunto de dados HGF, as classes M (91%), C (88%) e PS (88%) possuem o maior percentual de amostras classificadas corretamente ao avaliar a acurácia. Na classe C, 12% das amostras foram classificadas erroneamente como argilito (M). Considerando a classe CS, 18% foram preditas como arenitos (PS, MS) e 1% como M. Na classe FS, 18% das amostras foram classificadas como arenito (CS, MS, PS) e 1% como M. Para amostras pertencentes a M, 8% como arenito (PS, MS, CS). Na classe MS, 17% das amostras foram classificadas incorretamente como arenito (PS, FS, CS), 2% como M e 1% como siltito (S). Para as amostras de PS, 2% como argilito (M) e 10% como arenito (FS, MS, CS), enquanto para a classe S 18% forma consideradas arenitos (CS, FS, MS, PS), 3% como M.

A Figura 69 retrata os gráficos de barras comparando os resultados encontrados para cada classe DGF com aqueles disponíveis em Referência (2). Os resultados para esta base de dados foram semelhantes em relação à referência. Uma diferença mais significativa

Tabela 44 – Média e Desvio Padrão da Acurácia, F1 e Recall, para cada base de dados, para validação cruzada 5-fold. Um total de 50 iterações independentes foram avaliadas. Os melhores resultados estão em negrito enquanto * indica que a diferença observada não é estatisticamente significante com o respectivo melhor resultado. Um par de conjuntos de resultados são estatisticamente significantemente diferente quando o p-valor a partir do teste não-paramétrico de Wilcoxon é menor que 0.05.

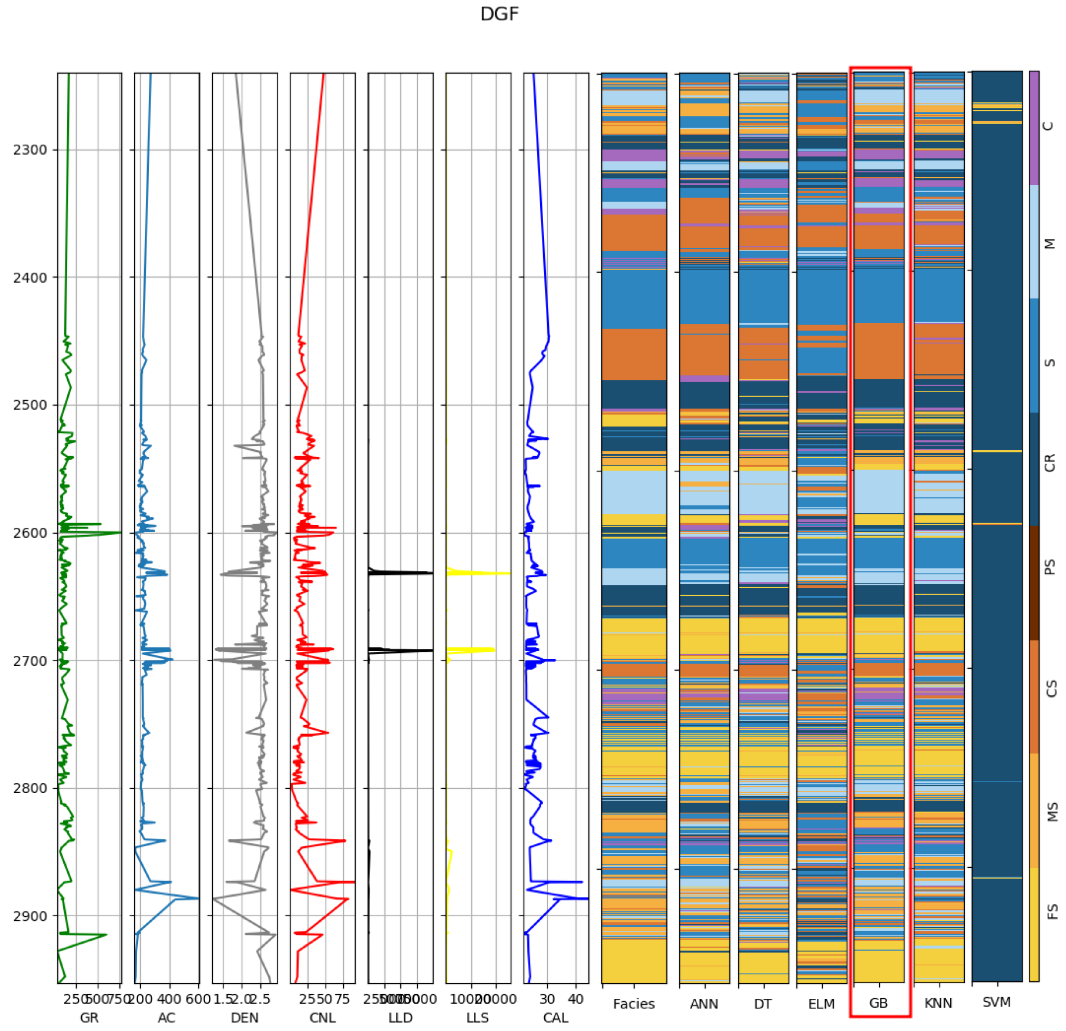
| Base de Dados | Método | Acurácia | F1 | Recall | Kappa |
|---|---|---|---|---|---|
| DGF | ANN | 0.679±0.010* | 0.676±0.010 | 0.679±0.010 | 0.622±0.012 |
| DGF | DT | 0.737±0.009* | 0.738±0.009 | 0.737±0.009 | 0.692±0.010 |
| DGF | ELM | 0.497±0.007 | 0.481±0.006 | 0.497±0.007 | 0.397±0.008 |
| DGF | GB | 0.819±0.006 | 0.818±0.006 | 0.819±0.006 | 0.787±0.007 |
| DGF | KNN | 0.725±0.007* | 0.724±0.007 | 0.725±0.007 | 0.678±0.008 |
| DGF | SVM | 0.314±0.140 | 0.213±0.186 | 0.314±0.140 | 0.119±0.195 |
| HGF | ANN | 0.654±0.016 | 0.651±0.017 | 0.654±0.016 | 0.562±0.021 |
| HGF | DT | 0.776±0.008* | 0.776±0.008 | 0.776±0.008 | 0.720±0.010 |
| HGF | ELM | 0.536±0.007 | 0.521±0.010 | 0.536±0.007 | 0.406±0.010 |
| HGF | GB | 0.850±0.006 | 0.850±0.006 | 0.850±0.006 | 0.812±0.008 |
| HGF | KNN | 0.789±0.006* | 0.789±0.006 | 0.789±0.006 | 0.737±0.008 |
| HGF | SVM | 0.344±0.082 | 0.215±0.134 | 0.344±0.082 | 0.072±0.129 |

Fonte: Elaborada pelo autor (2020).

foi observada para a classe CS, onde GB foi superior. A comparação para o conjunto de dados do HGF é apresentada na Figura 70. Pode-se observar que GB se mostra superior para classificar a classe C.

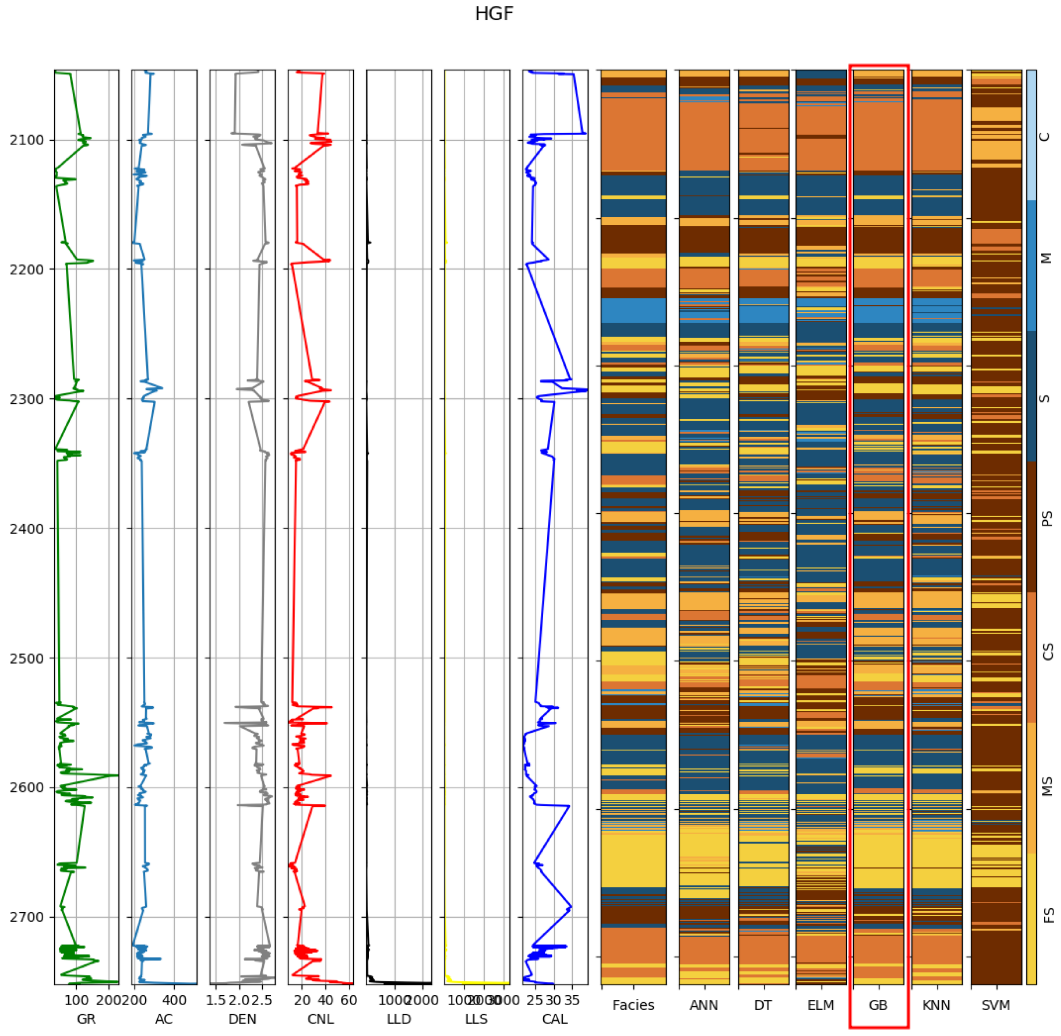
A Figura 71 exibe as distribuições de parâmetros em 50 execuções independentes para conjuntos de dados DGF e HGF. A partir dos resultados pode-se observar que o conjunto de dados DGF necessitou de maior fração de amostras (parâmetro Subamostra) quando comparado ao conjunto de dados HGF. No geral, a seleção do modelo GB leva a um conjunto semelhante de parâmetros para ambos os conjuntos de dados.

Figura 66 - Perfis e Facies Preditas - DGF




Fonte: Elaborada pelo autor (2020).

Figura 67 - Perfis e Facies Preditas - HGF




Fonte: Elaborada pelo autor (2020).

Tabela 45 – Melhores parâmetros do modelo (de acordo com F1) em 50 execuções independentes.

| Base de Dados | Parâmetros | Acurácia | F1 | Recall |
|---|---|---|---|---|
| DGF | Taxa de Aprendizado = 0.109, N$^o$ de Estimadores = 388, Máx. Profundidade = 53, Min. Amostras Divididas = 34, Min. Amostras nas Folhas = 21, Min. Folha de fração de peso = 0.013, Subamostras = 0.928 | 0.831 | 0.830 | 0.831 |
| HGF | Taxa de Aprendizado = 0.177, N$^o$ de Estimadores = 702, Máx. Profundidade = 32 79, Min. Amostras Divididas = 34, Min. Amostras nas Folhas = 21, Min. Folha de fração de peso = 0.024, Subamostras = 0.795 | 0.864 | 0.864 | 0.864 |

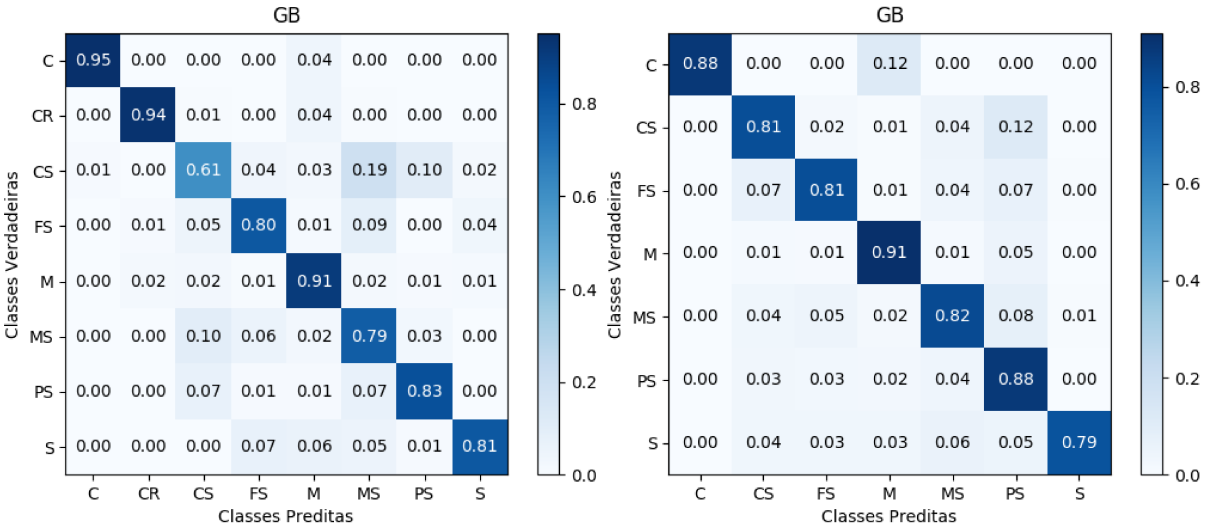
Fonte: Elaborada pelo autor (2020).

Tabela 46 – Média e Desvio Padrão da Acurácia, F1 e Recall, para cada classe, para validação cruzada 5-fold. Um total de 50 iterações independentes foram avaliadas.

| Classe | Base de Dados | F1 | Precisão | Recall |
|---|---|---|---|---|
| C | DGF | 0.953± 0.014 | 0.967± 0.009 | 0.953± 0.014 |
| CR | DGF | 0.942± 0.027 | 0.932± 0.016 | 0.942± 0.027 |
| CS | DGF | 0.606± 0.032 | 0.621± 0.023 | 0.606± 0.032 |
| FS | DGF | 0.805± 0.017 | 0.814± 0.016 | 0.805± 0.017 |
| M | DGF | 0.908± 0.020 | 0.880± 0.015 | 0.908± 0.020 |
| MS | DGF | 0.787± 0.011 | 0.782± 0.010 | 0.787± 0.011 |
| PS | DGF | 0.831± 0.021 | 0.831± 0.015 | 0.831± 0.021 |
| S | DGF | 0.806± 0.038 | 0.813± 0.029 | 0.806± 0.038 |
| C | HGF | 0.881± 0.055 | 0.925± 0.039 | 0.881± 0.055 |
| CS | HGF | 0.806± 0.021 | 0.818± 0.014 | 0.806± 0.021 |
| FS | HGF | 0.807± 0.023 | 0.810± 0.016 | 0.807± 0.023 |
| M | HGF | 0.911± 0.013 | 0.918± 0.009 | 0.911± 0.013 |
| MS | HGF | 0.815± 0.017 | 0.819± 0.012 | 0.815± 0.017 |
| PS | HGF | 0.878± 0.013 | 0.853± 0.010 | 0.878± 0.013 |
| S | HGF | 0.789± 0.043 | 0.849± 0.028 | 0.789± 0.043 |

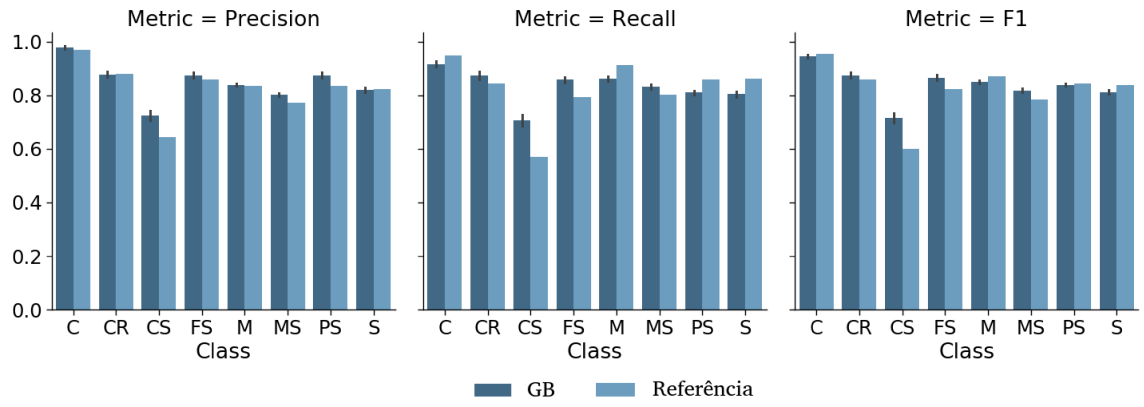
Fonte: Elaborada pelo autor (2020).

