Camila Martins Saporetti

Comparação de técnicas de inteligência computacional para classificação de dados petrográficos

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre em Modelagem Computacional.

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

Coorientador: Prof. D.Sc. Egberto Pereira

Juiz de Fora

2016

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.
Comparação de técnicas de inteligência computacional para classificação de dados petrográficos / Camila Martins Saporetti. -- 2016.
119 f. : il.

Orientador: Leonardo Goliatt da Fonseca
Coorientador: Egberto Pereira
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2016.

1. Inteligência Computacional. 2. Diagênese. 3. Petrografia Sedimentar. I. Fonseca, Leonardo Goliatt da , orient. II. Pereira, Egberto, coorient. III. Título.

Camila Martins Saporetti

Comparação de técnicas de inteligência computacional para classificação de dados petrográficos

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre em Modelagem Computacional.

Aprovada em 22 de Fevereiro de 2016.

BANCA EXAMINADORA

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

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

Prof. D.Sc. Marcus Vinicius Berao Ade
Universidade Estadual do Rio de Janeiro

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

Dedico este trabalho à minha mãe. 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.

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.

Ao Leonardo Goliatt agradeço a orientação, a confiança, a dedicação e claro a paciência. 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 a conclusão desta dissertação e planejamento de trabalhos futuros.

Ao Leonardo Oliveira pela colaboração. As suas sugestões foram importantes para desfecho de artigos e apresentações. Sem dúvidas foi importante para a construção deste trabalho.

À Ariane e à Jacqueliny agradeço pela companhia, pelos conselhos. Vocês se tornaram minhas irmãzinhas.

Às minhas meninas, 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.

À Alessandra por ser essa pessoa com o coração imenso, obrigada pela amizade e por sempre estar disposta a me ajudar.

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

Aos professores e técnicos do PPGMC por tornarem possível este 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.

"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

Modelos preditivos de distribuição de heterogeneidades e qualidade em reservatórios de hidrocarbonetos são de fundamental importância para exploração e otimização da produção de campos de óleo e gás. As heterogeneidades são determinadas através das distintas petrofácies sedimentares, 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. Recentemente, técnicas oriundas da área de inteligência computacional têm sido usadas para auxiliar na tomada de decisões de especialistas em diversos problemas de Geociências. O objetivo desta dissertação é avaliar o desempenho de diferentes técnicas baseadas em inteligência computacional para prever a classificação de amostras petrográficas pertencentes a uma mesma bacia sedimentar e propor o uso delas nesse tipo de problema. Para isso, desenvolveu-se um *framework* computacional para classificar petrofacies de acordo com seus constituintes. Os dados analisados são provenientes de três fontes distintas. A primeira base de dados é formada por amostras da região de Tibagi (PR) e a segunda da região de Dom Aquino (MS). Tais amostras são referentes a uma unidade litoestratigráfica formalizada na Bacia do Paraná como Membro Tibagi. A terceira é a junção das duas bases anteriores. A quarta por amostras do membro Mucuri da Bacia Sedimentar do Espírito Santo. A metodologia proposta envolve o uso de métodos de classificação, técnicas de validação cruzada, redução de dimensionalidade, seleção de características e o emprego de assembleia de constituintes. Os parâmetros envolvidos no ajuste dos métodos foram determinados por um processo de busca exaustiva com validação cruzada, e métricas de classificação adequadas foram usadas para avaliar e comparar os resultados. A metodologia apresentada, além de avaliar o desempenho de diversas técnicas de inteligência computacional, surge como uma alternativa para auxiliar o geólogo/especialista na determinação e caracterização das petrofácies, contribuindo para a redução do esforço no processo manual de individualização.

Palavras-chave: Inteligência Computacional. Diagênese. Petrografia Sedimentar.

ABSTRACT

Predictive models of heterogeneities distribution and quality in hydrocarbon reservoirs are of fundamental importance for exploration and production optimization of oil and gas fields. The heterogeneities are determined by the different sedimentary petrofacies, a set of petrographic characteristics that specifies a group of rocks. The identification and classification of petrofacies is usually a time consuming procedure, and the use of computational methods can reduce the time and effort spent in the analysis. Recently, techniques derived from the computational intelligence research area have been used to assist in making decisions experts in several problems in Geosciences. The purpose of this dissertation is evaluating the performance of different techniques based on computational intelligence to predict the classification of petrographic samples belonging to the same sedimentary basin. A computational framework was developed to classify petrofacies according to their constituents. The data was collected from three different sources. The first database is formed by thin sections of Tibagi region (PR). The second by thin sections of Dom Aquino region (MS). Such thin sections are for a lithostratigraphic unit formalized in the Paraná Basin as Member Tibagi. The third by thin sections from the two previous databases. The fourth database is a set of thin sections from Mucuri member of the Espírito Santo sedimentary basin. The proposed method involves the use of classifiers, cross validation, dimensionality reduction, feature selection and the use of ensemble of constituents. The parameters involved in adjusting methods were determined by an exhaustive search procedure with cross-validation and classification metrics were used to evaluate and compare the results. The presented methodology evaluates the performance of several computational intelligence techniques, and arises as an alternative to assist the geologist in the determination and characterization of petrofacies, helping to reduce the effort in the process of individualization.

Keywords: Computational Intelligence. Diagenesis. Sedimentary Petrography.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

Lista de Abreviações

AAM	Análise de Agrupamentos baseado em Modelo
AC	Acurácia
ACP	Análise de Componentes Principais
D	Darcy
DT	Decision Tree
ECA	Expectativa Condicional Alternada
EM	Escalonamento Multidimensional
EUA	Estados Unidos da América
KF	K-Fold
KNN	K-Nearest Neighbors
LinSVM	Linear Support Vector Machines
MAG	Modelo Aditivo Generalizado
md	milidarcys
MLP0	Multi-Layer Perceptron
MSE	Mean Squared Error
NFS	No Feature Selection
PAM	Partitioning Around Medoids
P&D	Pesquisa e Desenvolvimento
RBFSVM	RBF Support Vector Machines
RF	Random Forest
RNA	Redes Neurais Artificiais
RVR	Relevance Vector Regression
SC	Silhouette Coefficient
SFpr	Select False Positive Rate
SKB	Select K-Best
SKF	Stratified K-Fold
SP	Select Percentile
SVM	Support Vector Machines
SVR	Support Vector Regression
Swe	Family Wise Error

1 INTRODUÇÃO

1.1 Contextualização

Nessa dissertação iremos focar nas rochas reservatórios de um sistema petrlífero. O petróleo que é uma substância oleosa e inflamável, utilizado na produção de energia elétrica, calorífica ou mecânica. O petróleo originou a partir de restos orgânicos de zooplâncton e fitoplâncton depositados no fundo de lagos e mares através de transformações termoquímicas ao longo de milhares de anos.

Os reservatórios estão divididos em convencionais e não convencionais (Jacomo, 2014). Os convencionais tratam-se daqueles em que os hidrocarbonetos podem ser extraídos por processos de recuperação mais simples e com viabilidade econômica. São constituídos de reservatórios porosos e permeáveis, de baixa viscosidade e de densidade baixa e média. Os não convencionais são aqueles que não apresentam características que garantam que o hidrocarboneto acumulado possa ser extraído por procedimentos simples de recuperação, exigindo recursos tecnológicos mais avançados.

A maior parte da exploração ocorre nos reservatórios convencionais, mas as pesquisas em reservatórios não-convencionais vem aumentando, uma vez que o petróleo é uma fonte não-renovável. Com intuito de resolver este problema, pesquisadores estão desenvolvendo técnicas para auxiliar na extração de petróleo de reservas que não são comuns, como as areias asfálticas (areias betuminosas) do Canadá ou exploração em grandes profundidades nos oceanos. No Brasil, tem-se concentrado nos depósitos pré-sal, situados no oceano há mais de cinco quilômetros de profundidade, abaixo de uma camada de sal com cerca de dois quilômetros. Os EUA tem avançado na exploração, no desenvolvimento e na produção de reservas de gás natural e gás de folhelho, tratando-se de recursos não convencionais, sendo o principal responsável pela oferta do gás naquele país nos últimos anos (Bartlett, 2000).

Os recentes desafios encontrados na indústria de óleo e gás resultaram em problemas com difíceis resoluções onde as abordagens geralmente usadas são ineficazes ou inviáveis. Alguns exemplos são a delimitação de fronteiras de um reservatório de gás natural (Egging *et al.*, 2010) e a análise do potencial de reservatórios de petróleo (Zhang, 1999). Com o

objetivo de contribuir para a solução de tais problemas, nas últimas décadas vem crescendo o número de trabalhos utilizando modelos matemáticos e técnicas computacionais.

No processo de exploração de petróleo é fundamental ter conhecimento sobre a qualidade e o potencial de um reservatório. Para tal análise, o estudo da diagênese das rochas é de vital importância. O termo diagênese pode ser descrito como um conjunto de processos químicos e físicos sofridos pelos sedimentos desde a sua deposição até a sua consolidação. As propriedades das rochas são investigadas através das litofácies quando trata-se da análise em campo, da litologia ou petrofácies, quando se trata de laboratório.

As litofácies sedimentares são um aglomerado de características físicas e orgânicas macroscópicas das rochas, já as petrofácies sedimentares são um conjunto de características petrográficas, como por exemplo, cor, granulometria, estrutura sedimentar, geometria deposicional, presença de fósseis, paleocorrente, entre outras que individualizam um grupo de rochas, e sua determinação permite a inferência da heterogeneidade do reservatório.