Figura 68 - Matriz de confusão no conjunto de dados de teste, DGF e HGF, respectivamente. As entradas normalizadas foram medidas em 50 execuções independentes



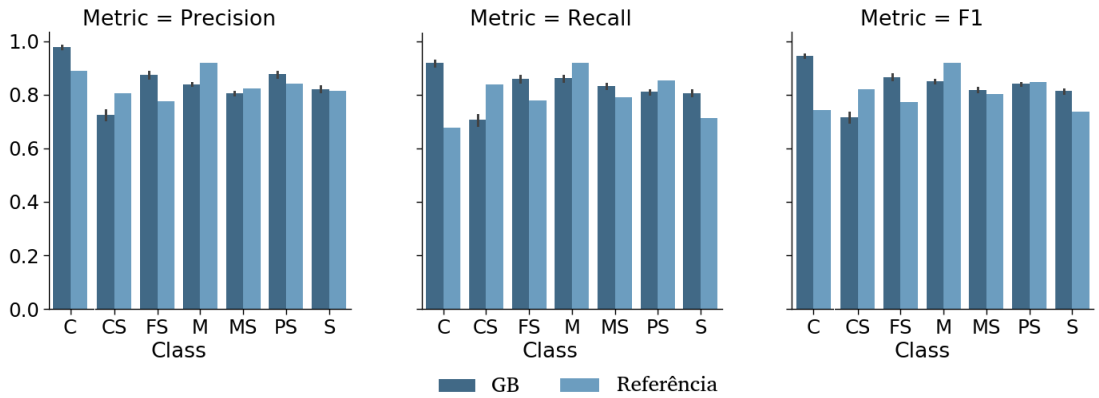

Fonte: Elaborada pelo autor (2020).

Figura 69 - Barplots para as classes de DGF com a referência (2).



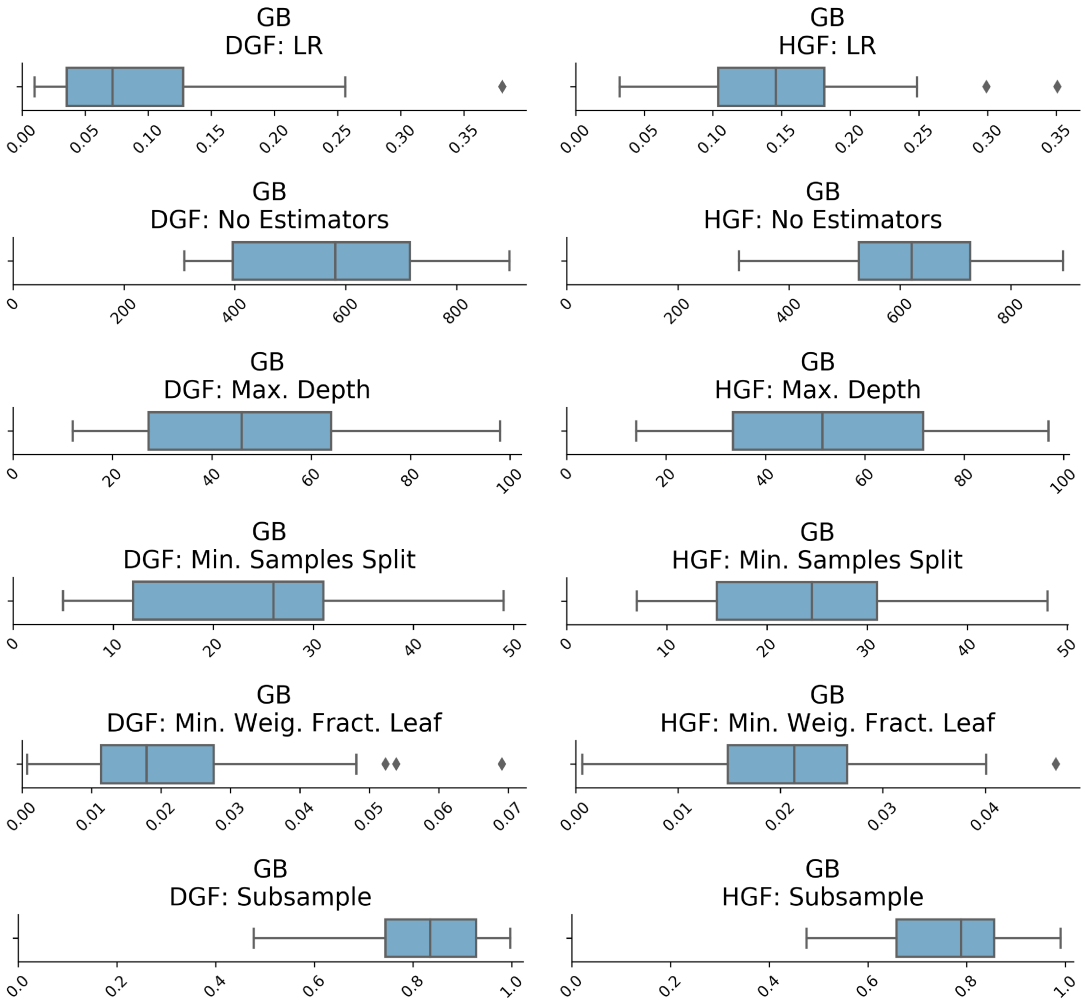

Fonte: Elaborada pelo autor (2020).

Figura 70 - Barplots para as classes de HGF com a referência (2).




Fonte: Elaborada pelo autor (2020).

Figura 71 - Distribuição dos parâmetros para 50 execuções.

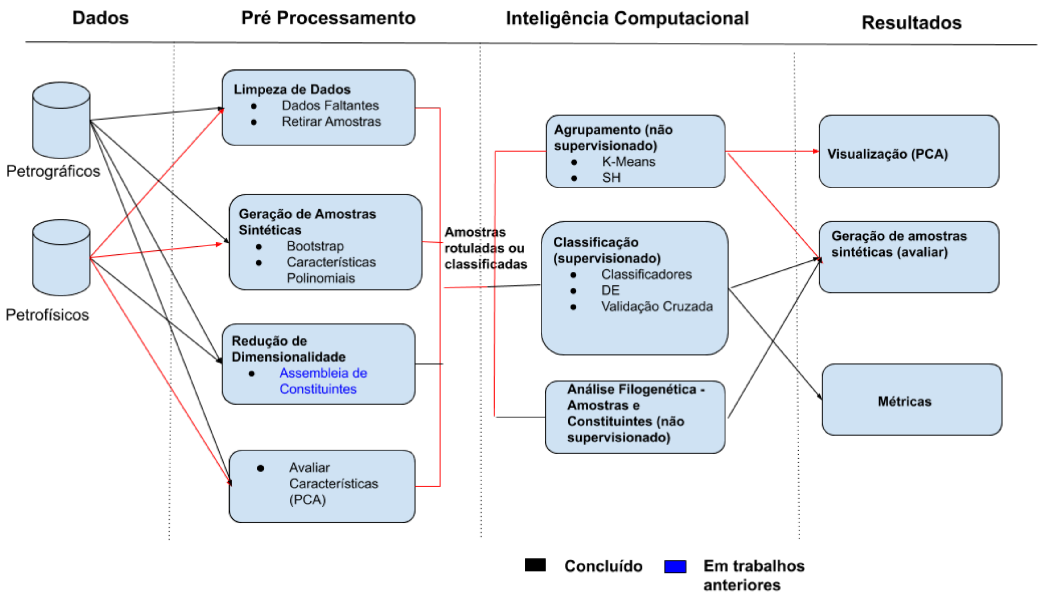



Fonte: Elaborada pelo autor (2020).

### 5.3.9 Dados Petrofísicos - Poço A

A Figura 72 mostra através das linhas vermelhas os procedimentos da metodologia descritos na Seção 5.2.

Figura 72 - Fluxograma ilustrando, através das linhas vermelhas, a metodologia aplicada a base de dados Poço A.



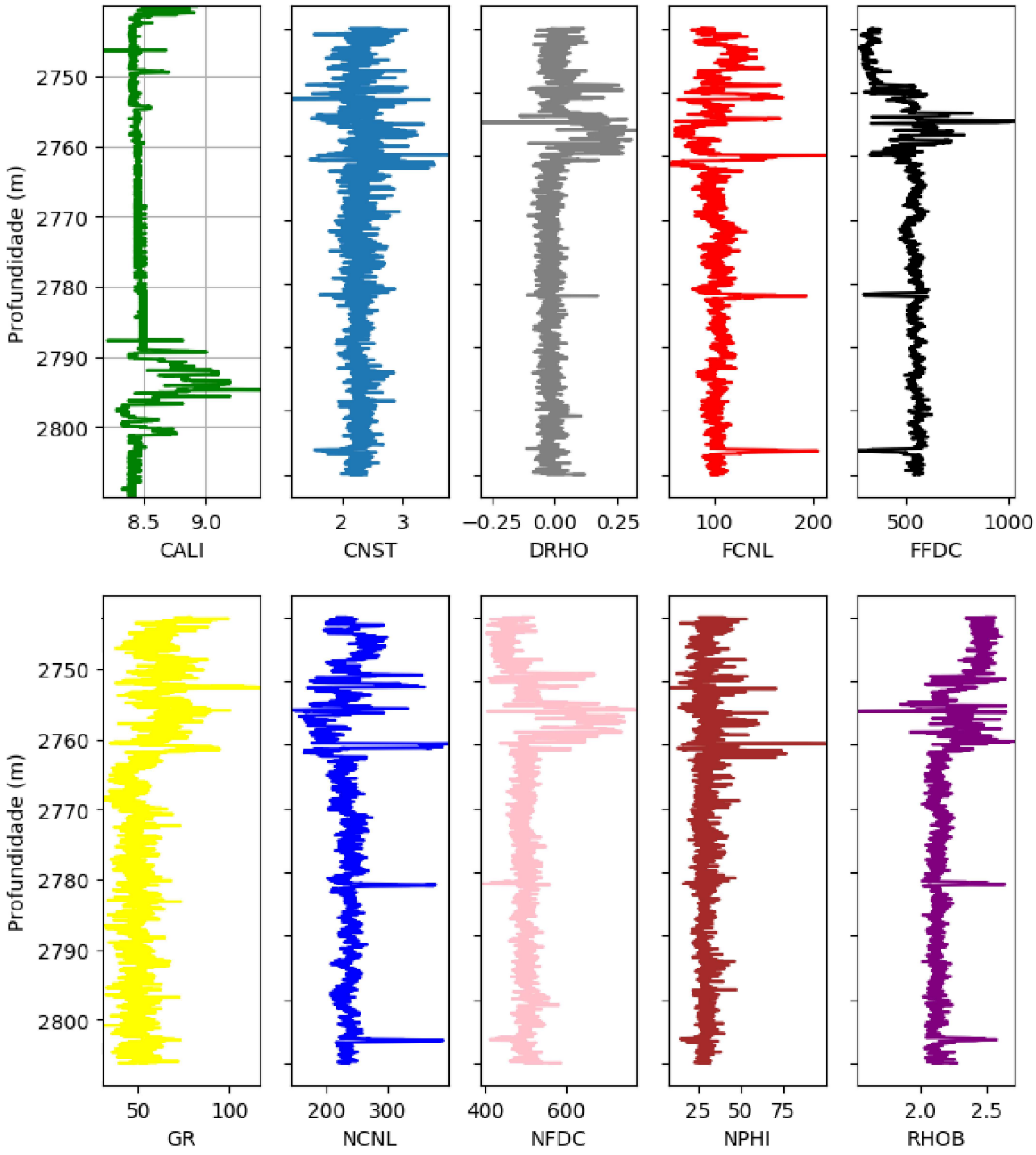

Fonte: Elaborada pelo autor (2020).

A Figura 73 apresenta as propriedades em relação a profundidade. Observa-se que há um aumento nos valores das características entre as profundidades 2790 e 2800 metros, enquanto a Figura 74 mostra a correlação entre as variáveis registradas em função da profundidade ao longo do processo de perfuração. Nota-se que há uma alta correlação entre as características NPHI e CNST que são características que estão relacionadas com a porosidade das rochas e entre DRHO e NFDC que são características que estão relacionadas com a densidade das rochas.

Na Tabela 47 encontram-se as componentes obtidas. Para a Componente 1 as características FFDC, NCNL e NFDC influenciam mais, na Componente 2 são as características FCNL, NCNL e NFDC que estão relacionadas a porosidade e a densidade de uma rocha. A Figura 75 apresenta a distribuição das amostras, nota-se que há um grupo bem denso.

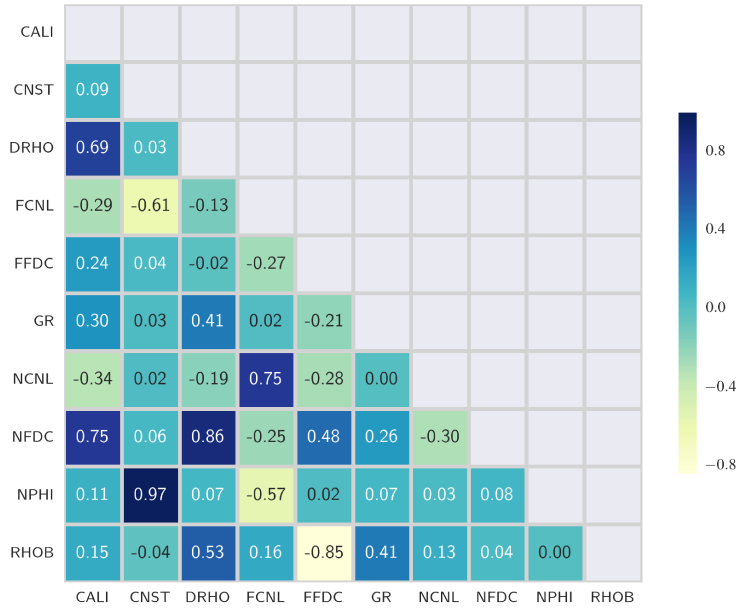
A Tabela 48 apresenta a média, o valor mínimo e o máximo para cada característica de acordo com as três classes litológicas geradas. Dentre as características que podem-se notar diferença entre os grupos estão DRHO, NFDC e NPHI que estão relacionadas com

Figura 73 - Propriedades em relação a profundidade para cada propriedade respectivamente - Poço A.




Fonte: Elaborada pelo autor (2020).

Figura 74 - Matriz de correlação entre os registros coletados - Poço A.




Fonte: Elaborada pelo autor (2020).

Tabela 47 – Componentes Principais em relação as propriedades petrofísicas. Os valores com * indicam os valores mais significativos obtidos para cada componente sugerindo qual propriedade petrofísica tem mais atuação sobre as mesmas - Poço A.

| Características | Componente 1 | Componente 2 |
|---|---|---|
| CALI | -6.74268787E-03 | 8.98905272E-04 |
| CNST | -1.60573376E-03 | 2.54982173E-04 |
| DRHO | -2.40105011E-04 | 1.30685655E-03 |
| FCNL | -1.77614567E-01 | -2.54969427E-01* |
| FFDC | 8.84238329E-01* | -8.03684520E-02 |
| GR | -1.03019566E-01 | 7.06628846E-02 |
| NCNL | -3.63797421E-01* | -5.43971276E-01* |
| NFDC | -2.07758007E-01* | 7.92124287E-01* |
| NPHI | -1.97103441E-02 | 1.29965584E-02 |
| RHOB | -3.74989694E-03 | 1.06468082E-03 |

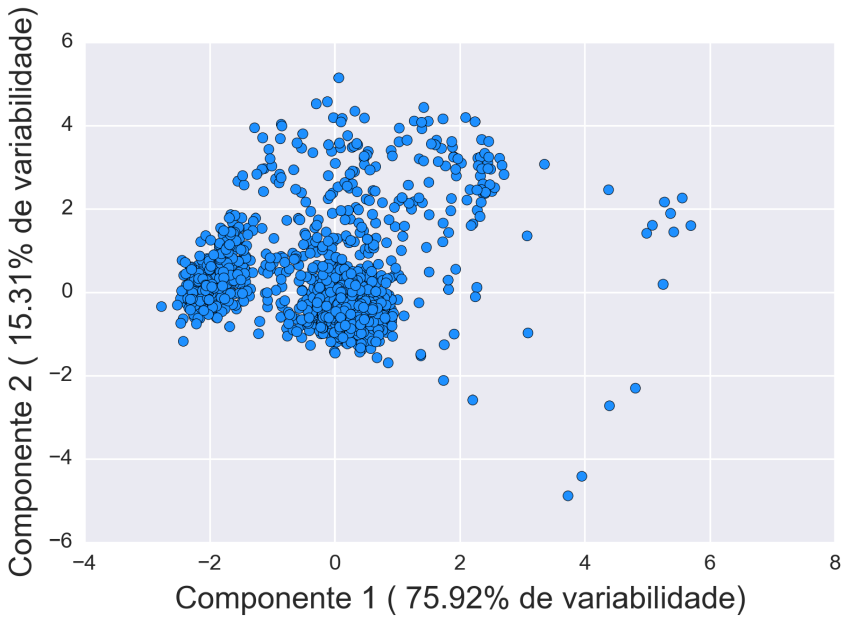
Fonte: Elaborada pelo autor (2020).

a porosidade e a densidade da rocha, sendo fatores que afetam o potencial de recuperação de hidrocarbonetos.

De acordo com a Tabela 48, para G0 DRHO tem valor entre aproximadamente -0.1381 e 0.3291, NFDC entre 443.5 e 719.75 e NPHI entre 7.3120 e 76.757. Para G1 DRHO tem valor entre aproximadamente -0.1093 e 0.2734, NFDC entre 387.5 e 646.0 e NPHI entre 11.9995 e 99.9023. Para G2 DRHO tem valor entre aproximadamente -0.2982 e 0.2897, NFDC entre 406.6665 e 773.25 e NPHI entre 15.1367 e 65.625.

A Figura 76 mostra o comportamento de cada propriedade em relação a profundi-

Figura 75 - Distribuição das Amostras - Poço A. A componente principal 1 expressa 75.92% da variabilidade dos dados e a componente principal 2 15.31%.




Fonte: Elaborada pelo autor (2020).

dade para os grupos encontrados pelo K-Means. Percebe-se que maioria das amostras do grupo G0 estão localizadas em profundidades maiores que 2800 metros. Em relação ao G1 estão abaixo de 2790 metros. As amostras associadas ao grupo G2 concentram-se onde houve um aumento no valor das propriedades, entre as profundidade 2790 e 2800 metros aproximadamente.

A Figura 77 mostra a distribuição das propriedades em função dos grupos determinados pela metodologia proposta. A mediana do três grupos estão próximas quando se trata de NPHI e CNST.

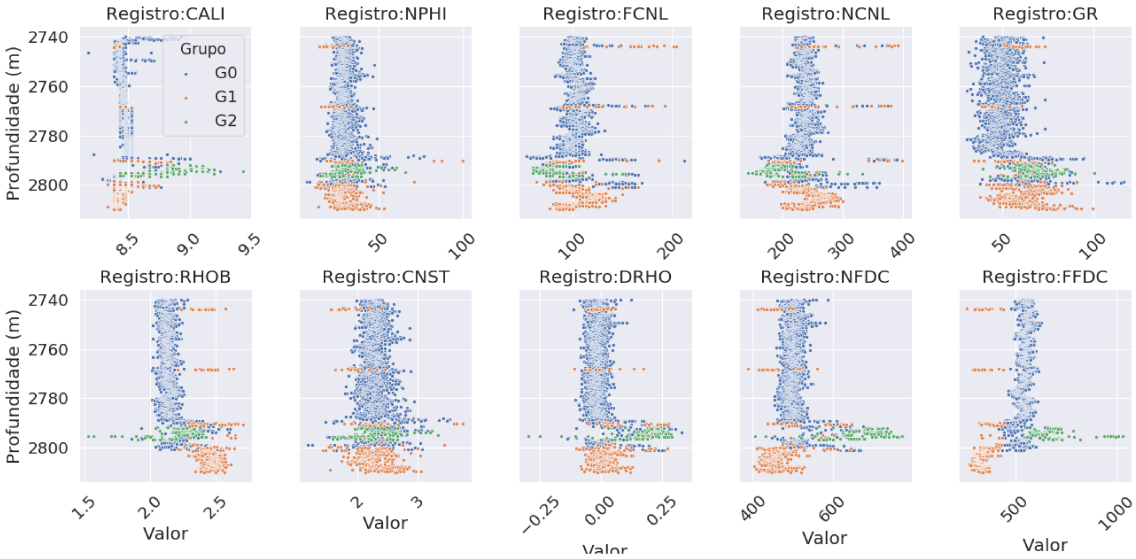
A Tabela 49 apresenta os resultados da análise de variância intergrupos para cada característica. Pode-se observar que $H_0$ é rejeitada para todas as características. O $H_{cal}$ é o valor calculado pelo método de *Kruskal-Wallis*, nos caso em que p-valor $< 5\%$, $H_{cal}$ é maior que o valor tabelado na Distribuição Qui-Quadrado. A Figura 78 exibe o resultado do método K-Means. A primeira componente principal expressa 75.92% de variabilidade dos dados, a segunda expressa 15.31% da variabilidade.

Tabela 48 – Grupos encontrados pelo K-Means. Para cada grupo tem-se a média de cada característica petrofísica e da profundidade. O número entre parênteses é o número de amostras atribuídas a cada grupo. O * indica as características que se diferenciaram entre os grupos - Poço A.

| Grupo | Logname | Média | Mínimo | Máximo |
|---|---|---|---|---|
| G0(1800) | CALI | 8.4835 | 8.1738 | 9.2460 |
| | CNST | 2.2943 | 1.1787 | 3.5195 |
| | DRHO* | -0.00794 | -0.1381 | 0.3291 |
| | FCNL | 102.8487 | 53.8750 | 212.5000 |
| | FFDC | 543.8936 | 421.0000 | 709.8750 |
| | GR | 51.1935 | 30.3750 | 116.7012 |
| | NCNL | 237.6139 | 163.7500 | 389.5938 |
| | NFDC* | 505.1021 | 443.5000 | 719.7500 |
| | NPHI* | 29.3162 | 7.3120 | 76.7578 |
| | RHOB | 2.1447 | 1.8554 | 2.6083 |
| G1(397) | CALI | 8.4576 | 8.3346 | 8.8476 |
| | CNST | 2.2466 | 1.4301 | 3.7416 |
| | DRHO | 0.02526 | -0.1093 | 0.2734 |
| | FCNL | 109.6860 | 63.5625 | 204.5313 |
| | FFDC | 339.5399 | 264.0000 | 444.0000 |
| | GR | 61.7102 | 34.3750 | 99.6448 |
| | NCNL | 246.4515 | 173.7500 | 399.0000 |
| | NFDC | 469.1577 | 387.5000 | 646.0000 |
| | NPHI | 28.5263 | 11.9995 | 99.9023 |
| | RHOB | 2.4590 | 2.2089 | 2.7089 |
| G2(100) | CALI | 8.9021 | 8.4307 | 9.4355 |
| | CNST | 2.3197 | 1.5976 | 3.3359 |
| | DRHO | 0.1733 | -0.2982 | 0.2897 |
| | FCNL | 83.7858 | 60.1875 | 153.2188 |
| | FFDC | 702.2404 | 556.6665 | 1029.7500 |
| | GR | 68.6862 | 39.0500 | 100.5000 |
| | NCNL | 194.9174 | 143.5000 | 332.7500 |
| | NFDC | 669.5732 | 406.6665 | 773.2500 |
| | NPHI | 30.6346 | 15.1367 | 65.6250 |
| | RHOB | 2.1813 | 1.5293 | 2.4242 |

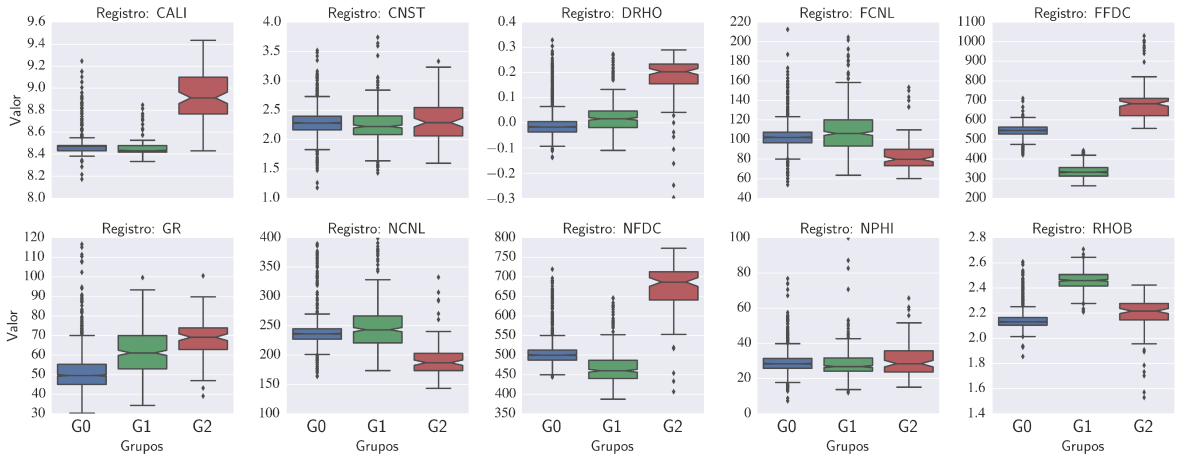
Fonte: Elaborada pelo autor (2020).

Figura 76 - Propriedades em relação a profundidade para cada grupo respectivamente - Poço A.




Fonte: Elaborada pelo autor (2020).

Figura 77 - Distribuição das propriedades para cada grupo determinado pelo procedimento computacional - Poço A.



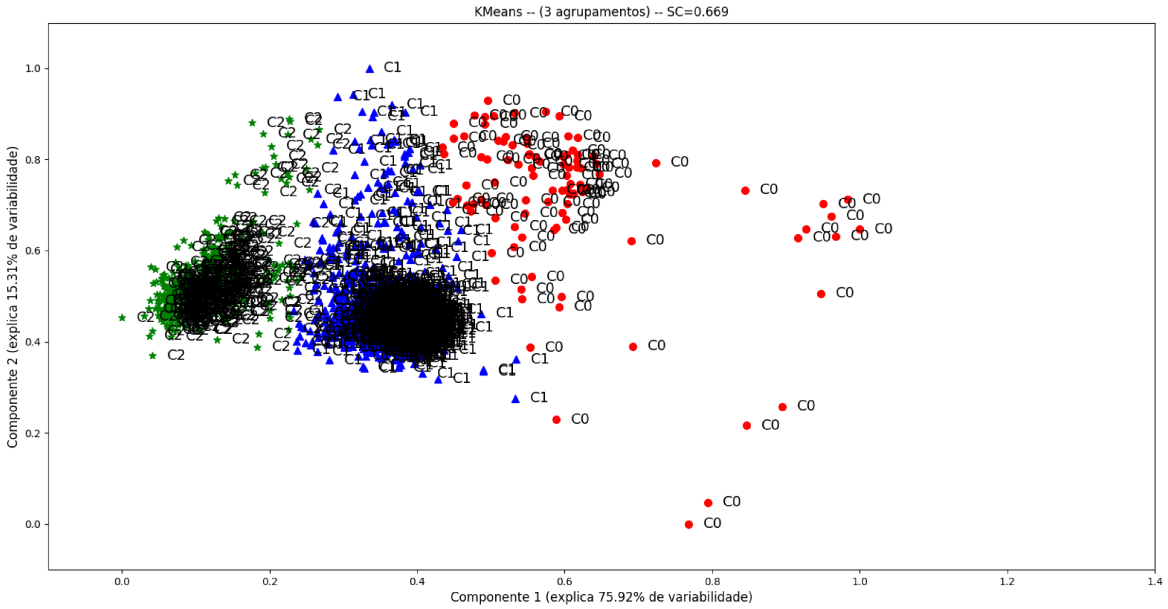

Fonte: Elaborada pelo autor (2020).

Tabela 49 – Análise de variância intergrupos para cada registro - Poço A. Os valores com * indicam os p-valores que rejeitaram a $H_0$. Os grupos foram determinados pelo procedimento computacional. O nível de significância é de 0.05.

| Características | $H_{cal}$ | p-valor |
|---|---|---|
| CALI | 331.952378 | 8.269076E-73* |
| CNST | 17.380082 | 1.682531E-04* |
| DRHO | 330.315620 | 1.874453E-72* |
| FCNL | 159.214816 | 2.672649E-35* |
| FFDC | 1162.285763 | 4.100655E-253* |
| GR | 437.028519 | 1.260268E-95* |
| NCNL | 194.169524 | 6.864731E-43* |
| NPHI | 21.057391 | 2.675751E-05* |
| RHOB | 967.062855 | 1.011508E-210* |

Fonte: Elaborada pelo autor (2020).

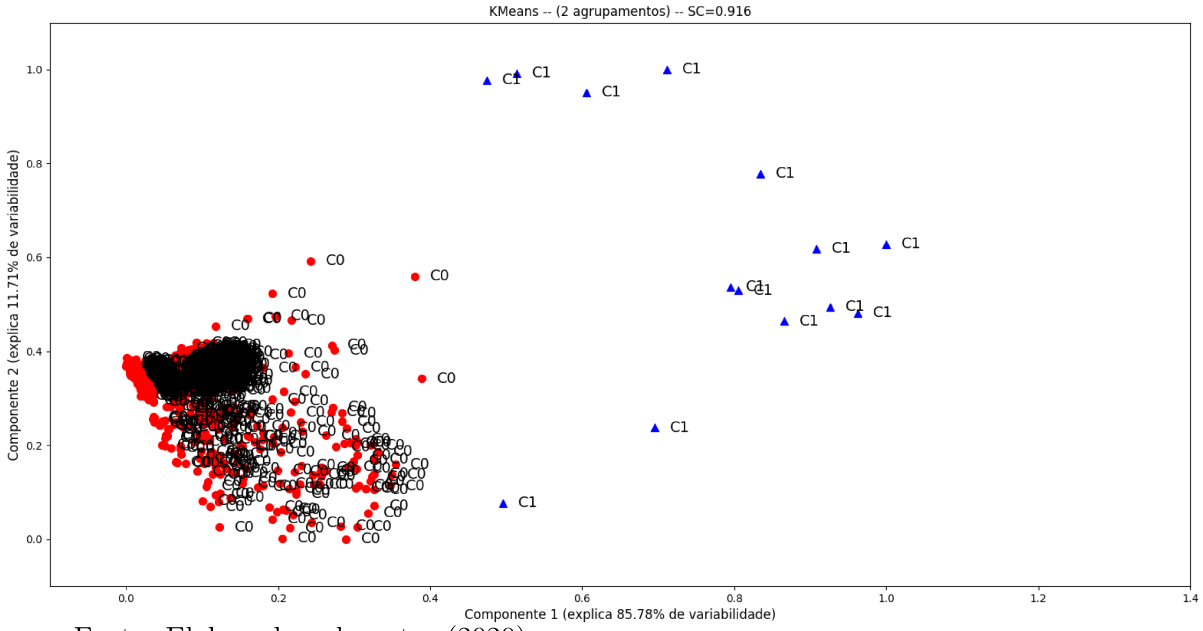
Figura 78 - Resultado K-Means - Poço A. SC = 0.669




Fonte: Elaborada pelo autor (2020).