Um reservatório de petróleo, no enfoque econômico, é julgado de bom redimento, se dispõe, além de uma abundante quantidade de óleo, condições ótimas de recuperação dos fluidos, que estão intimamente ligadas à porosidade e à permeabilidade da formação (Azevedo, 2005). A permeabilidade é um fator geométrico que define características de transmissão de fluidos em um meio poroso, representando a área de fluxo efetiva na escala dos poros.

Com o objetivo de prever propriedades de reservatórios petrolíferos, vários trabalhos na literatura podem ser citados. Maraschin e Mizusaki (2008) desenvolveram um método alternativo para prever permeabilidade em dados de perfis de poços, que se baseia na integração entre teoria *wavelet* e Redes Neurais Artificiais. Silva *et al.* (2015) realizaram a classificação petrográfica de rochas carbonato-siliciclástica usando um algoritmo de retropropagação de rede neural suportado por informações mineralógica e textural a partir de um conjunto de dados coletados da Bacia Provence do Sul, no sudoeste da França. Xie (2008) apresentou uma comparação de três métodos de análise variada para prever a permeabilidade baseado em dados de perfis de poços de dois reservatórios carbonáticos. O primeiro é uma combinação do algoritmo de seleção *stepwise* com a técnica de Expectativa Condicional Alternada. O segundo é a aplicação de árvore de regressão e validação cruzada. O terceiro emprega *splines* de regressão adaptativa multivariada. Wang e Carr

(2012) utilizaram um modelo baseado em Redes Neurais Artificiais para prever litofácies a partir de dados de perfis de poços da bacia sedimentar Appalachian.

1.2 Justificativa

A Bacia do Paraná é uma bacia intracratônica com uma área total de aproximadamente 1400000 km^2 que ocupa partes do Brasil, Argentina, Paraguai e do norte do Uruguai, como pode ser visto na Figura 1.1. Possui reservas significativas de carvão mineral, sendo objeto de estudo em muitos trabalhos.

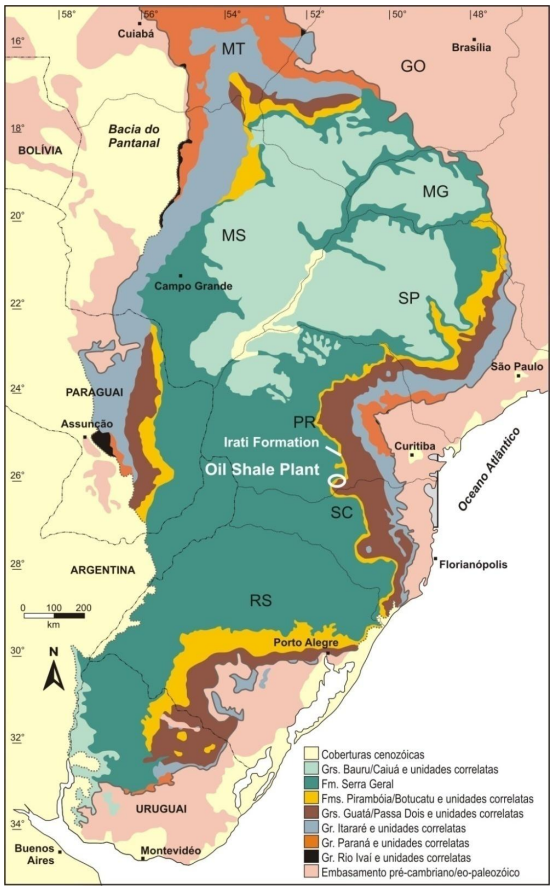


Figura 1.1: Localização da Bacia do Paraná (extraído de (França e Oliveira, 2010)).

A Formação Ponta Grossa, sistema petrolífero mais importante da Bacia do Paraná (Milani e Zalan, 1999). Esta formação é subdividida em três membros: Membro Jaguariaíva (inferior), Membro Tibagi (médio) e Membro São Domingos (superior). Segundo Grahn (1999), as idades para os três membros são, respectivamente praguiana-emsiana (+/-410.8 ma a +/-393.3 ma), emsiana (+/-407.3 ma a +/-393.3 ma) e eifeliana-eofrasniana (+/-393.3 ma a +/-382.7 ma). Essa consiste de folhelhos argilosos, micáceos,

finamente laminados, cinzentos, e folhelhos sílticos a arenosos, com siltitos e arenitos muito finos subordinados (Petri e Fúlvaro, 1983). Os arenitos finos a muito finos, depositados no Membro Tibagi no Eifeliano (+/-390 ma), compreendem possíveis reservatórios para um hipotético sistema petrolífero Ponta Grossa-Ponta Grossa (Oliveira, 2009). A superposição de variados e intensos processos diagenéticos sobre os sedimentos depositados ocasionou em uma alta heterogeneidade nesses reservatórios. Isso dificulta a realização de procedimentos de produção e faz com que a recuperação de hidrocarbonetos seja limitada.

As bacias do Espírito Santo e Mucuri ocupam o litoral norte do Espírito Santo e o sul da Bahia. Esta bacia possui campos petrolíferos de grande importância, com reservas significativas de gás natural e óleo leve. Em 2008, a Petrobras anunciou a descoberta de óleo de boa qualidade no campo de Golfinho, na Bacia do Espírito Santo e a partir dessa data a área passou a despertar interesse de pesquisadores, sendo mais estudada.

A definição da distribuição de heterogeneidades de reservatórios de hidrocarbonetos são primordiais para exploração e otimização da produção de campos de petróleo. Para uma rocha ser julgada como um bom reservatório, ela deve possuir as seguintes características: uma extensão considerável, boa porosidade (de 15% a 20%), uma apreciável permeabilidade e eficiência de recuperação de hidrocarbonetos. Essas características são ditas petrofísicas, e indicam o resultado de toda a história geológica dos sedimentos depositados e em particular das condições de sedimentação e dos fenômenos de diagênese, sendo de fundamental importância para definição da qualidade do reservatório.

O estudo da diagênese das rochas vem sendo incentivado pelas empresas petrolíferas, com intuito de entender a distribuição da porosidade em arenitos. O interesse vem do fato que estes arenitos podem originar em 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 como reservatório (Maraschin e Mizusaki, 2008).

A partir da análise petrográfica é possível identificar, por meio de um microscópio de luz polarizada, os constituintes de uma rocha. Dessa forma, pode-se realizar uma avaliação das implicações futuras de suas propriedades sobre o comportamento dos produtos gerados, como o petróleo (de Menezes, 1999). 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 a fim de avaliar as heterogeneidades do reservatório. Este processo usualmente é muito longo, pois envolve o processo de amostragem, geração dos dados e posterior interpretação destes (Cevolani *et al.*, 2011). Além disso, devido à grande quantidade de dados, nem toda informação obtida pode ser adequadamente aproveitada no procedimento manual.

A permeabilidade geralmente é determinada em laboratório com permeabilímetros pelos métodos de nível constante e de nível variável. Em ensaios de campo a permeabilidade é identificada pelos métodos do bombeamento em dois poços, pelo bombeamento em um poço (pontual), método de recuperação, de infiltração ou de rebaixamento e por perfis geofísicos complexos. Desse modo, a permeabilidade é determinada no laboratório pela medida da pressão relacionada ao fluxo do fluido viscoso que permeia uma amostra de rocha. Os testes em laboratório, todavia, não são empregados para os materiais de testemunhos mal recuperados ou para amostras de calha (Jones e Owens, 1980).

Tendo em vista as exposições acima, nota-se a necessidade de automatizar os procedimentos de caracterização do reservatório com o objetivo de melhorar a produtividade ou a interpretabilidade dos dados. Nesse contexto, técnicas de Inteligência Computacional aparecem como um mecanismo útil para auxiliar na classificação de petrofácies e da permeabilidade.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo principal desta dissertação é comparar técnicas de inteligência computacional para a classificação de dados petrográficos e propor o uso delas nesse tipo de problema através do desenvolvimento de um método computacional. Dessa maneira pretende-se avaliar algumas técnicas com intuito de auxiliar o geólogo/petrólogo na análise de dados petrográficos.

1.3.2 Objetivos Específicos

Os seguintes objetivos específicos são enumerados:

1. Estudar e implementar classificadores, seletores de características e métodos de validação cruzada;
2. Implementar uma busca exaustiva a fim de encontrar os parâmetros dos classificadores que retorna um melhor desempenho;
3. Estudar e implementar métricas para avaliar as técnicas de inteligência computacional;
4. Aplicar a abordagem em diferentes bases de dados para avaliar o seu comportamento em diferentes aplicações importantes para determinar o potencial de um reservatório de petróleo.

1.4 Organização da Dissertação

Esse trabalho subdivide-se em seis capítulos. No Capítulo 1 encontra-se uma apresentação dessa dissertação que é composta pela contextualização, justificativa, hipótese da dissertação 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.

Os conceitos de Inteligência Computacional são apresentados no Capítulo 3, onde encontra-se revisão bibliográfica, no qual constam trabalhos que utilizaram técnicas de mineração de dados com o objetivo de descobrir relações úteis nos dados. Nesse capítulo estão contidos conceitos relacionados a classificação, validação cruzada, seleção de características e métricas para a seleção de modelos.

No Capítulo 4 estão descritas as bases de dados, procedimentos realizados para o desenvolvimento do projeto em questão, os resultados e discussões. O Capítulo 5 apresenta a conclusão com base nos resultados obtidos e as indicações para trabalhos futuros e no Apêndice A encontram-se as bases de dados utilizadas para testar o desempenho das técnicas de Inteligência Computacional.