A Figura 79 mostra o resultado do K-Means aplicado nos dados após o pré-processamento o acréscimo das características polinomiais. O grau do polinômio foi 4 e os dados processados ficaram com 1001 características. A primeira componente principal expressa 71.26% de variabilidade dos dados, a segunda expressa 25.43% da variabilidade. De acordo com o que é mostrado na figura, pode-se observar que houve um aumento considerável no valor do coeficiente de silhueta em relação aos dados originais.

Figura 79 - Resultado K-Means (Características Polinomiais) - Poço A. SC = 0.916

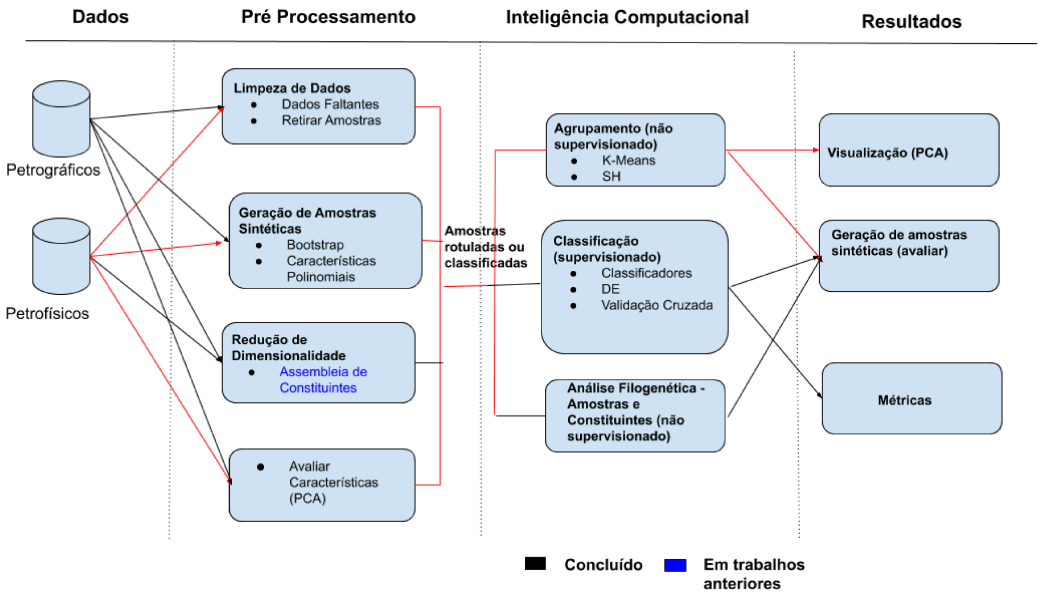



Fonte: Elaborada pelo autor (2020).

### 5.3.10 Dados Petrofísicos - Poço B

A Figura 80 mostra, através das linhas vermelhas, quais procedimentos da metodologia descritos na Seção 5.2.

Figura 80 - Fluxograma ilustrando, através das linhas vermelhas, a metodologia aplicada a base de dados Poço B.



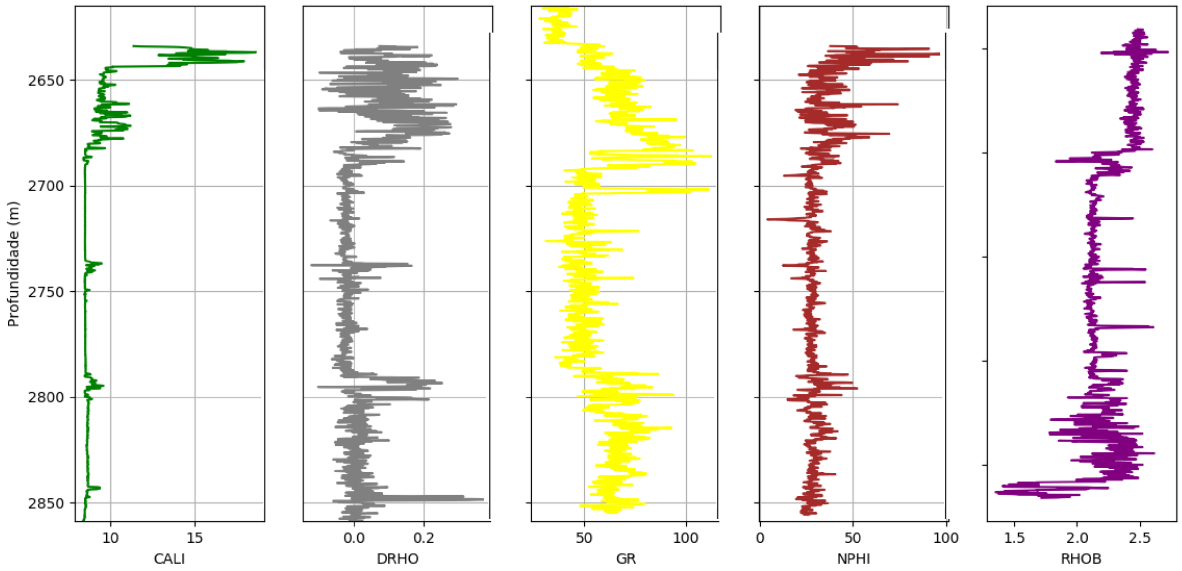

Fonte: Elaborada pelo autor (2020).

A Figura 81 apresenta as propriedades em relação a profundidade. Observa-se que há um aumento nos valores das características entre as profundidades 2650 e 2700 metros. As correlações entre as variáveis registradas em função da profundidade ao longo do processo de perfuração são mostradas na Figura 82. Nota-se que há uma alta correlação entre as características NPHI e CALI e entre RHOB e CALI o que mostra que o diâmetro do furo de sondagem está relacionado com a porosidade.

Na Tabela 50 encontram-se as componentes obtidas com o procedimento de Análise de Componentes Principais. Para a Componente 1 as características GR e NPHI influenciam mais em relação às outras, enquanto na Componente 2 são as características CALI, GR e NPHI que se destacam. O GR é uma das característica que mais contribui para identificação de litologia e o NPHI está relacionado a porosidade que é um fator importante para definir o potencial de recuperação de hidrocarbonetos de uma rocha. A Figura 83 apresenta a distribuição das amostras, onde nota-se a existência de um grupo bem denso.

A Tabela 51 apresenta a média, o valor mínimo e o máximo para cada característica de acordo com as três classes litológicas geradas. Dentre as características que podem-

Figura 81 - Propriedades em relação a profundidade para cada propriedade respectivamente - Poço B.




Fonte: Elaborada pelo autor (2020).

Tabela 50 – Componentes Principais em relação as propriedades petrofísicas. Os valores com * indicam os valores mais significativos obtidos para cada componente sugerindo qual propriedade petrofísica tem mais atuação sobre as mesmas - Poço B.

| Características | Componente 1 | Componente 2 |
|---|---|---|
| CALI | -8.59859881E-02 | 1.90843845E-01* |
| DRHO | -1.54564864E-03 | 2.42427337E-04 |
| GR | -8.68519511E-01* | -4.95212940E-01* |
| NPHI | -4.87789005E-01* | 8.47267463E-01* |
| RHOB | -1.84327654E-02 | 2.19216548E-02 |

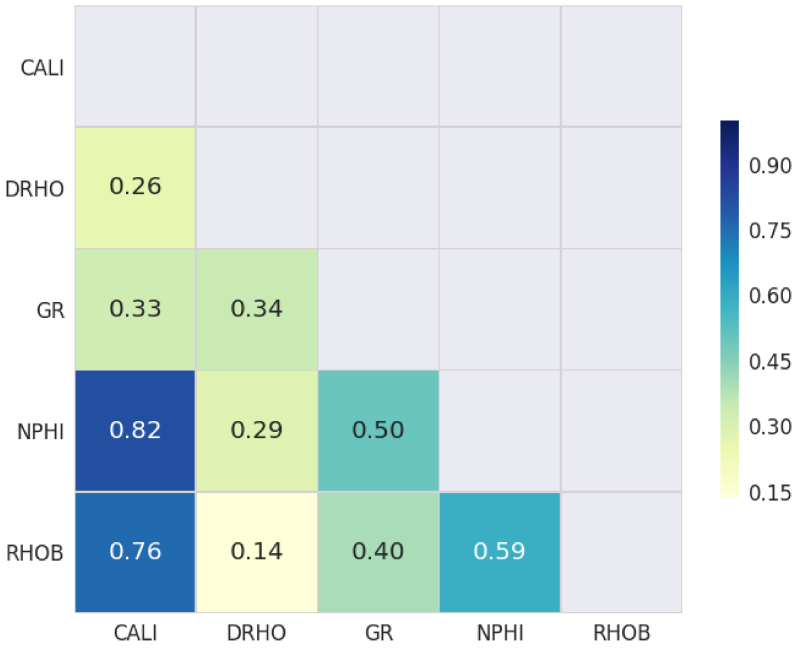
Fonte: Elaborada pelo autor (2020).

se notar diferença entre os grupos estão DRHO, GR e NPHI que são características relacionadas a porosidade e a densidade de uma rocha.

De acordo com a Tabela 51, para G0 DRHO tem valor entre aproximadamente -0.0380 e 0.2924, GR entre 46.5937 e 73.2363 e NPHI entre 40.6005 e 96.6796. Para G1 DRHO tem valor entre aproximadamente -0.1210 e 0.3134, GR entre 30.8828 e 62.3004 e NPHI entre 7.2265 e 40.3137. Para G2 DRHO tem valor entre aproximadamente -0.1015 e 0.3717, GR entre 60.9218 e 113.0 e NPHI entre 15.3770 e 58.8989. Para G3 DRHO tem valor entre aproximadamente -0.0406 e 0.0683, GR entre 0.0 e 57.6543 e NPHI entre 0.0 e 2.5628.

A Figura 84 mostra o comportamento de cada propriedade em relação a profundi-

Figura 82 - Matriz de correlação entre os registros coletados - Poço B.




Fonte: Elaborada pelo autor (2020).

dade para os grupos encontrados pelo K-Means. Percebe-se que maioria das amostras do grupo G0 estão localizadas em profundidades abaixo de 2650 metros. Em relação ao G1 estão entre 2600 e 2850. As amostras associadas ao grupo G2 concentram-se entre 2600 e 2850. Tratando-se de G3 as amostras estão entre 2700 e 2800.

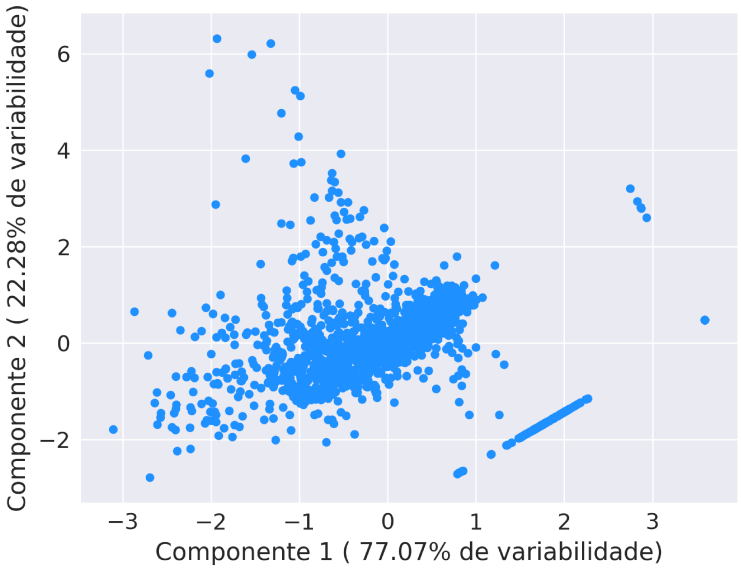
A Figura 85 mostra a distribuição das propriedades em função dos grupos determinados pela metodologia proposta. A mediana dos quatro grupos estão próximas quando se trata de DRHO e RHOB.

A Tabela 52 apresenta os resultados da análise de variância intergrupos para cada característica. Pode-se observar que $H_0$ é rejeitada para todas as características. O $H_{cal}$ é o valor calculado pelo método de *Kruskal-Wallis*, nos caso em que p-valor $< 5\%$, $H_{cal}$ é maior que o valor tabelado na Distribuição Qui-Quadrado.

A Figura 86 exibe o resultado do K-Means. A primeira componente principal expressa 77.07% de variabilidade dos dados, a segunda expressa 22.28% da variabilidade.

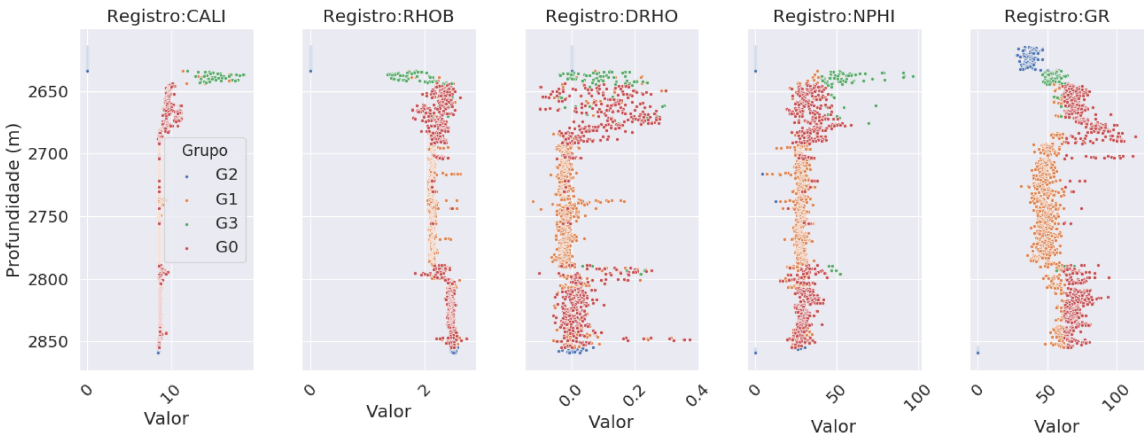
A Figura 87 exibe o resultado do K-Means aplicado nos dados após o pré-processamento o acréscimo das características polinomiais. O grau do polinômio foi 4 e os dados ficaram com 126 características. A primeira componente principal expressa 90.49% de variabilidade dos dados, a segunda expressa 8.42% da variabilidade. Pode-se observar que houve um aumento no valor do coeficiente de silhueta em relação aos dados originais.

Figura 83 - Distribuição das Amostras - Poço B. A componente principal 1 expressa 77.07% da variabilidade dos dados e a componente principal 2 22.28%.




Fonte: Elaborada pelo autor (2020).

Figura 84 - Propriedades em relação a profundidade para cada grupo respectivamente - Poço B.



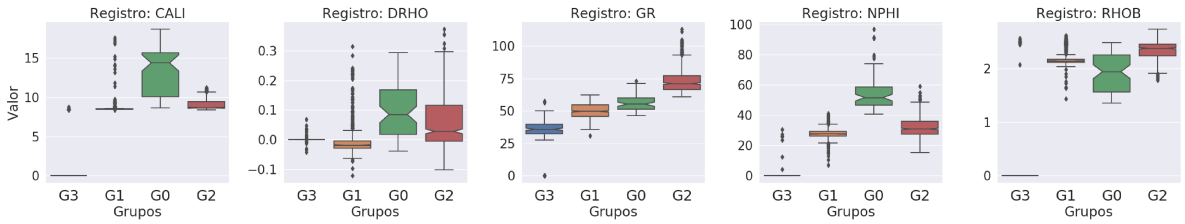

Fonte: Elaborada pelo autor (2020).

Tabela 51 – Grupos encontrados pelo K-Means. Para cada grupo tem-se a média de cada característica petrofísica e da profundidade. O número entre parênteses é o número de amostras atribuídas a cada grupo. O * indica as características que se diferenciaram entre os grupos - Poço B.

| Grupo | Logname | Média | Mínimo | Máximo |
|---|---|---|---|---|
| G0(77) | CALI | 13.4586 | 8.6298 | 18.6684 |
| | DRHO* | 0.0959 | -0.0380 | 0.2924 |
| | GR* | 56.2860 | 46.5937 | 73.2363 |
| | NPHI* | 54.5451 | 40.6005 | 96.6796 |
| | RHOB | 1.9296 | 1.3492 | 2.4826 |
| G1(792) | CALI | 8.6672 | 8.3994 | 17.0683 |
| | DRHO | -0.0057 | -0.1210 | 0.3134 |
| | GR | 50.1998 | 30.8828 | 62.3004 |
| | NPHI | 27.7404 | 7.2265 | 40.3137 |
| | RHOB | 2.1850 | 1.4314 | 2.611 |
| G2(579) | CALI | 9.0504 | 8.4087 | 11.2055 |
| | DRHO | 0.0593 | -0.1015 | 0.3717 |
| | GR | 73.3168 | 60.9218 | 113.0 |
| | NPHI | 31.9885 | 15.3770 | 58.8989 |
| | RHOB | 2.3241 | 1.7795 | 2.7285 |
| G3(153) | CALI | 1.5502 | 0.0 | 8.7407 |
| | DRHO | 0.0013 | -0.0406 | 0.0683 |
| | GR | 31.4844 | 0.0 | 57.6543 |
| | NPHI | 1.1514 | 0.0 | 30.4809 |
| | RHOB | 0.4535 | 0.0 | 2.5628 |

Fonte: Elaborada pelo autor (2020).

Figura 85 - Distribuição das propriedades para cada grupo determinado pelo procedimento computacional - Poço B.



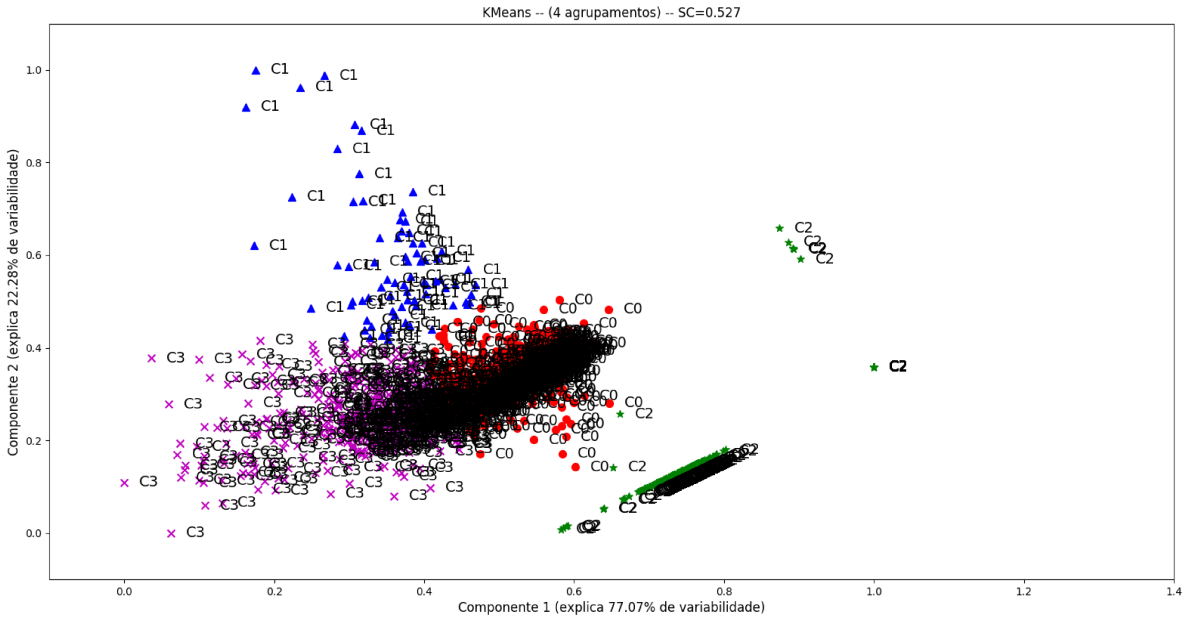

Fonte: Elaborada pelo autor (2020).

Tabela 52 – Análise de variância intergrupos para cada característica - Poço B. Os valores com * indicam os p-valores que rejeitaram a $H_0$. Os grupos foram determinados pelo procedimento computacional. O nível de significância é de 0.05.

| Característica | $H_{cal}$ | p-valor |
|---|---|---|
| CALI | 867.859380 | 8.286943E-188* |
| DRHO | 431.770551 | 2.902586E-93* |
| GR | 1245.328754 | 1.072080E-269* |
| NPHI | 708.612831 | 2.847312E-153* |
| RHOB | 441.065357 | 2.812208E-95* |

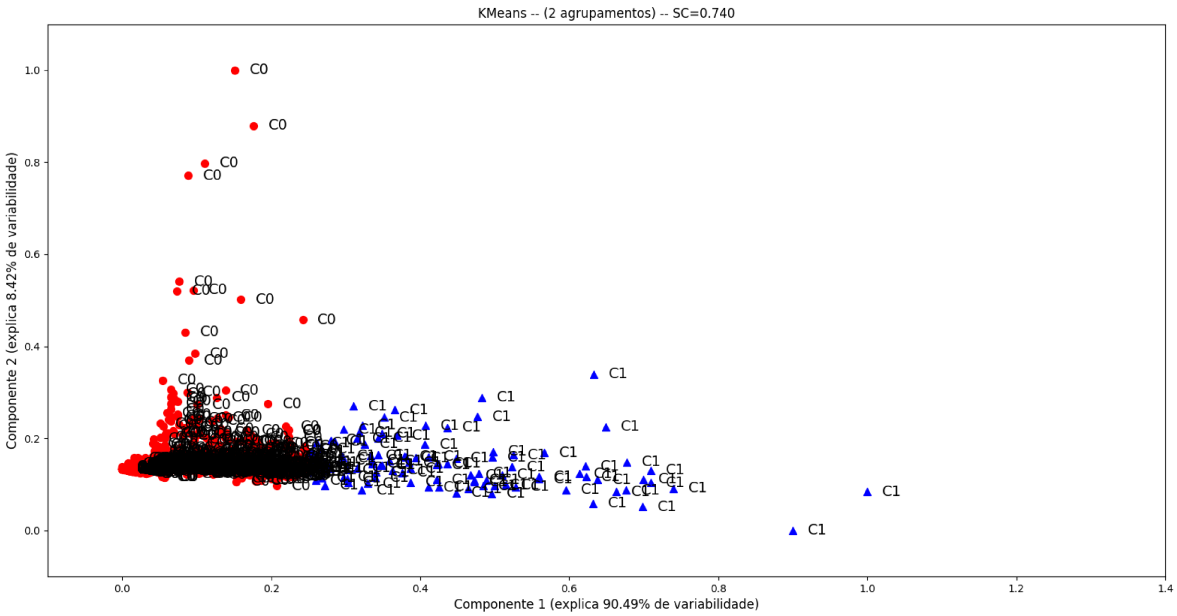
Fonte: Elaborada pelo autor (2020).

Figura 86 - Resultado K-Means - Poço B. SC = 0.527




Fonte: Elaborada pelo autor (2020).

Figura 87 - Resultado K-Means (Características Polinomiais) - Poço B. SC = 0.740

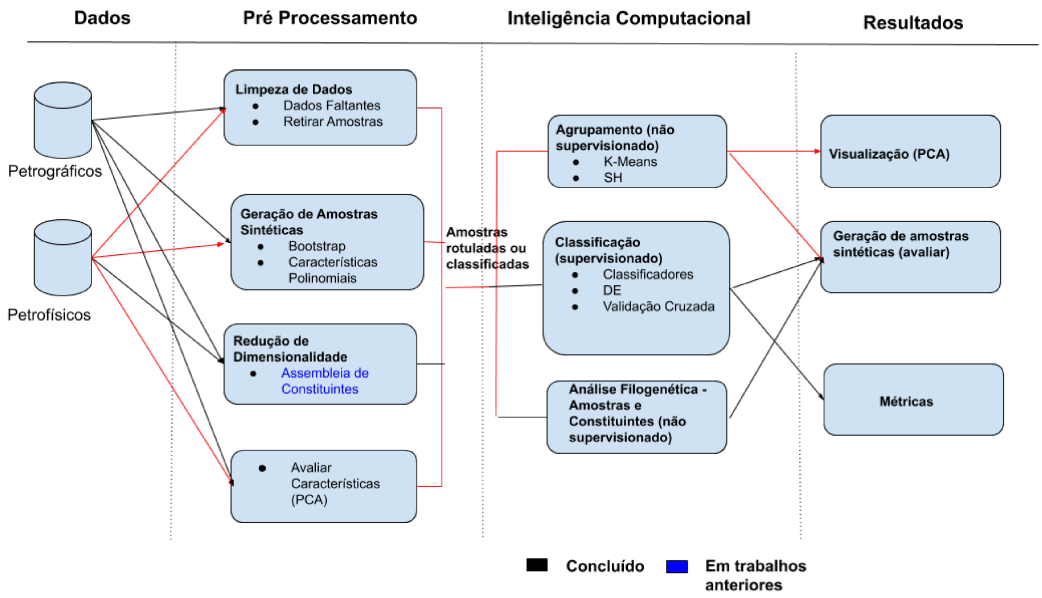



Fonte: Elaborada pelo autor (2020).

### 5.3.11 Dados Petrofísicos - Poço C

A Figura **??** mostra, através das linhas vermelhas, quais procedimentos da metodologia descritos na Seção 5.2.

Figura 88 - Fluxograma ilustrando, através das linhas vermelhas, a metodologia aplicada a base de dados Poço C.



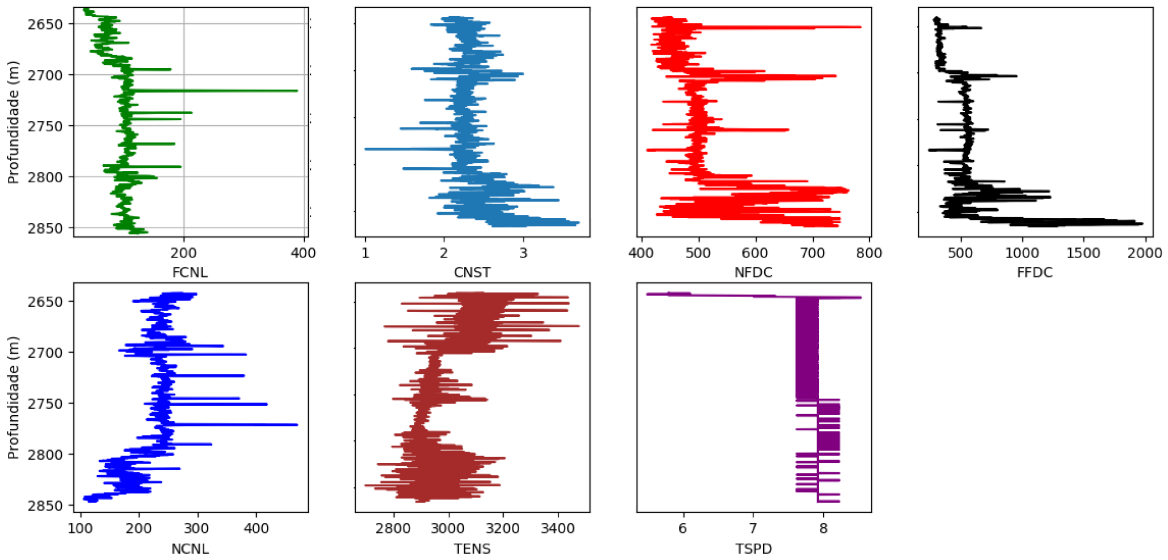

Fonte: Elaborada pelo autor (2020).

A Figura 89 apresenta como as propriedades se comportam em relação a profundidade. A Figura 90 mostra a correlação entre as variáveis registradas em função da profundidade ao longo do processo de perfuração. Nota-se que há uma alta correlação entre as características NCNL e FCNL o que mostra que há relação entre porosidade da rocha medida através dos dois detectores, próximo e longe.

Na Tabela 53 encontram-se as componentes obtidas. Para a Componente 1 as características FFDC, NFDC e TENS influenciam mais, na Componente 2 são as características FFDC, NCNL e TENS. Estes atributos estão relacionados a porosidade e a densidade de uma rocha. A Figura 91 apresenta a distribuição das amostras, nota-se que há um grupo bem denso.

A Tabela 54 apresenta a média, o valor mínimo e o máximo para cada característica de acordo com as três classes litológicas geradas. Dentre as características que podem-se notar diferença entre os grupos estão FCNL, FFDC, NCNL e NFDC o mostra que a densidade e a porosidade entre os grupos diferem sendo fatores importantes para a separação desses.

Figura 89 - Propriedades em relação a profundidade para cada propriedade respectivamente - Poço C.




Fonte: Elaborada pelo autor (2020).

Tabela 53 – Componentes Principais em relação as propriedades petrofísicas. Os valores com * indicam os valores mais significativos obtidos para cada componente sugerindo qual propriedade petrofísica tem mais atuação sobre as mesmas - Poço C.

| Característica | Componente 1 | Componente 2 |
|---|---|---|
| CNST | 6.52896420E-04 | -8.50663466E-05 |
| FCNL | -5.18495475E-02 | -7.97377872E-02 |
| FFDC | 9.47481359E-01* | 1.46313119E-01* |
| NCNL | -8.76867506E-02 | -1.42927710E-01* |
| NFDC | 2.43544006E-01* | 8.28758301E-02 |
| TENS | -1.80520159E-01* | 9.72078705E-01* |
| TSPD | 3.08314674E-04 | -7.79870506E-04 |

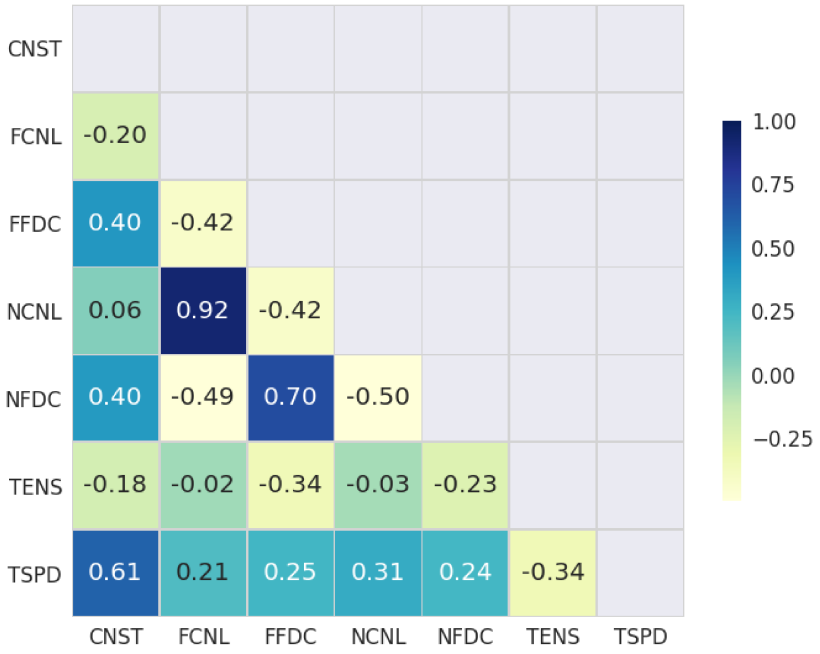
Fonte: Elaborada pelo autor (2020).

De acordo com a Tabela 54, para G0 FCNL tem valor entre aproximadamente 0.0 e 389.3711, FFDC entre 244.1875 e 876.5, NCNL entre 0.0 e 470.375 e NFDC entre 410.375 e 784.5156. Para G1 FCNL tem valor entre aproximadamente 32.8046 e 82.1406, FFDC entre 870.75 e 1976.406, NCNL entre 106.25 e 233.125 e NFDC entre 612.5 e 757.9609.

A Figura 92 mostra o comportamento de cada propriedade em relação a profundidade para os grupos encontrados pelo K-Means. Percebe-se que maioria das amostras do grupo G0 estão localizadas em profundidades abaixo de 2650 metros. Em relação ao G1 estão entre 2650 e 2850.