2 CONTEXTO GEOLÓGICO

2.1 Sedimentologia

A sedimentologia é o estudo dos depósitos sedimentares e suas origens. Pode ser aplicada em diversos tipos de depósitos: antigos ou modernos, marinhos ou continentais, minerais, texturas e estruturas, diagênese e evolução temporal e espacial (Suguio, 2003).

Com base em observação e descrição das feições em sedimentos moles e duros 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.

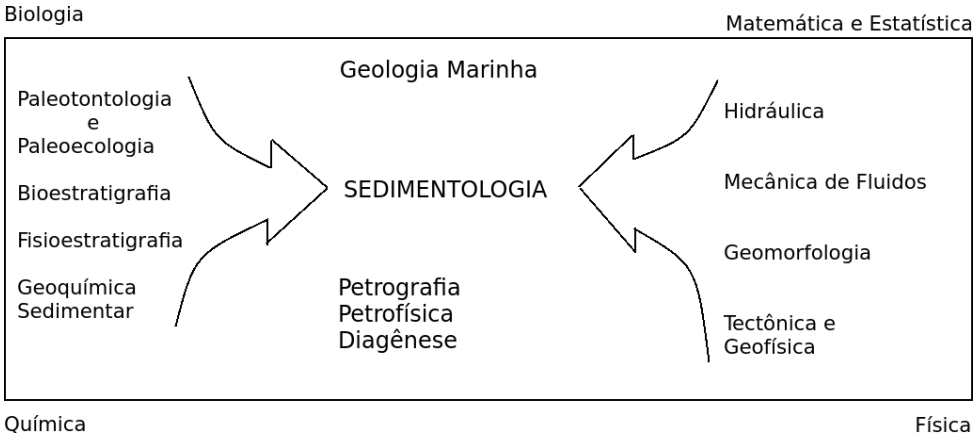


Figura 2.1: Sedimentologia e disciplinas relacionadas (modificado de (Suguio, 2003))

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 da porosidade e da permeabilidade de um sedimento, influenciando na capacidade de armazenamento e no fluxo de fluidos, como água, petróleo e gás (Suguio, 2003).

Independente dos tipos de rochas que formam os detritos, elas passam pelo intemperismo que realiza a desintegração e/ou decomposição, sucedida da erosão. Os detritos gerados são transportados e posteriormente sedimentados e litificados. A seguir são descritas algumas dessas ações.

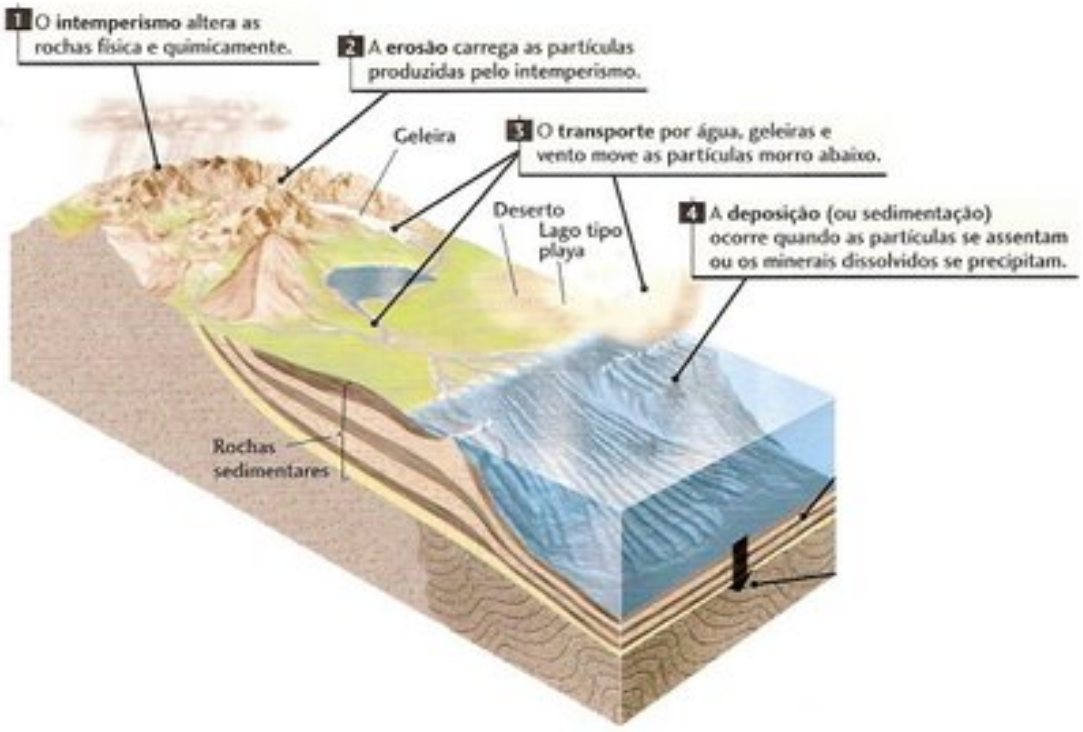


Figura 2.2: Intemperismo, Erosão, Transporte e Deposição (retirado de (Press e Menegat, 2006))

2.1.1 Intemperismo

O intemperismo é o processo no qual as rochas são destruídas na superfície terrestre. O intemperismo gera as argilas, os solos e as substâncias dissolvidas e levadas 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 (Teixeira *et al.*, 2003).

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 (Roth, 1965).

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 (Teixeira *et al.*, 2003).

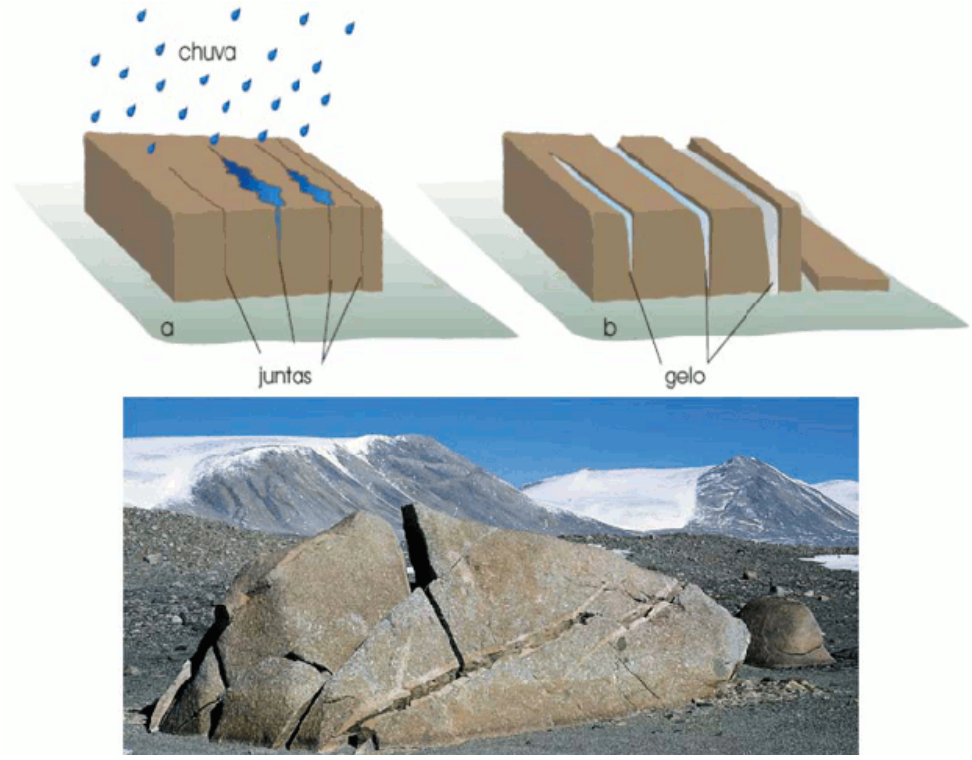


Figura 2.3: Fragmentação pela ação do gelo (retirado de (Teixeira *et al.*, 2003))

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 (Toledo, 2000).

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 (Suguio, 2003).

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 o 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.

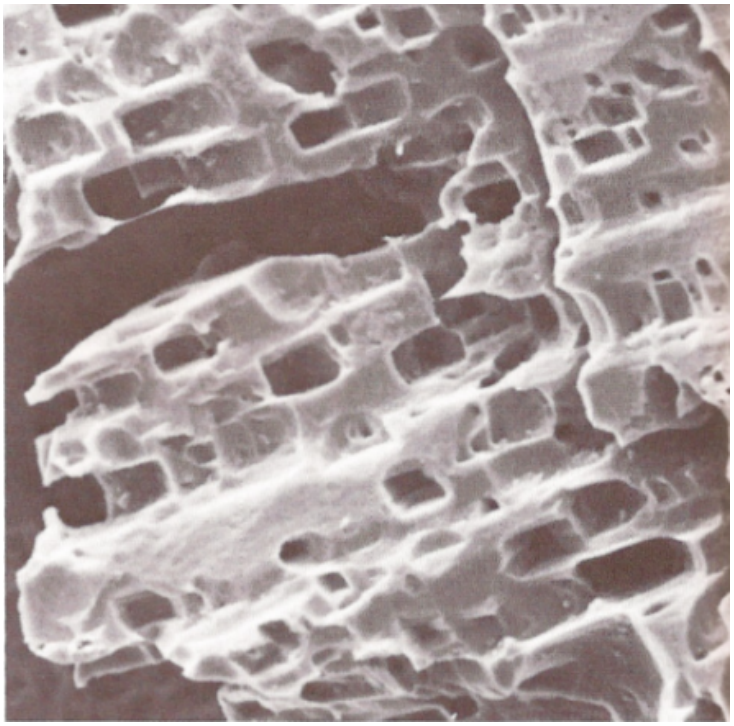


Figura 2.4: Fotomicrografia, obtida por microscópio eletrônico de varredura, de um feldspato marcado e corroído pelo intemperismo químico no solo.(retirado de (Berner e Holdren, 1977))

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.

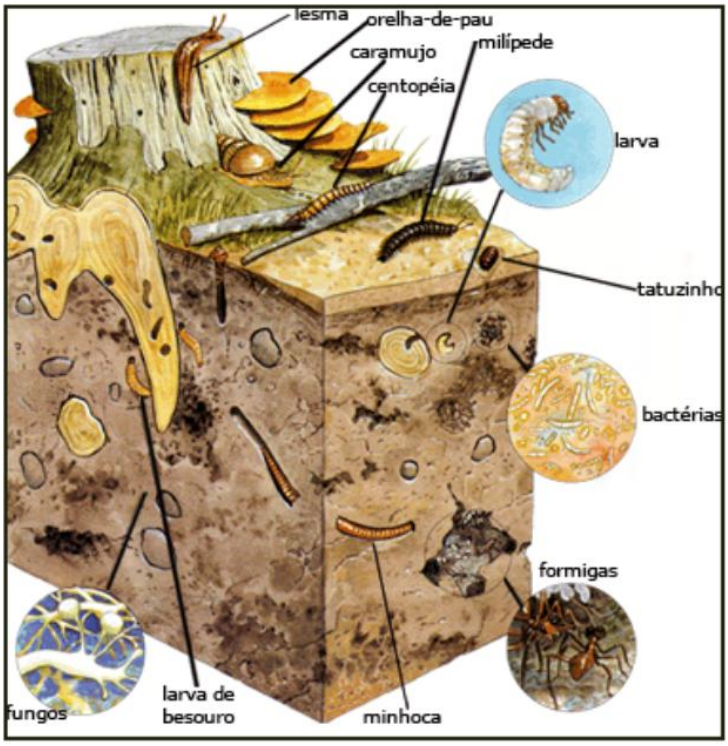


Figura 2.5: A ação dos organismos vivos no solo que geram o intemperismo biológico (retirado de (Teixeira *et al.*, 2003))

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 (Imeson, 2005).

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 (Suguio, 2003). 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 (Suguio, 2003).

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 (Suguio, 2003).

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.

Movimentos de massa também denominados fluxos gravitacionais referem-se aos mecanismos de transporte de sedimentos paralelamente ao substrato, com maior ou menor atuação da gravidade. (Crozier, 1987). 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.

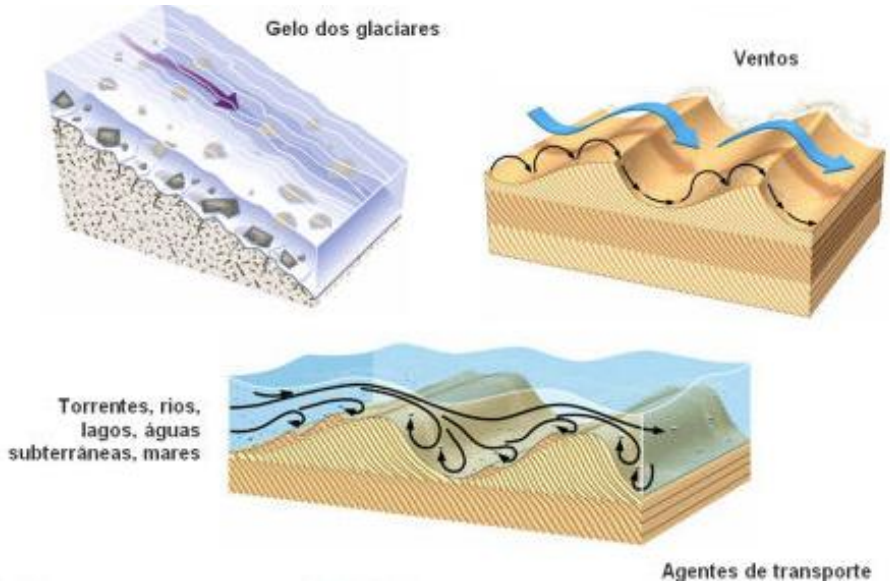


Figura 2.6: Agentes de Transporte (retirado de (Kastro, 2010))

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 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 (Press e Menegat, 2006).

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 2.7 exibe esse processo.

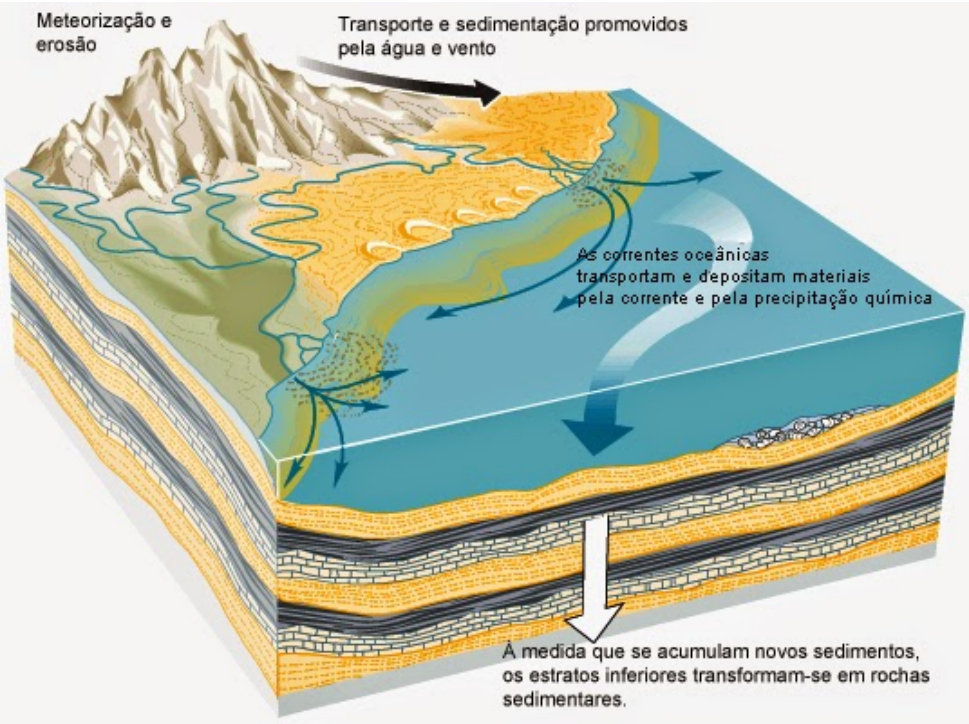


Figura 2.7: Formação de Rochas Sedimentares (retirado de (Press e Menegat, 2006))

As bacias sedimentares são áreas de extensão considerável, onde sofreu subsidência continuada e sedimentação, gerando uma espessa acumulação de sedimentos e rochas sedimentares.

2.3 Diagênese e Litificação

Segundo Walther (1984) a diagênese abrange todos os processos físicos e químicos que atuam sobre os sedimentos após a sua deposição, com exceção da pressão litostática devida superposição dos sedimentos, além do calor magmático. 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

interticiais 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.