A Figura 93 mostra a distribuição das propriedades em função dos grupos determi-

Figura 90 - Matriz de correlação entre os registros coletados - Poço C.




Fonte: Elaborada pelo autor (2020).

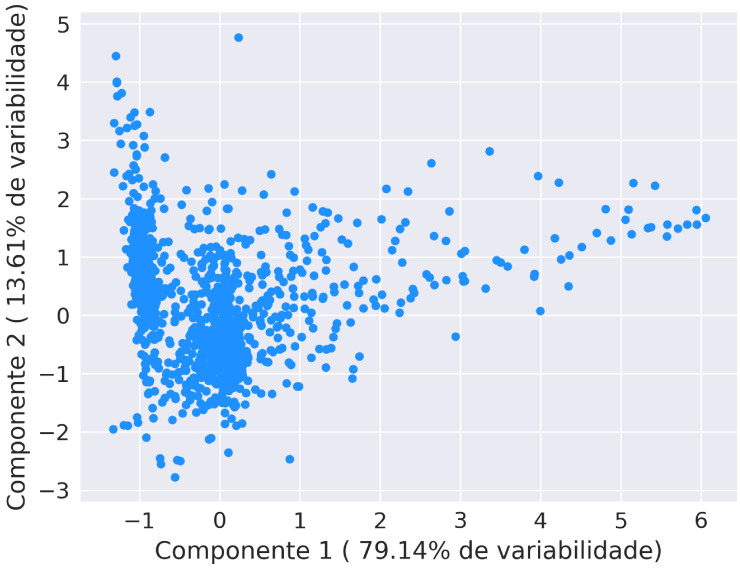
nados pela metodologia proposta. A mediana dos dois grupos estão próximas quando se trata de TENS e TSPD.

A Tabela 55 apresenta os resultados da análise de variância intergrupos para cada característica. Pode-se observar que $H_0$ é rejeitada para todas as características. O $H_{cal}$ é o valor calculado pelo método de *Kruskal-Wallis*, nos caso em que p-valor $< 5\%$, $H_{cal}$ é maior que o valor tabelado na Distribuição Qui-Quadrado.

A Figura 94 exibe o resultado do K-Means. A primeira componente principal expressa 79.14% de variabilidade dos dados, a segunda expressa 13.61% da variabilidade.

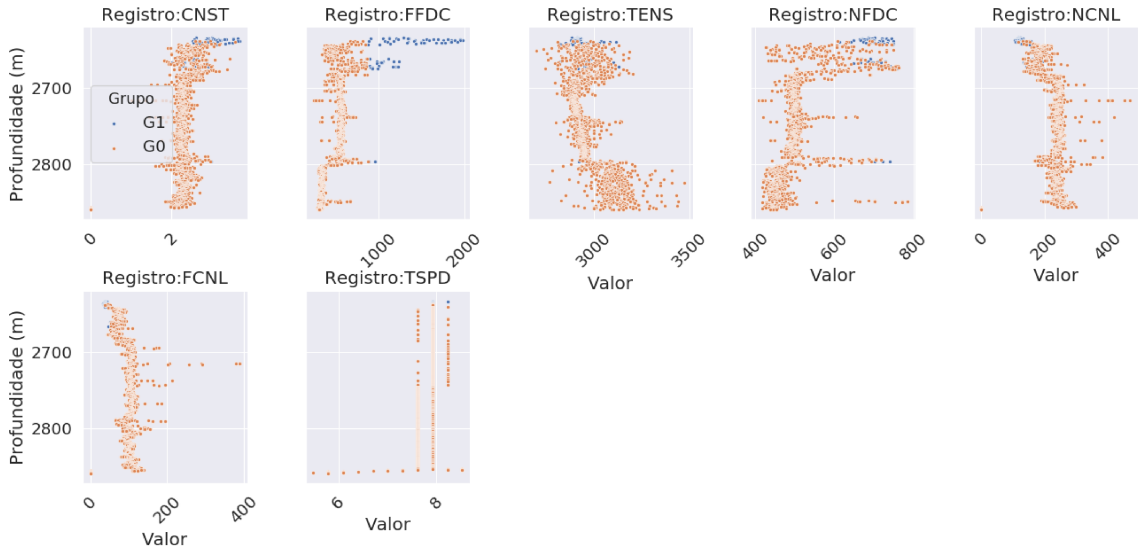
A Figura 95 exibe o resultado do K-Means aplicado nos dados após o pré-processamento o acréscimo das características polinomiais. O polinômio encontrado apresentou grau 4 e os dados processados ficaram com 330 características. A primeira componente principal expressa 71.26% de variabilidade dos dados, a segunda expressa 25.43% da variabilidade. Nesse caso, houve um decréscimo no valor do coeficiente de silhueta provavelmente em função da baixa correlação entre as características originais.

Figura 91 - Distribuição das Amostras - Poço C. A componente principal 1 expressa 79.14% da variabilidade dos dados e a componente principal 2 22.28%.




Fonte: Elaborada pelo autor (2020).

Figura 92 - Propriedades em relação a profundidade para cada grupo respectivamente - Poço C.



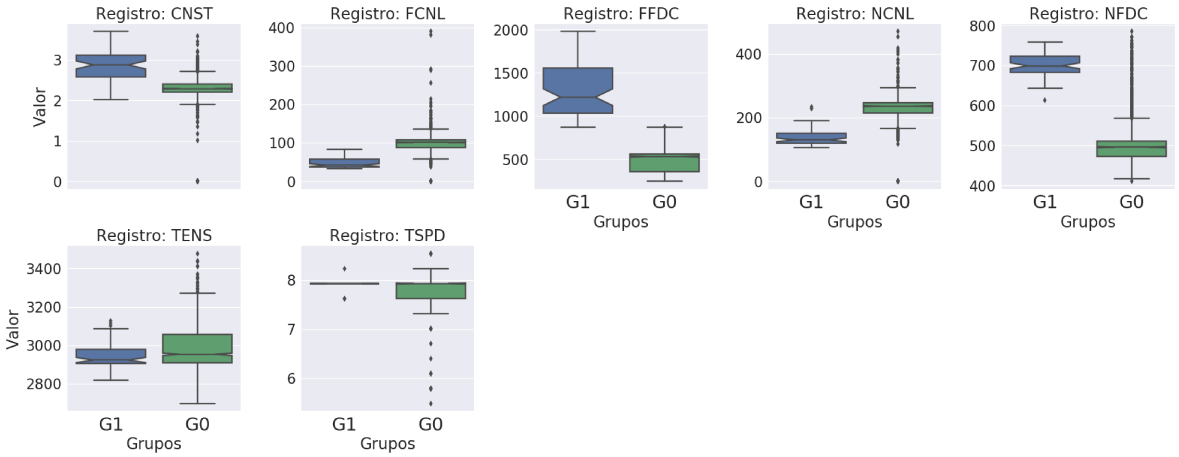

Fonte: Elaborada pelo autor (2020).

Tabela 54 – Grupos encontrados pelo K-Means. Para cada grupo tem-se a média de cada característica petrofísica e da profundidade. O número entre parênteses é o número de amostras atribuídas a cada grupo. O * indica as características que se diferenciaram entre os grupos - Poço C.

| Grupo | Logname | Média | Mínimo | Máximo |
|---|---|---|---|---|
| G0(1403) | CNST | 2.2868 | 0.0 | 3.5800 |
| | FCNL* | 97.1219 | 0.0 | 389.3711 |
| | FFDC* | 487.8023 | 244.1875 | 876.5 |
| | NCNL* | 226.1240 | 0.0 | 470.375 |
| | NFDC* | 508.9723 | 410.375 | 784.5156 |
| | TENS | 2986.4202 | 2698.03 | 3476.084 |
| | TSPD | 7.7805 | 5.4864 | 8.5344 |
| G1(73) | CNST | 2.8572 | 2.0180 | 3.6982 |
| | FCNL | 46.9892 | 32.8046 | 82.1406 |
| | FFDC | 1310.3705 | 870.75 | 1976.406 |
| | NCNL | 137.6453 | 106.25 | 233.125 |
| | NFDC | 701.1280 | 612.5 | 757.9609 |
| | TENS | 2944.4156 | 2819.725 | 3127.775 |
| | TSPD | 7.9164 | 7.62 | 8.2296 |

Fonte: Elaborada pelo autor (2020).

Figura 93 - Distribuição das propriedades para cada grupo determinado pelo procedimento computacional - Poço C.



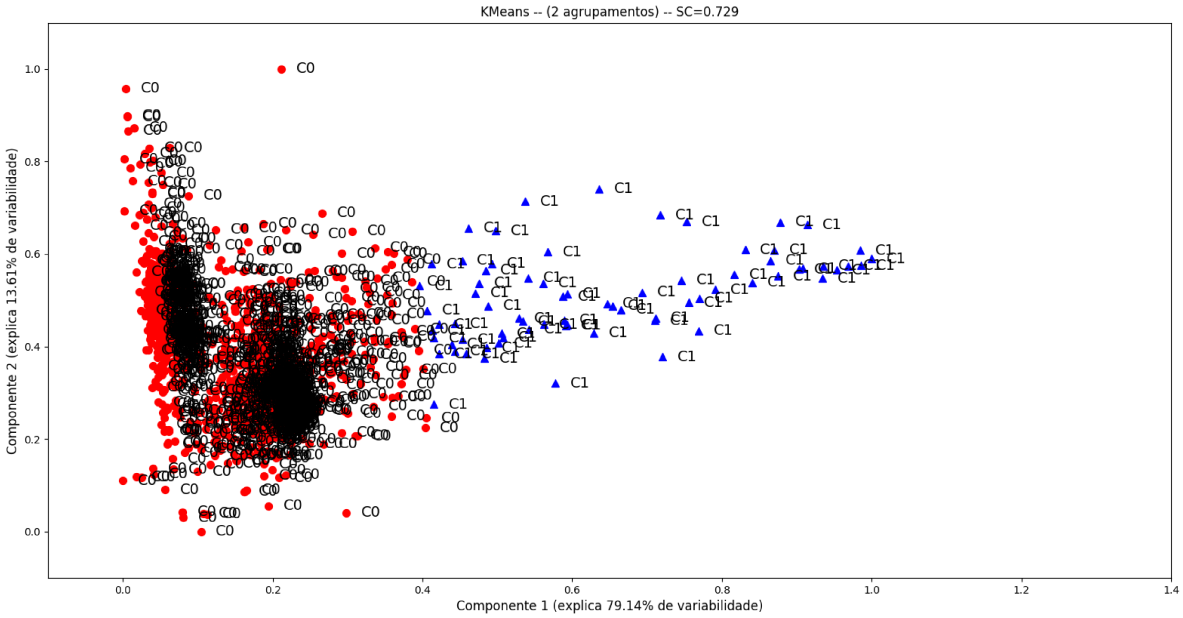

Fonte: Elaborada pelo autor (2020).

Tabela 55 – Análise de variância intergrupos para cada registro - Poço C. Os valores com * indicam os p-valores que rejeitaram a $H_0$. Os grupos foram determinados pelo procedimento computacional. O nível de significância é de 0.05.

| Característica | $H_{cal}$ | p-valor |
|---|---|---|
| CNST | 120.826693 | 4.170160E-28* |
| FCNL | 179.686999 | 5.672331E-41* |
| FFDC | 208.019669 | 3.714732E-47* |
| NCNL | 175.852939 | 3.898964E-40* |
| NFDC | 180.548395 | 3.678616E-41* |
| TENS | 9.893159 | 1.658943E-03* |
| TSPD | 35.818736 | 2.165554E-09* |

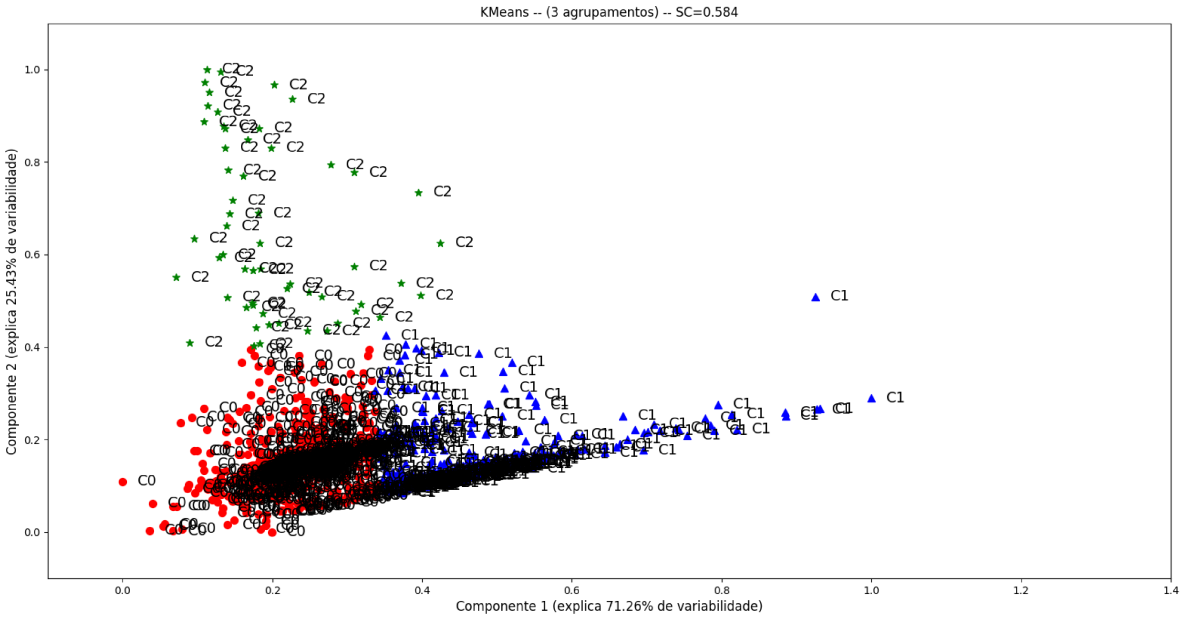
Fonte: Elaborada pelo autor (2020).

Figura 94 - Resultado K-Means - Poço C. SC = 0.728




Fonte: Elaborada pelo autor (2020).

Figura 95 - Resultado K-Means (Características Polinomiais) - Poço C. SC = 0.584




Fonte: Elaborada pelo autor (2020).

# 6 CONCLUSÕES

Neste trabalho foi proposto uma metodologia computacional para auxiliar na caracterização de reservatórios petrolíferos. A metodologia é dividida em duas partes: módulo supervisionado e módulo não supervisionado. O procedimento supervisionado consiste na aplicação de um algoritmo evolutivo, nesse caso escolheu-se o algoritmo de Evolução Diferencial pela simplicidade de implementação e pelo desempenho relatado na literatura para encontrar os parâmetros ótimos para os classificadores. Para classificar empregou-se os métodos Rede Neural Artificial (ANN), Árvore de Decisão (DT), Máquina de Aprendizado Extremo (ELM), Gradient Boosting (GB), K-Vizinhos mais próximos (KNN) e Máquina de Vetor Suporte (SVM). Usou-se o Bootstrap juntamente com o K-Means nas bases e dados Tibagi e Paleosul, este procedimento foi incluso na parte supervisionada por necessitar das classes previamente. O procedimento não supervisionado consiste na aplicação de duas estratégias: K-Means + ACP e Características Polinomiais + K-Means. Este procedimento foi empregado nas Poço A, Poço B e Poço C. Ademais, iniciou-se um estudo, na base de dados Tibagi, do uso de Análise Filogenética como forma de entender a diagênese.

Ao analisar a parte supervisionada pode-se perceber que a ANN e o GB são classificadores que alcançaram altos índices de precisão para classificar dados petrográficos e petrofísicos, respectivamente. Além disso, observa-se que o DE é uma opção eficiente para encontrar parâmetros dos modelos de aprendizado de máquina. A integração do algoritmo evolutivo com os modelos de aprendizado de máquina mostrou-se uma estratégia eficiente para a classificação de litologia. O método $k$-fold ($k = 5$) é usado como critério de particionamento de dados para testar e treinar a separação de conjuntos de dados. Em relação ao GB, embora seja considerado robusto para o ajuste, gasta bastante tempo para realizar o treinamento do modelo, especialmente quando se trata de grandes bancos de dados.