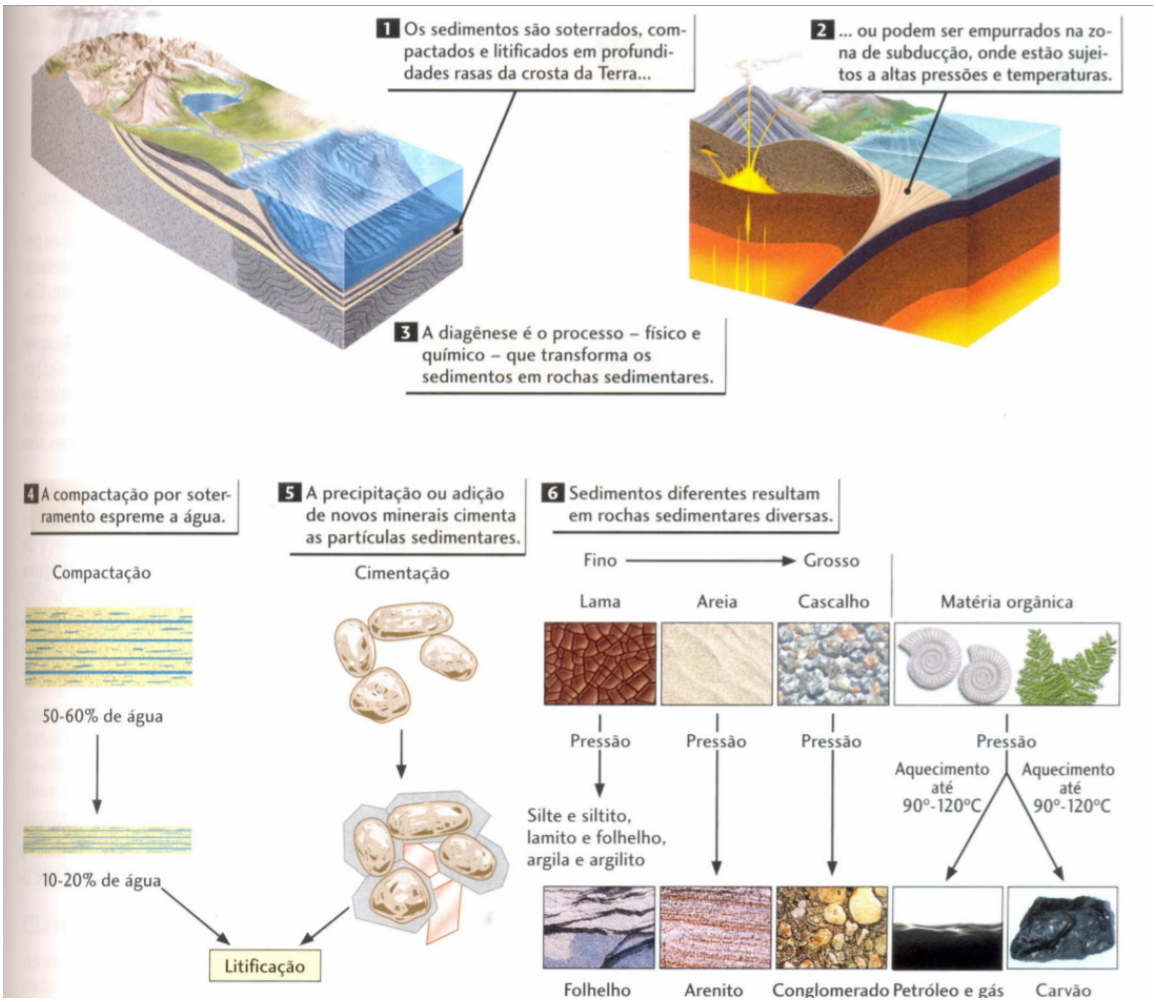


Figura 2.8: Processos Diagenéticos (retirado de (Press e Menegat, 2006))

2.4 Tipos de Rochas Sedimentares

As rochas sedimentares são constituídas, principalmente, por três componentes: terrígenos, aloquímicos e ortoquímicos (Folk, 1957). 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, oólicos, 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 (Folk, 1957):

- **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. O estudo dessas características juntamente com a simulação introdutória de condições ótimas foi denominada sistema petrolífero (Milani *et al.*, 2000).

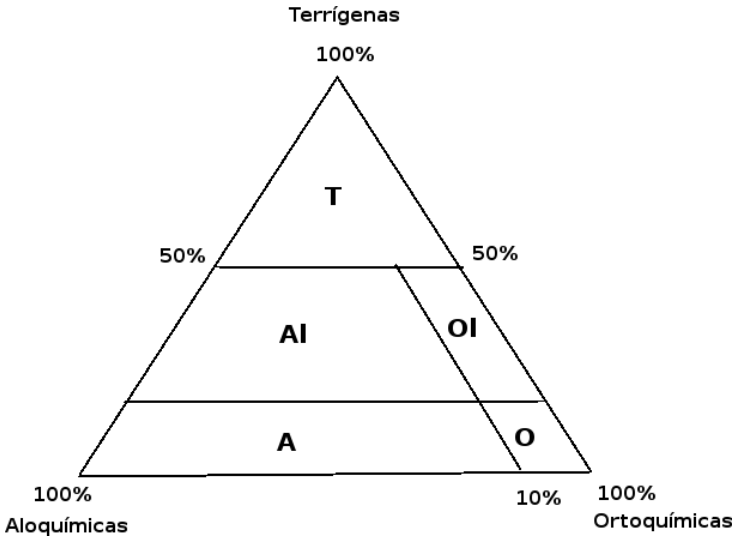


Figura 2.9: Diagrama triangular de classificação geral das rochas sedimentares segundo (Folk, 1957)

Um sistema petrolífero 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) (Milani *et al.*, 2000).

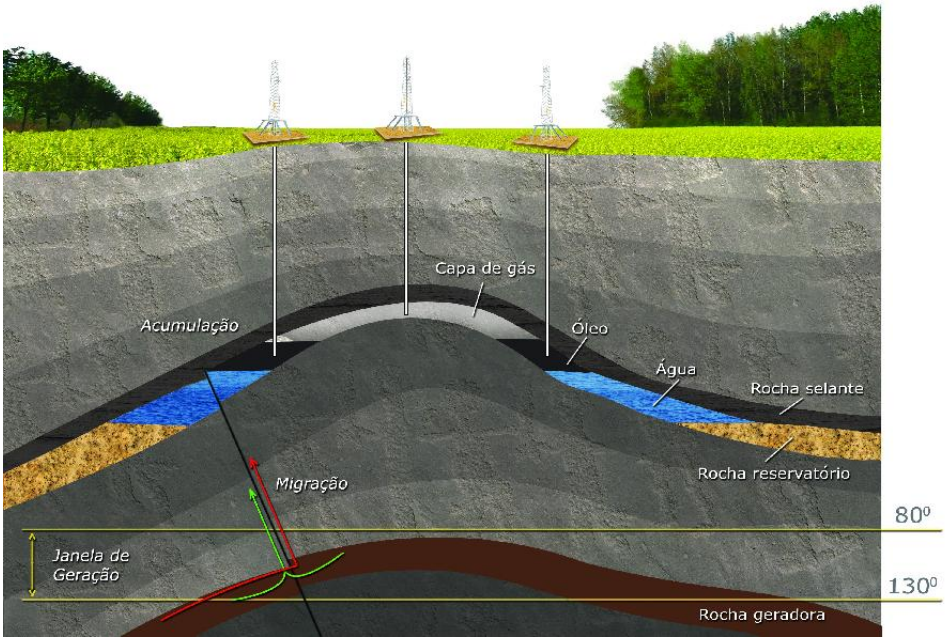


Figura 2.10: Sistema Petrolífero (retirado de (de Oliveira, 2014))

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.

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 (Milani *et al.*, 2000).

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 (Milani *et al.*, 2000). 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 (Tucker e Wright, 2009).

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 (Tucker e Wright, 2009).

2.5.4.2 Carbonáticas

As rochas reservatórios carbonáticas são formados principalmente por carbonatos provindos de processos biológicos e bioquímicos, ou seja, de origem 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.

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 (Tucker e Wright, 2009).

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% (Technology, 2011). 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 $(>20\%)$ (Technology, 2011).

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 (Technology, 2011).

2.7 Ambientes de Sedimentação, Litologia, Fácies e Petrofácies Sedimentares

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 (Suguio, 2003). 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 (Brazil, 2004) e (Suguio, 2003).

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 (Hyne, 2014). 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 de 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 (1938) 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 (Fávera, 2001).

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 Ros e Goldberg (2007), 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 (Ros e Goldberg, 2007), 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 (Ros e Goldberg, 2007). 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 (Suguio, 2003).

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 (Suguio, 2003).

2.8.1 Petróleo e Gás Natural

O petróleo originou-se através 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. Além de gerar a gasolina, muitos produtos são derivados do petróleo como a parafina, produtos asfálticos, querosene, óleo diesel e combustível de aviação.

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.

2.8.1.1 Exploração de Petróleo no Brasil

No início do século XIX a exploração de petróleo começou a se tornar atrativa devido ao interesse econômico. Nesse período o petróleo era utilizado como fonte de energia para a iluminação pública. Mas seu uso para tal finalidade durou até meados da década de 1870, quando se deu o início do uso de energia elétrica. Como consequência, a busca pelo fóssil diminuiu rapidamente, retornando somente no final do mesmo século, particularmente no século seguinte devido a criação dos motores a gasolina e a diesel. A partir de então, a matéria-prima passou a ter fundamentos comerciais para ser explorado (Ortiz Neto e Costa, 2007).

Esta recente utilidade do petróleo fez com que o emprego da ciência nas práticas

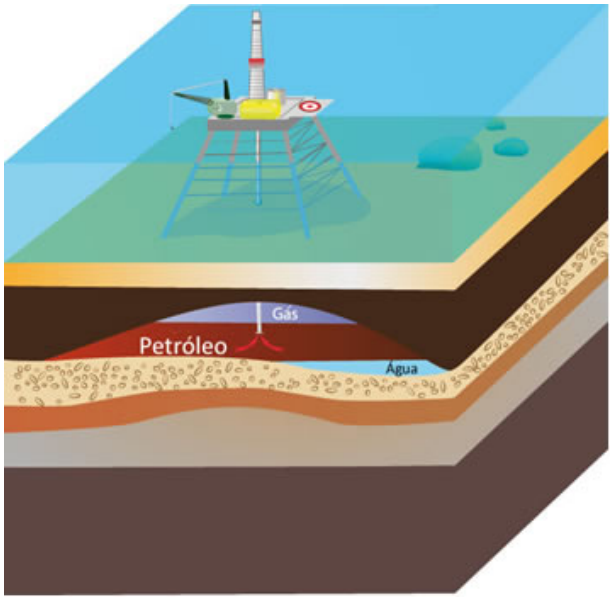


Figura 2.11: Posicionamento Gás e Petróleo

industriais aumentasse. A atividade petrolífera em conjunto com a atividade química, tornaram-se as primeiras a aplicar a ciência a partir de programas de Pesquisa e Desenvolvimento (P&D) para o desenvolvimento econômico. Desde então, o uso e P&D nas várias indústrias tem sido expressivo, devido ao fato de ser fundamental para a produção de novos itens e procedimentos profissionais das entidades (Ortiz Neto e Costa, 2007).

No Brasil, a primeira sondagem ocorreu em São Paulo, em meados de 1892, por Eugênio Ferreira de Camargo, quando ele fez a primeira perfuração na profundidade de 488 metros. No entanto, o poço só esguichou água sulfurosa. No ano de 1939 que foi descoberto o óleo de Lobato na Bahia.

A Petrobras foi criada em 1954 com o intuito de centralizar a exploração do petróleo no Brasil. Muitos poços passaram a ser perfurados. Nos dias de hoje, a Petrobras é uma das maiores empresas petrolíferas do mundo.

3 INTELIGÊNCIA COMPUTACIONAL

Inteligência Computacional é o estudo e projeto de agentes inteligentes (Poole *et al.*, 1997). Um agente é um ser que atua em algum ambiente. Agentes incluem animais, humanos, meios de transporte, organizações. Um agente inteligente é um sistema que age de maneira inteligente, isso quer dizer que o que ele faz é adequado para as ocasiões e objetivos, é variável de acordo com os ambientes e objetivos no decorrer do tempo, aprende através da experiência vivenciada, e toma as decisões apropriadas de acordo com restrição sensorial e a capacidade computacional. O principal objetivo científico da Inteligência Computacional é compreender os princípios que levam um ser comportar-se de forma inteligente, em sistemas naturais e artificiais. A hipótese fundamental parte de que o raciocínio corresponde à computação (Poole *et al.*, 1997).

A partir do conceito descrito anteriormente entende-se que os procedimentos que são realizados com base no domínio, a experiência de uma pessoa pode ser realizada por meio de técnicas de Inteligência Computacional. Essa necessidade aparece à medida que alguns fatores podem prejudicar o procedimento manual, como tempo, custo e introdução de erros. Nos últimos anos vários trabalhos têm usado técnicas baseadas em Inteligência Computacional para assistir na solução de problemas nas áreas de ciência e engenharia, inclusive na caracterização de reservatórios. A Seção 3.1 apresenta alguns destes trabalhos.

3.1 Revisão Bibliográfica

Lee e Datta-Gupta (1999) propuseram uma abordagem para prever a permeabilidade utilizando regressão em conjunto com análise estatística multivariada. Primeiro foi realizada a classificação dos dados de perfis de poços em eletrofácies, usando análise discriminante. Em seguida foi aplicada técnica de regressão para prever a permeabilidade utilizando perfis de poços dentro de cada eletrofácies. Foram examinadas três abordagens não-paramétricas Expectativa Condicional Alternada (ECA), o Modelo Aditivo Generalizado (MAG) e Redes Neurais Artificiais. Um exame das taxas de erro

associadas a análise discriminante para poços uncored indica que a classificação dos dados com base na caracterização de eletrofácies é mais robusto em comparação com outras abordagens. No caso das previsões de permeabilidade, o modelo ECA parece ser a melhor entre as três abordagens não-paramétricas.

Um modelo de previsão de porosidade baseado em Regressão de Vetores Suporte (Support Vector Regression, SVR) para um reservatório heterogêneo de arenitos foi desenvolvido no trabalho de Al-Anazi e Gates (2010). O método SVR tem sido comparado a perceptron multicamadas, a Redes Neurais com Regressão Generalizada e Redes Neurais com Função de Base Radial. Os resultados revelaram que o método SVR exibe precisão superior e maior robustez com relação a esses métodos de redes neurais, especialmente no que diz respeito à precisão ao generalizar a dados de porosidade inéditas. Ponte (2010) utilizou Redes Neurais Artificiais para identificar litofácies, estimando colunas estratigráficas a partir de análises combinadas de perfis geofísicos e testemunhos de poços. As Redes Neurais foram utilizadas em trechos não testemunhados para constituir várias medições geoelétricas e fornecer como resposta qual a litofácies que melhor caracteriza os parâmetros geoelétricos medidos.

Cevolani *et al.* (2011) desenvolveram um procedimento para identificação e classificação de petrofácies sedimentares. O processo era dividido em duas partes: 1) elaboração de um procedimento computacional para determinação do número de petrofácies e posterior agrupamento das amostras nas petrofácies determinadas, 2) visualização dos agrupamentos encontrados em duas dimensões. Para a primeira parte foi aplicado o algoritmo de agrupamento de partição em torno de medoides (PAM), a fim de identificar as petrofácies. Para a segunda parte foi utilizado a técnica de Análise de Componentes Principais para reduzir a dimensionalidade dos dados para duas dimensões. Para visualização dos dados foi empregado o algoritmo de Escalonamento Multidimensional (EM). O código computacional facilitou a visualização dos agrupamentos das petrofácies dessas rochas e permitiu que se obtenham de maneira rápida informações sobre a similaridade entre as amostras em um mesmo poço.

Olatunji (2011) usou Fuzzy do tipo 2 como uma nova abordagem para prever a permeabilidade de perfis de poços para a caracterização de um reservatório. Os resultados empíricos de simulação mostraram que Fuzzy do tipo 2 supera outras abordagens em geral e em particular na área de estabilidade e capacidade de lidar com dados em situações de

incerteza, que são características comuns de dados perfis de poços.

Raeesi *et al.* (2012) utilizou Redes Neurais Artificiais (RNAs) para interpretar dados sísmicos 3D com intuito de posteriormente identificar litofácies e determinar suas mudanças para exploração de reservatórios de hidrocarbonetos. As RNAs oferecem um manuseio superior quando tem-se não-linearidade nos dados, como é o caso dos dados sísmicos, o que proporciona uma melhor análise em relação a outras técnicas de análise matemática. Foi aplicada uma análise multi-característica com base em RNAs em dados de perfis de poços para determinar a alteração de litofácies e a heterogeneidade em um dos campos de petróleo estruturais estratigráficos no Golfo Pérsico.

Sharma *et al.* (2012) propuseram uma abordagem para melhorar a estimativa da permeabilidade de um reservatório de hidrocarbonetos através da caracterização de eletrofácies. A abordagem foi dividida em duas partes. A primeira é classificar os dados de perfis de poços em tipos de eletrofácies. Esta classificação é realizada com base nas medidas das características que individualizam um tipo. Foi aplicado uma combinação dos métodos Análise de Componentes Principais (ACP), Análise de Agrupamentos baseado em Modelo (AAM) e Análise Discriminante, este último usado para identificar e classificar os tipos de eletrofácies. Na segunda parte é utilizado a técnica de regressão não paramétrica, Expectativa Condicional Alternada (ECA), para prever a permeabilidade usando perfis de poços dentro de cada eletrofácies encontrada.

Um método para estimar a permeabilidade de perfis de poços foi proposto por Lacentre (2013). O método faz uso de modelos matemáticos, que se mostraram eficazes em campos da ciência e da engenharia, como Análise de Componentes Principais, Redes Neurais e Análise de Agrupamentos.

Gholami *et al.* (2014) empregaram Regressão de Vetores Relevantes (Relevance Vector Regression, RVR) e Regressão de Vetores Suporte (Support Vector Regression, SVR) combinados com Algoritmos Genéticos na previsão da permeabilidade em dados de perfis retirados de três poços localizados em um reservatório carbonático na parte sul do Irã. Foi realizada a comparação dos resultados do RVR com a do SVR, foi constatado que o RVR obteve maior acurácia, na previsão de permeabilidade. No entanto, SVR ainda pode ser considerado como uma segunda opção para a previsão de propriedades petrofísicas devido à sua eficiência confiável.

Iturraran-Viveros e Parra (2014) aplicaram modelos de Redes Neurais Artificiais

(RNAs)para prever a permeabilidade e a porosidade em dados de perfis de poços de um aquífero no sudeste da Flórida e a atenuação intrínseca de um reservatório de óleo de areia xistosa no nordeste do Texas. As estimativas de 1/Q dá um vislumbre da capacidade das RNAs de generalizar e mostrou-se ser aplicável com sucesso em RNAS para estimar 1/Q para os poços próximos, no mesmo nordeste campo de petróleo no Texas.

O trabalho de da Silva *et al.* (2015) compara o desempenho de diversas técnicas de Inteligência Computacional para a tarefa de classificação da permeabilidade em um conjunto de dados petrofísicos com 78 amostras de rochas de seis reservatórios carbonáticos diferentes. Os melhores comportamentos foram gerados pelos algoritmos Floresta Aleatória e SMO combinados com técnicas de discretização e seleção de atributos melhorando o desempenho Precisão em mais de 24% em relação ao modelo Kenyon e mais de 28% em relação ao modelo de Timur-Coates.

Asfahani *et al.* (2015) propôs o uso de lógica Fuzzy para interpretar a combinação de dados de perfis, nuclear e elétrico, de poços que inclui raio gama natural, densidade e porosidade neutrão, enquanto os perfis elétricos de poços incluem profundidade de invest rasa e profunda. O objetivo desse trabalho é descrever, caracterizar e estabelecer a litologia de áreas com grandes extensões basálticas no sul da Síria. A lógica fuzzy é aplicado com sucesso nos dados de perfis de Kodana e, portanto, pode ser utilizado como uma ferramenta poderosa para interpretar enorme de dados de perfis com maior número de variáveis necessárias para estimativas litológicas.

O uso de técnicas de Inteligência Computacional para resolver problemas envolve algumas etapas, como: Pré-Processamento, aplicação da técnica supervisionada ou não (de acordo com objetivo, se o problema é classificar ou agrupar), Validação Cruzada e Grid-Search são opções se o problema for de Classificação ou outro método afim de encontrar os parâmetros ótimos para a metodologia e Métrica para a seleção de modelos. Para visualizar os agrupamentos e a distribuição das amostras utilizou-se Análise de Componentes Principais. Estas etapas serão descritas nas próximas seções.