O Bootstrap é uma alternativa eficaz para encontrar os parâmetros do K-Means quando as característica da base de dados original possuem correlação alta. No caso da base Tibagi não houve alteração no valor do coeficiente de silhueta encontrado, pois os parâmetros sugeridos pelo Bootstrap foram similares ao se utilizar a base de dados original. Já utilizando apenas os constituintes diagenéticos houve um aumento no coeficiente de silhueta e quando se retira as amostras pertencentes as petrofácies I-1 e I-2 há um aumento mais significativo no valor da silhueta, o que mostra que as amostras que foram retiradas interferiam na correlação. Quando utilizou-se apenas os constituintes diagenéticos na base de dados sem I-1 e I-2 o oposto aconteceu, o valor da silhueta diminuiu. Em relação a base Paleosul o valor da silhueta diminuiu, quando aplicou-se somente nos constituintes diagenéticos o valor aumentou.

A parte não supervisionada quando aplicada nos dados petrofísicos torna-se uma ferramenta para auxiliar na caracterização de litologia, determinando possíveis classes petrofísicas. Para o Poço A na primeira estratégia (K-Means + ACP) três classes petrofísicas foram geradas onde foram diferenciadas pelas características DRHO, NFDC e NPHI, uma dentre estas se mostrou influente nas componentes principais. Ao aplicar o teste Kruskal-Wallis, foi constatado que a média dos grupos são diferentes, ou seja, há heterogeneidade entre eles. Na segunda estratégia (Características Polinomiais + K-Means) duas classes petrofísicas foram geradas e o valor do coeficiente de silhueta aumentou consideravelmente. Para o Poço B na primeira estratégia foram obtidas quatro classes petrofísicas que se distinguiram pelos atributos DRHO, GR e NPHI, duas dessas se mostraram importantes nas componentes principais. Pelo Kruskal-Wallis constatou que a média dos grupos encontrados são diferentes. Na segunda estratégia duas classes petrofísicas foram obtidas e o valor da silhueta aumentou em comparação com os dados originais. No Poço C na primeira estratégia três classes petrofísicas foram encontradas que foram particularizadas pela diferença entre as características FCNL, FFDC, NCNL e NFDC, duas destas mostraram ter influência sobre as componentes principais. Na segunda estratégia ocorreu um decréscimo no valor da silhueta, acredita-se que foi pela correlação baixa entre as características.

Nos dados petrográficos foram empregadas nas bases Tibagi e Paleosul. Na base Tibagi na primeira estratégia foram identificadas três petrofácies. As características que expressam maior importância nas componentes principais são Bioclasto, Porosidade Intergranular e Opaco. Ao aplicar o teste Kruskal-Wallis, constatou-se que para algumas características a média dos grupos são consideradas estatisticamente iguais. Na segunda estratégia três petrofácies foram geradas e o valor do coeficiente de silhueta aumentou. Na base Paleosul na primeira estratégia duas petrofácies foram obtidas. As características que expressam maior influência nas componentes principais são Minerais Pesados, Mica e Pirita. Ao aplicar o teste Kruskal-Wallis, constatou-se que para algumas características a média dos grupos são consideradas estatisticamente iguais. Na segunda estratégia duas petrofácies foram encontradas e o valor da silhueta aumentou consideravelmente. Outras técnicas de validação de agrupamentos devem ser avaliadas para proporcionar um resultado mais robusto.

A Análise Filogenética foi aplicada nas bases Tibagi e Paleosul. Duas abordagens foram utilizadas: amostras e constituintes diagenéticos. Em relação às amostras foi possível visualizar a distribuição das amostras e as respectivas petrofácies indicando as amostras que apresentaram processos diagenéticos semelhantes. Alguns resultados concordaram com as análises Intra-Poço e Inter-Poço. Tratando-se dos constituintes foi possível identificar os eventos que ocorreram nas fases eodiagenética, mesodiagenética e telodiagenética. Os eventos encontrados foram semelhantes aos que foram propostos na literatura.

A abordagem proposta permitiu avaliar o desempenho da análise de agrupamentos

na identificação de classes petrográficas que representam diferentes petrofácies e dessa forma, surge como uma alternativa para assistir o geólogo/especialista na determinação e caracterização da qualidade do reservatório, contribuindo para a redução de custo e aumento na precisão do modelo geológico final.

Com base na série de experimentos realizados, pode se concluir que sistema computacional proposto e desenvolvido nesta tese pode auxiliar o geólogo/petrólogo na tarefa de caracterização de reservatórios, seja em dados petrográficos ou petrofísicos. A abordagem computacional tem o potencial de assistir na identificação e classificação de petrofácies e classes litológicas. O sistema permite também avaliar as características através das análises oferecidas reduzindo o tempo de análise em comparação com o processo manual. A análise filogenética pode auxiliar na compreensão do processo diagenético e na classificação das petrofácies sedimentares. Além disso, os especialistas podem aplicar os métodos para analisar um banco de dados de alta dimensão durante a exploração geológica, o que também proporciona uma melhoria na eficiência da avaliação econômica de áreas potencialmente produtoras.

## 6.1 PERSPECTIVAS DE TRABALHOS FUTUROS

Apresenta-se, a seguir, uma relação dos desenvolvimentos previstos para a continuidade da pesquisa. Entre os principais desenvolvimentos destacam-se:

- Análise filogenética: aplicar em outras bases de dados..
- Aplicar a metodologia em uma base de dados com grande número de amostras que tenha classificação.
- Melhoria contínua do arcabouço computacional.

# REFERÊNCIAS

1 FOLK, R. L. *Petrology of sedimentary rocks*. [S.l.]: Hemphill Publishing Company, 1957.

2 XIE, Y. et al. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *Journal of Petroleum Science and Engineering*, Elsevier, v. 160, p. 182–193, 2018.

3 FOURNIER, F.; BORGOMANO, J. Critical porosity and elastic properties of microporous mixed carbonate-siliciclastic rocks. *Geophysics*, Society of Exploration Geophysicists, v. 74, n. 2, p. E93–E109, 2009.

4 FERNANDES, L. Petróleo e gás natural. *Departamento Nacional de Produção Mineral*, 2009.

5 ROSA, S. E. S. d.; GOMES, G. Pico de hubbert e o futuro da produção mundial de petróleo. Banco Nacional de Desenvolvimento Econômico e Social, 2004.

6 CEPAL, N. Recursos naturais na união das nações sul-americanas (unasul): Situação e tendências para uma agenda de desenvolvimento regional. CEPAL, 2013.

7 COELHO, J. M. F. José mauro ferreira coelho. Empresa de Pesquisa Energética, n. 2, 2017.

8 VIEIRA, P. A.; BUAINAIN, A. M.; FIGUEIREDO, E. V. C. O brasil alimentará a china ou a china engolirá o brasil? *Revista Tempo do Mundo*, v. 2, n. 1, p. 51–81, 2016.

9 BIROL, F. Gas 2018: Analysis and forecasts to 2023. *Market Report Series*, International Energy Agency, 2018.

10 SILVA, D. C. da; SILVA, R. S. M. da; GONçALO, V. M. C. *Boletim de P&D*. [S.l.], 2013.

11 PONTES, C. de S. Boletim anual de exploração e produção de petróleo e gás. *Boletim SBGf*, Ministério de Minas e Energia, 2018.

12 MADEIRA, T. P. Boletim da produção de petróleo e gás natural. Superintendência de Desenvolvimento e Produção, n. 112, 2019.

13 SILVA, F. V. da. Bases de dados na indústria do petróleo e gás. *Boletim SBGf*, Sociedade Brasileira de Geofísica, n. 2, 2011.

14 XAVIER, A. R.; FARIA, G. F. *Boletim SBGf*. [S.l.], 2011.

15 YAMAMOTO, T. M. *Métodos de Determinação de Geóide*. Dissertação (Mestrado) — Universidade Federal do Paraná, 1983.

16 SILVA, C. G. et al. Bampetro–banco de dados ambientais para a indústria de petróleo. In: *8th International Congress of the Brazilian Geophysical Society*. [S.l.: s.n.], 2003.

17 SYSTEMES, . D. Digitalrock product brief. *The 3D Experience Company*, 2018.

18 DIGITAL, . B. Petrovisual. *LLC*, 2018.

19 MARASCHIN, A. J.; MIZUSAKI, A. M. Datação de processos diagenéticos em arenitos-reservatório de hidrocarbonetos: Uma revisão conceitual. v. 35(1), p. 27–41, 2008.

20 MENEZES, M. R. F. de. *Estudos sedimentológicos e o contexto estrutural da formação Serra do Martins, nos platôs de Portaalegre, Martins e Santana, RN*. Dissertação (Mestrado) — Programa de Pesquisa e Pós-Graduação em Geodinâmica e Geofísica da UFRN, Natal, 1999.

21 CEVOLANI, J. T. et al. Visualização e classificação automática de petrofácies sedimentares. In: *6º Congresso Brasileiro de Pesquisa e Desenvolvimento em Petróleo e Gás*. [S.l.: s.n.], 2011.

22 SOARES, A. C. C. P. P. *Métodos Geofísicos em Obras Lineares*. Tese (Doutorado) — Dissertação (Mestrado em Engenharia Civil)–Pontifícia Universidade Católica . . . , 2009.

23 MORAES, P. C. X. d. et al. Determinação de propriedades petrofísicas e geológicas utilizando uma técnica de análise digital de rochas. [sn], 2018.

24 JUDD, W. S. et al. *Sistemática Vegetal-: Um Enfoque Filogenético*. [S.l.]: Artmed Editora, 2009.

25 SUGUIO, K. *Geologia Sedimentar*. [S.l.]: Blucher, 2003. 400 p.

26 PRESS, F.; MENEGAT, R. *Para entender a Terra*. [S.l.]: Bookman Porto Alegre, 2006. v. 656.

27 TEIXEIRA, W.; TOLEDO, M. C. M. de; FAIRCHILD, T. R. *Decifrando a terra*. [S.l.]: Oficina de textos, 2003.

28 ROTH, E. S. Temperature and water content as factors in desert weathering. *The Journal of Geology*, The University of Chicago Press, v. 73, n. 3, p. 454–468, 1965. ISSN 00221376, 15375269.

29 TOLEDO, M. C. M. e. a. Intemperismo e formação do solo. in: Decifrando a terra. *Decifrando a Terra*, p. 140–166, 2000.

30 BERNER, R. A.; HOLDREN, G. R. Mechanism of feldspar weathering: Some observational evidence. *Geology*, Geological Society of America, v. 5, n. 6, p. 369–372, 1977.

31 IMESON, A. C. Addressing soil erosion in europe: proceedings of the scape workshop in alicante, spain, june 2003. *Land Degradation & Development*, John Wiley & Sons, Ltd., v. 16, n. 6, p. 505–508, 2005. ISSN 1099-145X.

32 CROZIER, M. J. Landslides: Causes, consequences and environment. *Géographie physique et Quaternaire*, v. 41, n. 3, p. 409–410, 1987.

33 KASTRO, A. *A geologia, os geólogos e seus métodos*. [S.l.]: Oficina de textos, 2010.

34 WALTHER, J. *Einleitung in die Geologie als historische Wissenschaft*. [S.l.]: Gustav Fischer, 1984. 536 p.

35 GALLOWAY, W. E. Hydrogeologic regimes of sandstone diagenesis: Part 1. concepts and principles. AAPG Special Volumes, 1984.

36 CHOQUETTE, P. W.; PRAY, L. C. Geologic nomenclature and classification of porosity in sedimentary carbonates. *AAPG bulletin*, American Association of Petroleum Geologists (AAPG), v. 54, n. 2, p. 207–250, 1970.

37 MILANI, E. J. et al. Petróleo na margem continental brasileira: geologia, exploração, resultados e perspectivas. *Revista Brasileira de Geofísica*, SciELO Brasil, v. 18, p. 352 – 396, 00 2000. ISSN 0102-261X.

38 OLIVEIRA, V. A. A. de. *Caracterização de Reservatórios Não Convencionais/ Tight Gas*. [S.l.]: Projeto de conclusão do curso de graduação em Geofísica - Universidade Federal Fluminense, 2014.

39 TUCKER, M. E.; WRIGHT, V. P. Carbonate sediments and limestones: Constituents. In: _____. *Carbonate Sedimentology*. [S.l.]: Blackwell Publishing Ltd., 2009. p. 1–27. ISBN 9781444314175.

40 TECHNOLOGY, P. G. *Geologia do Petróleo*. [S.l.: s.n.], 2011.

41 BRAZIL, F. A. F. *Estratigrafia de Sequências e Processo Diagenético: Exemplo dos Arenitos Marinho-Rasos da Formação Ponta Grossa, Noroeste da Bacia do Paraná*. Dissertação (Mestrado) — Programa de pós-graduação em Análise de Bacias e Faixas Móveis. Universidade do Estado do Rio de Janeiro, 2004.

42 HYNE, N. *Dictionary of petroleum exploration, drilling & production*. [S.l.]: PennWell Corporation, 2014.

43 GRESSLY, A. Observations géologiques sur le jura soleurois. *Neue Denkschr*, v. 2, p. 1–112, 1938.

44 FáVERA, J. C. D. *Fundamentos de estratigrafia moderna*. [S.l.]: EdUERJ, 2001.

45 ROS, L. F. d.; GOLDBERG, K. Reservoir petrofacies: a tool for quality characterization and prediction. In: *AAPG, Annual Convention and Exhibition, Long Beach, Abstracts Volume*. [S.l.: s.n.], 2007. p. 1.

46 THOMAS, J. E. *Fundamentos de engenharia de petróleo*. [S.l.]: Interciência, 2001.

47 NIKRAVESH, M.; AMINZADEH, F. Soft computing for reservoir characterization. In: _____. *Fuzzy Partial Differential Equations and Relational Equations: Reservoir Characterization and Modeling*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. p. 1–79. ISBN 978-3-540-39675-8.

48 WUERGES, A. F. E.; BORBA, J. A. Redes neurais, lógica nebulosa e algoritmos genéticos: aplicações e possibilidades em finanças e contabilidade. *JISTEM - Journal of Information Systems and Technology Management*, scielo, v. 7, p. 163 – 182, 00 2010. ISSN 1807-1775.

49 AMINZADEH, F. Applications of ai and soft computing for challenging problems in the oil industry. *Journal of Petroleum Science and Engineering*, v. 47, n. 1, p. 5 – 14, 2005. ISSN 0920-4105. Intelligent Computing in Petroleum Engineering.

50 NIKRAVESH, M.; ADAMS, R. D.; LEVEY, R. A. Soft computing: tools for intelligent reservoir characterization (iresc) and optimum well placement (owp). *Journal of Petroleum Science and Engineering*, v. 29, n. 3, p. 239 – 262, 2001. ISSN 0920-4105. Soft Computing and Earth Sciences.

51 NIKRAVESH, M. Soft computing-based computational intelligent for reservoir characterization. *Expert Systems with Applications*, v. 26, n. 1, p. 19 – 38, 2004. ISSN 0957-4174. Intelligent Computing in the Petroleum Industry, ICPI-02.

52 ABEL, M. et al. PetroGrapher: Managing petrographic data and knowledge using an intelligent database application. *Expert Systems with Applications*, v. 26, n. 1 SPEC.ISS., p. 9–18, jan. 2004. ISSN 09574174.

53 ROS, L. F. D.; GOLDBERG, K. Reservoir petrofacies: a tool for quality characterization and prediction. In: *AAPG, Annual Convention and Exhibition, Long Beach, Abstracts Volume*. [S.l.: s.n.], 2007. p. 1.

54 GHALLAB, S. A. et al. A fuzzy expert system for petroleum prediction. *WSEAS, Croatia*, v. 2, p. 77–82, 2013.

55 CRACKNELL, M. J.; READING, A. M. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, Elsevier, v. 63, p. 22–33, 2014.

56 JÚNIOR, J. M. d. O. *Classificação de litofácies através da análise automática de perfis elétricos de poços de petróleo da Amazônia*. Dissertação (Mestrado) — Universidade Federal do Amazonas, 2014.

57 SEBTOSHEIKH, M. A.; SALEHI, A. Lithology prediction by support vector classifiers using inverted seismic attributes data and petrophysical logs as a new approach and investigation of training data set size effect on its performance in a heterogeneous carbonate reservoir. *Journal of Petroleum Science and Engineering*, Elsevier, v. 134, p. 143–149, 2015.

58 HORROCKS, T.; HOLDEN, E.-J.; WEDGE, D. Evaluation of automated lithology classification architectures using highly-sampled wireline logs for coal exploration. *Computers & Geosciences*, Elsevier, v. 83, p. 209–218, 2015.

59 DONG, S.; WANG, Z.; ZENG, L. Lithology identification using kernel fisher discriminant analysis with well logs. *Journal of Petroleum Science and Engineering*, Elsevier, v. 143, p. 95–102, 2016.

60 MAHMOODI, O.; SMITH, R. S.; TINKHAM, D. K. Supervised classification of down-hole physical properties measurements using neural network to predict the lithology. *Journal of Applied Geophysics*, Elsevier, v. 124, p. 17–26, 2016.

61 SAPORETTI, C. M.; FONSECA, L. G. da; PEREIRA, E. A lithology identification approach based on machine learning with evolutionary parameter tuning. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 16, n. 12, p. 1819–1823, 2019.

62 MIN, X.; PENGBO, Q.; FENGWEI, Z. Research and application of logging lithology identification for igneous reservoirs based on deep learning. *Journal of Applied Geophysics*, v. 173, p. 103929, 2020. ISSN 0926-9851.

63 HE, M.; GU, H.; WAN, H. Log interpretation for lithology and fluid identification using deep neural network combined with mahakil in a tight sandstone reservoir. *Journal of Petroleum Science and Engineering*, v. 194, p. 107498, 2020. ISSN 0920-4105.

64 LIU, H. et al. Well logging based lithology identification model establishment under data drift: A transfer learning method. *Sensors*, v. 13, p. 3643, 2020.

65 RUZGAR, E.; ERCIYES, K. Clustering based distributed phylogenetic tree construction. *Expert Systems with Applications*, Elsevier, v. 39, n. 1, p. 89–98, 2012.

66 OUAFI, A. Z. et al. Use of cluster analysis method in log's data processing: prediction and rebuilding of lithologic facies. *Recent Advances in Enviromental Science and Geoscience*, p. 98, 2014.

67 MAHMOODI, O.; SMITH, R. Clustering of downhole physical property measurements at the victoria property, sudbury for the purpose of extracting lithological information. *Journal of Applied Geophysics*, Elsevier, v. 118, p. 145–154, 2015.

68 SILVA, R. R. *Caracterização Petrofísica de Reservatório Carbonático Albiano a partir de Perfis de Poços e Medições de Laboratório*. Dissertação (Mestrado) — Universidade Estadual do Norte Fluminense, 2016.

69 METHE, P.; GOEPEL, A.; KUKOWSKI, N. Testing the results of estimating lithological stratigraphy through cluster analysis on geophysical borehole logging data through multi sensor core logging data. In: *EGU General Assembly Conference Abstracts*. [S.l.: s.n.], 2017. v. 19, p. 9642.

70 OLOSO, M. A. et al. Hybrid functional networks for oil reservoir pvt characterisation. *Expert Systems with Applications*, Elsevier, v. 87, p. 363–369, 2017.

71 WANG, X. et al. Lithology identification using an optimized knn clustering method based on entropy-weighed cosine distance in mesozoic strata of gaoqing field, jiyang depression. *Journal of Petroleum Science and Engineering*, v. 166, p. 157 – 174, 2018. ISSN 0920-4105.

72 SAPORETTI, C. M. et al. Computational intelligence techniques and phylogenetic trees for identification of sedimentary petrofacies. In: SBC. *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2018. p. 437–448.

73 ABDIDEH, M.; AMERI, A. Cluster analysis of petrophysical and geological parameters for separating the electrofacies of a gas carbonate reservoir sequence. *Natural Resources Research*, Springer, p. 1–14, 2019.

74 MOHAGHEGH, S.; RICHARDSON, M.; AMERI, S. Use of intelligent systems in reservoir characterization via synthetic magnetic resonance logs. *Journal of Petroleum Science and Engineering*, Elsevier, v. 29, n. 3-4, p. 189–204, 2001.

75 RODOLFO, S. B. et al. How to improve reservoir characterization models using intelligent systems. In: _____. *Soft Computing for Reservoir Characterization and Modeling*. Heidelberg: Physica-Verlag HD, 2002. p. 387–417. ISBN 978-3-7908-1807-9.

76 MOHAGHEGH, S. D. et al. *An Intelligent Systems Approach to Reservoir Characterization*. [S.l.], 2005.

77 ANIFOWOSE, F.; ABDULRAHEEM, A. Fuzzy logic-driven and svm-driven hybrid computational intelligence models applied to oil and gas reservoir characterization. *Journal of Natural Gas Science and Engineering*, Elsevier, v. 3, n. 3, p. 505–517, 2011.

78 ARAÚJO, E. H. S. d. *Sistema inteligente para estimar a porosidade em sedimentos a partir da análise de sinais GPR*. Tese (Doutorado) — Universidade Federal do Rio Grande do Norte, 2013.

79 PANJA, P. et al. Application of artificial intelligence to forecast hydrocarbon production from shales. *Petroleum*, Elsevier, 2017.

80 ZHANG, Y. et al. Improvement of petrophysical workflow for shear wave velocity prediction based on machine learning methods for complex carbonate reservoirs. *Journal of Petroleum Science and Engineering*, Elsevier, p. 107234, 2020.

81 HESLOP, K. et al. Log ascii standard (las) version 3.0. *The Log Analyst*, Society of Petrophysicists and Well-Log Analysts, v. 40, n. 06, 1999.

82 VICTORINE, J. R. Las version 2.0: A digital standard for logs. Canadian Well Logging Society, 2017.

83 CRAIN, E. *Crain's petrophysical handbook*. [S.l.]: Spectrum 2000 Mindware Limited, 2002.

84 HEARST, J. R.; NELSON, P. H. Well logging for physical properties. 1985.

85 GIBSON, C. R. et al. *Basic well log analysis for geologists*. [S.l.]: American Association of Petroleum Geologists, 1982.

86 HIGASHI, N. Natural gas in china. *Energy markets and security*, 2009.

87 ZHU, X. Sedimentary petrology. *Petroleum Industry Press, Beijing*, p. 130–256, 2008.

88 ABEL, M. *Estudo da perícia em petrografia sedimentar e sua importância para a engenharia de conhecimento. 2001. 239 f.* Tese (Doutorado) — Instituto de Informática, UFRGS, Porto Alegre, 2001.

89 ROS, L. F. D. et al. Petrographer: Uma aplicação de banco de dados inteligente para a descrição petrográfica e interpretação petrogenética de rochas-reservatórios de petróleo. In: *2º Congresso Brasileiro de Pesquisa e Desenvolvimento em Petróleo e Gás*. [S.l.: s.n.], 2003.

90 OLIVEIRA, L. C.; PEREIRA, E. Aplicação de parâmetros diagenéticos para a caracterização do arcabouço estratigráfico do devoniano da bacia do paraná. In: *5º Congresso Brasileiro de Pesquisa e Desenvolvimento em Petróleo e Gás*. [S.l.: s.n.], 2009.

91 PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

92 EFRON, B. Censored data and the bootstrap. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 76, n. 374, p. 312–319, 1981.

93 EFRON, B.; TIBSHIRANI, R. et al. [bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy]: rejoinder. *Statistical science*, Institute of Mathematical Statistics, v. 1, n. 1, p. 77–77, 1986.

94 PEREIRA, G. A. de A. *Modelos de Memória Longa Para Geração de Cenários Hidrológicos Sintéticos*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio de Janeiro, 2011.

95 DAVIDSON, I.; SATYANARAYANA, A. Speeding up k-means clustering by bootstrap averaging. In: *IEEE data mining workshop on clustering large data sets*. [S.l.: s.n.], 2003.

96 HAN, J.; LUO, M. Bootstrapping k-means for big data analysis. In: IEEE. *2014 IEEE International Conference on Big Data (Big Data)*. [S.l.], 2014. p. 591–596.

97 HOFMANS, J. et al. On the added value of bootstrap analysis for k-means clustering. *Journal of classification*, Springer, v. 32, n. 2, p. 268–284, 2015.

98 GARETH, J. et al. *An introduction to statistical learning: with applications in R*. [S.l.]: Spinger, 2013.

99 NETO, J. M.; MOITA, G. C. Uma introdução à análise exploratória de dados multivariados. *Química Nova*, SciELO Brasil, v. 21, n. 4, p. 467–469, 1998.

100 XIONG, H.; WU, J.; CHEN, J. K-means clustering versus validation measures: A data distribution perspective. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, v. 39, 2009.

101 LINDEN, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, v. 4, p. 18–36, 2009.

102 VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, v. 3(4), p. 209–235, 2010.

103 ROUSSEEUW, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 1987.

104 KRUSKAL, W. H.; WALLIS, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, Taylor & Francis Group, v. 47, n. 260, p. 583–621, 1952.

105 MOTTA, C. G. L. *Sistema Inteligente para Avaliação de Riscos em Vias de Transporte Terrestre*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, 2004.

106 HASSOUN, M. H. et al. *Fundamentals of artificial neural networks*. [S.l.]: MIT press, 1995.

107 STORN, R.; PRICE, K. Differential evolution-a simple and efficient adaptive scheme for global optimization over continuous spaces. *Tech. Rep. TR-95-012*, 1995.

108 VAPNIK, V. N.; KOTZ, S. *Estimation of dependences based on empirical data*. [S.l.]: Springer-Verlag New York, 1982. v. 40.

109 HAYKIN, S. *Redes Neurais*. 2. ed. [S.l.]: Bookman, 2001.

110 ZIEN, A. et al. Engineering suppot vector machine kernels that recognize translation initiation sites in dna. *Bioinformatics*, n. 16, p. 906 – 914, 2000.

111 CHEN, J.; LI, Z.; BIAN, B. Application of data mining in multi-geological-factor analysis. In: CAI, Z. et al. (Ed.). *Advances in Computation and Intelligence*. [S.l.]: Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6382). p. 402–411. ISBN 978-3-642-16492-7.

112 CHAVES, A. d. C. F. Extração de regras fuzzy para máquinas de vetor de suporte (svm) para classificação em múltiplas classes. *Rio de Janeiro*, 2006.

113 FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics Springer, Berlin, 2001. v. 1.

114 QUINLAN, J. R. Induction of decision trees. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 1, n. 1, p. 81–106, mar 1986. ISSN 0885-6125.

115 SILVA, L. M. O. d. *Uma Aplicação de Árvores de Decisão, Redes Neurais e KNN para a Identificação de Modelos ARMA Não-Sazonais e Sazonais*. Tese (Doutorado) — Pós-graduação em Engenharia Elétrica, PUC-Rio, Rio de Janeiro, 2005.

116 SILVA, L. Uma aplicação de árvores de decisão, redes neurais e knn para a identificação de modelos arma não sazonais e sazonais. *Rio de Janeiro. 145p. Tese de Doutorado-Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro*, 2005.

117 FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001.

118 BÜHLMANN, P.; YU, B. Boosting with the $L_2$ loss: regression and classification. *Journal of the American Statistical Association*, Taylor & Francis, v. 98, n. 462, p. 324–339, 2003.

119 HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. Extreme learning machine: a new learning scheme of feedforward neural networks. In: IEEE. *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*. [S.l.], 2004. v. 2, p. 985–990.

120 GUO, P.; CHENG, W.; WANG, Y. Hybrid evolutionary algorithm with extreme machine learning fitness function evaluation for two-stage capacitated facility location problems. *Expert Systems with Applications*, v. 71, p. 57 – 68, 2017. ISSN 0957-4174.

121 HUANG, G.-B. What are extreme learning machines? filling the gap between frank rosenblatt's dream and john von neumann's puzzle. *Cognitive Computation*, Springer, v. 7, n. 3, p. 263–278, 2015.

122 HUANG, G. et al. Trends in extreme learning machines: A review. *Neural Networks*, v. 61, n. Supplement C, p. 32 – 48, 2015. ISSN 0893-6080.

123 KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*. [S.l.: s.n.], 1995. v. 14, n. 2, p. 1137–1145.

124 PLAGIANAKOS, V.; TASOULIS, D.; VRAHATIS, M. A review of major application areas of differential evolution. In: *Advances in differential evolution*. [S.l.]: Springer, 2008. p. 197–238.

125 CHENG, S.-L.; HWANG, C. Optimal approximation of linear systems by a differential evolution algorithm. *IEEE Transactions on Systems, man, and cybernetics-part a: systems and humans*, IEEE, v. 31, n. 6, p. 698–707, 2001.

126 POWERS, D. M. W. Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation. In: . [S.l.]: Technical Report SIE-07-001, 2007.

127 LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. International Biometric Society, v. 33, p. 159–174, 1977.

128 WILCOXON, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, v. 1, n. 6, p. 80–83, 12 1945. ISSN 00994987.

129 SIEGEL, S.; JR, N. J. C. *Estatística não-paramétrica para ciências do comportamento*. [S.l.]: Artmed Editora, 1975.

130 JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. [S.l.]: 6.ed.ED. Person, 2007.

131 JOLLIFFE, I. Principal component analysis. In: *International encyclopedia of statistical science*. [S.l.]: Springer, 2011. p. 1094–1096.

132 SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, v. 4, n. 4, p. 406–425, 1987.

133 HUERTA-CEPAS, J.; SERRA, F.; BORK, P. Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, Society for Molecular Biology and Evolution, v. 33, n. 6, p. 1635–1638, 2016.

134 SOUZA, P. H. R. de; ROCHA, M. B. Sistemática filogenética em revista de divulgação científica: análise da scientific american brasil. *Alexandria: Revista de Educação em Ciência e Tecnologia*, Universidade Federal de Santa Catarina (UFSC), v. 8, n. 1, p. 75–99, 2015.

135 BARBANÇON, F. et al. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica*, John Benjamins Publishing Company, v. 30, n. 2, p. 143–170, 2013.

136 GRAY, R. D.; ATKINSON, Q. D. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, Nature Publishing Group, v. 426, n. 6965, p. 435, 2003.

137 RZHETSKY, A.; NEI, M. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular biology and evolution*, v. 10, n. 5, p. 1073–1095, 1993.

138 SAPORETTI, C. M. et al. Comparação de métodos de agrupamento para classificação de dados petrográficos. *Anais do SIMMEC/EMMCOMP 2014*, 2014.

139 OLIVEIRA, L. *Estudo das relações entre o arcabouço estratigráfico e as alterações diagenéticas observadas na seção Devoniana da Bacia do Paraná*. Dissertação (Mestrado) — Faculdade de Geologia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2009.

140 PRICE, K.; STORN, R. M.; LAMPINEN, J. A. *Differential evolution a practical approach to global optimization*. [S.l.]: Springer, 2005. (Natural Computing Series). ISBN 9783540209508,3540209506.

141 SILVA, A. A. et al. Artificial neural networks to support petrographic classification of carbonate-siliciclastic rocks using well logs and textural information. *Journal of Applied Geophysics*, v. 117, p. 118 – 125, 2015. ISSN 0926-9851.

142 POLLOCK, D. W.; BARRON, O. V.; DONN, M. J. 3d exploratory analysis of descriptive lithology records using regular expressions. *Computers & Geosciences*, v. 39, p. 111 – 119, 2012. ISSN 0098-3004.

143 GRAY, J. M.; BISHOP, T. F.; WILFORD, J. R. Lithology and soil relationships for soil modelling and mapping. *CATENA*, v. 147, p. 429 – 440, 2016. ISSN 0341-8162.

144 SAPORETTI, C. M.; GOLIATT, L.; PEREIRA, E. Neural network boosted with differential evolution for lithology identification based on well logs information. *Earth Science Informatics*, Springer, p. 1–8, 2020.

# APÊNDICE A – Conjuntos de Dados

Os dados petrográficos estão disponíveis em <https://goo.gl/Vc5mms>. Os dados petrofísicos foram fornecidos pela ANP e não podem ser disponibilizados.