3.2 Pré-Processamento

A etapa de pré-processamento consiste na aplicação de técnicas para captação, organização, tratamento e a preparação dos dados. É um passo que tem importante

relevância no processo de classificação. Essa etapa engloba desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de Inteligência Computacional. Nas bases de dados utilizadas realizou-se a normalização, empregou-se técnicas de seleção de características e efetuou-se o balanceamento dos dados. Estes procedimentos serão descritos a seguir.

3.2.1 Normalização

As características foram transformadas por escalonamento de cada característica para um determinado intervalo. Estima as escalas e traduz cada característica individualmente de modo que ela esteja na gama dada no conjunto, ou seja, entre zero e um.

A Tabela 3.1 apresenta as variáveis utilizadas no procedimento, para maior esclarecimento.

Tabela 3.1: Variáveis Normalização

Variáveis	Descrição
X	base de dados utilizada (valores)
X.min	o valor mínimo das características
X.max	o valor máximo das características
min	0
max	1

A transformação é dada por:

$$X_{std} = (X - X.min)/(X.max - X.min) \tag{3.1}$$

$$X_{escala} = X_{std} * (max - min) + min$$

Como max foi definido como 1 e min como 0, a segunda equação se torna dispensável.

3.2.2 Seleção de Características

Grande parte das aplicações existentes de classificação, previsão e reconhecimento de padrões possuem bases de dados com muitas características. Em geral, quanto maior a quantidade de variáveis envolvidas melhor pode se tornar a representação de um

problema (Contreras, 2002). Porém, em várias situações, muitas destas características são insignificantes ou há uma forte correlação entre elas. Nesse contexto, o uso de técnicas seleção de características entra como uma alternativa para reduzir o número de variáveis e detectar as mais importantes.

A seleção de características é um procedimento em que um espaço de dados é transformado em um espaço de características com menor dimensão, ou seja, a base de dados sofre uma redução de dimensionalidade. Neste novo espaço deve estar contida as características que armazenam informações mais relevantes para os dados. As técnicas de seleção de características tratam da separação, dentre todas as características da base de dados, daquelas mais importantes de acordo com os dados. Nesta dissertação empregou-se seletores que utilizam seleção de características univariada. Este procedimento e os seletores aplicados são brevemente descritos a seguir.

3.2.2.1 Seleção de Características Univariada

A Seleção de Características Univariada seleciona as melhores características baseando-se em testes de estatística univariada. A aplicação de seleção de características pode ser vista como um passo de preprocessamento do classificador. Ela examina cada característica individualmente para determinar a força da relação da característica com a variável de saída. Estes métodos são simples de executar e compreender e são, em geral, particularmente bons para obter uma melhor compreensão dos dados. Existem distintas opções para a seleção univariada.

3.2.2.2 Seletores de Características

Fez-se uso de quatro métodos de seleção de características univariada: Select K-Best, Select Percentile, Select False Positive Rate e Select Family Wise Error, disponíveis no pacote *Scikit-Learn* em Python (Pedregosa *et al.*, 2011). O Select K-Best seleciona as k características com maiores pontuações (F-valor), segundo o método ANOVA. No Select Percentile as características são escolhidas 10% das características com maiores pontuações (F-valor ANOVA). O Select False Positive Rate escolhe através dos p-valores abaixo de α (α geralmente 0.05) baseado no teste FPR. O Teste FPR (False Positive Rate) avalia a

quantidade total de falsas detecções segundo a expectativa.

$$E(V/m_0)$$

onde E é a expectativa, V é o número de falsos positivos (erros do tipo I) e m_0 é o número de hipóteses nulas verdadeiras. A técnica Select Family Wise Error seleciona os p-valores correspondentes a taxa de Family Wise Error (FWE). Family Wise Error é probabilidade de fazer uma ou mais falsas descobertas ou erros do tipo I, entre todas as hipóteses ao realizar testes de hipóteses múltiplas.

$$FWE = Pr(V \geq 1)$$

onde V é o número de erros do tipo I.

3.2.3 Balanceamento dos Dados

Em problemas reais de classificação é comum se deparar com dados desbalanceados, ou seja, dados que possuem classes com um número de amostras muito superior, ditas majoritárias, as demais classes, conhecidas como minoritárias. O desbalanceamento de classe pode prejudicar o desempenho dos classificadores. Em decorrência, os classificadores podem apresentar uma tendência em responder bem as classes majoritárias em relação as minoritárias. Em vários casos, o importante é possuir um bom desempenho para as classes minoritárias; neste caso pode-se citar a detecção de fraudes em cartões de créditos.

Existem muitos tipos de técnicas para tratar o desbalanceamento dos dados. As mais utilizadas são as denominadas *oversampling* e *undersampling*. O *oversampling* refere-se ao ato de replicar as amostras das classes minoritárias até que todas as classes possuam o mesmo/aproximadamente o mesmo número de amostras. O *undersampling* é a redução do número de amostras da(s) classe(s) majoritárias até que o número de amostras das classes seja aproximadamente o mesmo. Este procedimento pode acarretar em perda de informação, um vez que pode haver a exclusão de muitas amostras. Nas bases de dados utilizadas empregou-se *oversampling* pois se trata de uma base de dados com muitas classes e a exclusão de muitas amostras poderia acarretar em perda de informações essenciais.

3.3 Classificação

A tarefa de classificação consiste na descoberta de regras de previsão para auxílio 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 (Motta, 2004).

3.3.1 Classificadores

3.3.1.1 Máquinas de Vetor Suporte

Máquinas de Vetor Suporte (Support Vector Machine - SVM) (Vapnik e Kotz, 1982) é uma técnica de aprendizado estatístico que vem ganhando destaque nos últimos anos. Os resultados o uso desta técnica são comparáveis aos gerados por outros métodos de aprendizado, como as Redes Neurais Artificiais (RNAs) (Haykin, 2001), e em algumas aplicações têm se mostrado superiores, tal como em Bioinformática (Zien *et al.*, 2000) e em Geologia (Chen *et al.*, 2010).

Supondo um problema com duas classes, o SVM tem o intuito de separar as amostras das duas classes a partir de uma função que será obtida através das amostras conhecidas na

fase de treinamento. O intuito é gerar um classificador que funcione de maneira adequada com amostras desconhecidas, ganhando capacidade de predizer as classes de futuras novas amostras.

SVMs realizam a combinação linear de atributos por meio de funções, denominadas funções kernel, para atribuir uma classe a uma determinada amostra. Pode-se utilizar diferentes tipos de funções kernel e diferentes parâmetros a variar de acordo com o kernel selecionado. O SVM é comumente formulado como um problema de otimização da seguinte forma,

$$Maximizar \sum_{i=1}^{n} \alpha_i \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(z_i).\phi(z_j) \rangle \tag{3.2}$$

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

onde y_i são as classes da amostra de treinamento z_i, ϕ é usado para transformar os dados em um espaço de alta dimensão, h é o número de amostras, w representa os coeficientes da função de decisão, a constante $C > 0$ é o erro de separação do hiperplano. O produto interno $\phi(z_i)'\phi(z_i)$ 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 3.2.

Tabela 3.2: 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	$\tanh(\beta_0 \langle z_i.z_j \rangle) + \beta_1$

Nesta dissertação foram utilizados o kernel linear (Polinomial com $p = 1$) e o kernel de base radial (RBF). O desempenho dos métodos acima depende da escolha adequadas dos parâmetro C para o kernel linear e γ e C para o kernel RBF.

3.3.1.2 K-Nearest Neighbors

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 3.1 tem-se um exemplo de classificação pelo KNN.

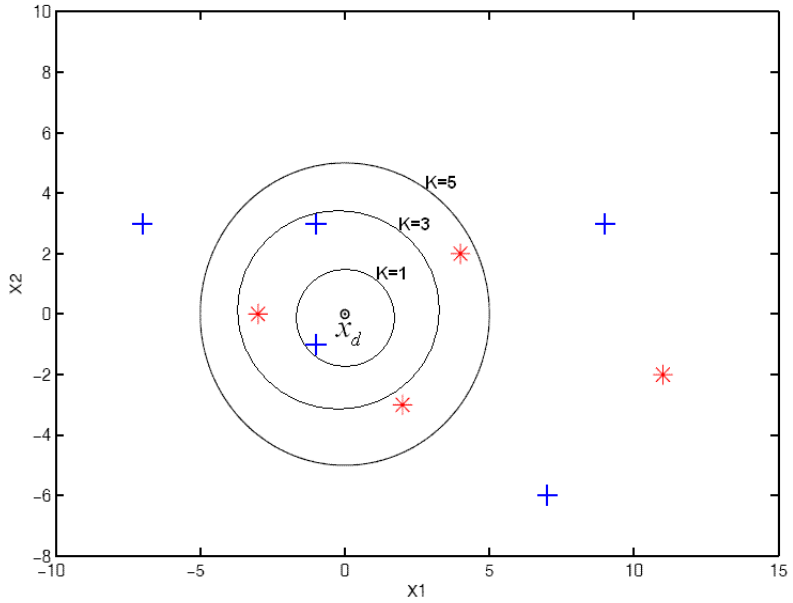


Figura 3.1: Para uma amostra desconhecida x_d entre amostras da classe 1 e 2. Dependendo do número de vizinhos mais próximos, x_d pode ser classificada como segue: se $K = 1$, x_d é classificado como "+", se $K = 3$, x_d é classificado como "+", se $K = 5$, x_d é classificado como "*".

3.3.1.3 Árvore de Decisão

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 complicado 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 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 (Friedman *et al.*, 2001).

A Figura 3.2 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 hiperretângulo no qual a interseção destes é vazia e a união é todo o espaço (Quinlan, 1986).

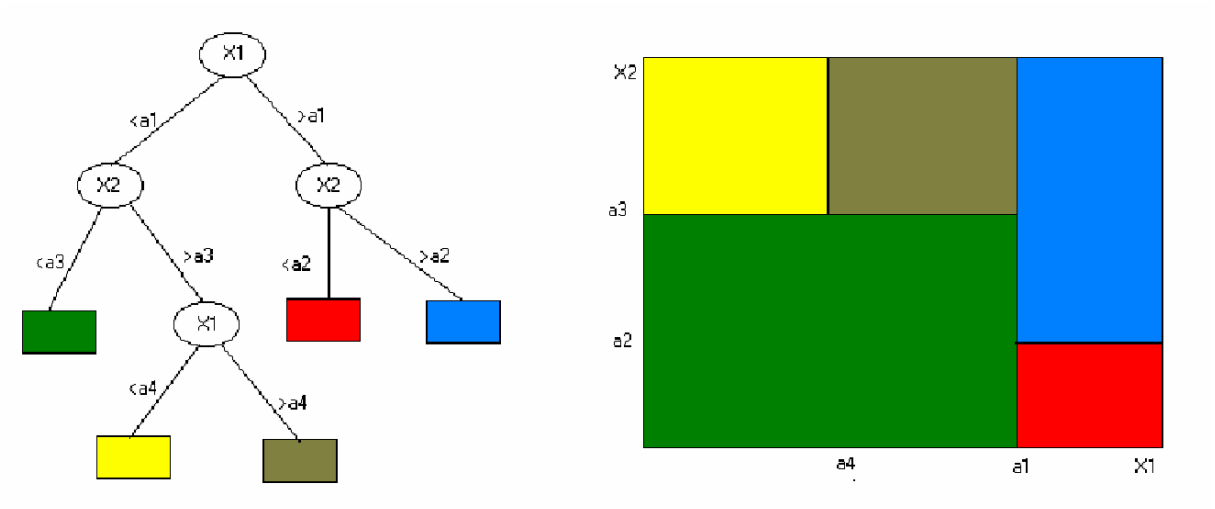


Figura 3.2: Árvore de decisão e sua respectiva exibição no espaço.

3.3.1.4 Random Forest

O algoritmo de Floresta Aleatórias (Random Forest, RF) foi introduzido por Breiman (2001). No algoritmo de florestas aleatórias, o conjunto de dados é dividido aleatoriamente em diversos subconjuntos de tamanho menor, diferentemente do método árvore de decisão, onde é feito uso de todo o conjunto de dados para formular a árvore. Cada um destes

conjuntos é construído por um tipo de amostragem denominada de bootstrap (Han, 2005), onde cada novo conjunto pode ter algumas amostras incluídas mais de uma vez e outras nenhuma vez. A amostragem bootstrap garante que 1/3 das amostras são utilizadas para testar as árvores depois de sua criação.

A partir de cada subconjunto reproduzido, uma árvore de decisão é construída. Em cada operação, primeiramente seleciona de forma randômica um conjunto de amostras do conjunto de treinamento. Para reproduzir uma árvore de decisão através desse subconjunto, o RF escolhe aleatoriamente um subconjunto de características como as características candidatas para cada nó. Dessa maneira, cada árvore de decisão é construída através do conjunto empregando subconjuntos aleatórios independentes de ambas as características e amostras. Uma floresta aleatória é uma coleção dessas árvores de decisão.

Depois que a floresta está gerada, há muitas árvores de decisão a serem testadas, sendo que todas contribuem para a classificação da amostra, através de um voto sobre qual classe a amostra deve pertencer. A classe mais votada é a atribuída a amostra.

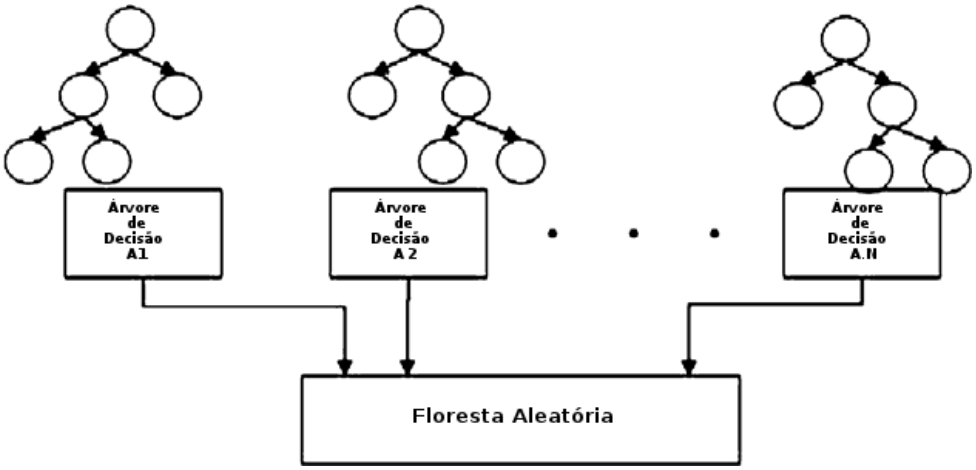


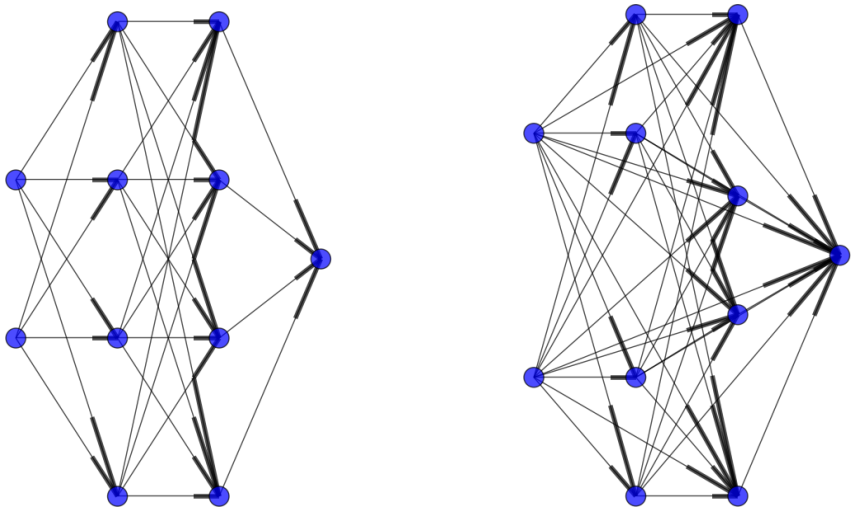
Figura 3.3: Floresta Aleatória

3.3.1.5 Multi-Layer Perceptron

As redes Perceptron de Múltiplas Camadas (Multi-Layer Perceptron - MLP0) (Haykin, 2001) são uma rede neural que têm sido aplicadas em variadas áreas, desempenhando funções de reconhecimento de padrões, controle e processamento de sinais, dentre outras. Uma rede do tipo MLP0 possui um conjunto de nós fontes que formam a camada de entrada, uma ou mais camadas ocultas e uma camada de saída. As redes de

múltiplas camadas diferenciam-se das redes de camada simples pelo número de camadas intermediárias, camadas entre a camada de entrada e a de saída. Essa arquitetura tem uma ou mais camadas ocultas, que são constituídas por neurônios computacionais, também conhecidos como neurônios ocultos. A atividade dos neurônios ocultos é intervir entre a camada de entrada externa e a saída da rede de forma útil.

Este algoritmo tem como parâmetro o número de camadas ocultas e neurônios, o algoritmo, a conectividade e renormalização. O número de camadas ocultas e neurônios é representado como [5,5], por exemplo, que indica 2 camadas ocultas com 5 neurônios cada. O algoritmo são os métodos usados para otimizar os pesos (Nocedal e Wright, 1999): l-bfgs é da família dos métodos Quase-Newton, sgd refere-se ao Método Descida Gradiente Estocástico, tnc trata a informação do gradiente no algoritmo de Newton truncado. A conectividade pode ser simplesmente conectada (mlgraph) ou totalmente conectada (tmlgraph). A Figura 3.4 exemplifica os tipos de conectividade.



(a) rede unidirecional simplesmente conectada

(b) rede unidirecional totalmente conectada

Figura 3.4: Com duas entradas, duas camadas escondidas (4 neurônios em cada uma) e uma saída. Em (a) o esquema de conexões simples com os neurônios são conectados somente aos neurônios da camada anterior, e em (b) o esquema totalmente conectado onde os neurônios estão conectados a todos os antecessores.

3.4 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 (Neto e Moita, 1998).

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 dentro e heterogeneidade entre os grupos (Neto e Moita, 1998).

O procedimento de agrupar pode ser dividido, de um modo geral, em duas fases: a primeira se refere à escolha de uma medida de dissimilaridade entre os objetos e a segunda, à aplicação de uma técnica de obtenção de grupos. 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. Neste trabalho foi realizado um teste inicial com a análise não-supervisionada com o intuito de introduzir alguns trabalhos futuros. O método utilizado foi o K-Means que será descrito a seguir.

3.4.1 K-Means

O K-Means (Xiong *et al.*, 2009), é 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 da função. A medida de dissimilaridade utilizada foi a distância Euclidiana. A Figura 3.5 ilustra a solução final de um agrupamento para dois grupos.

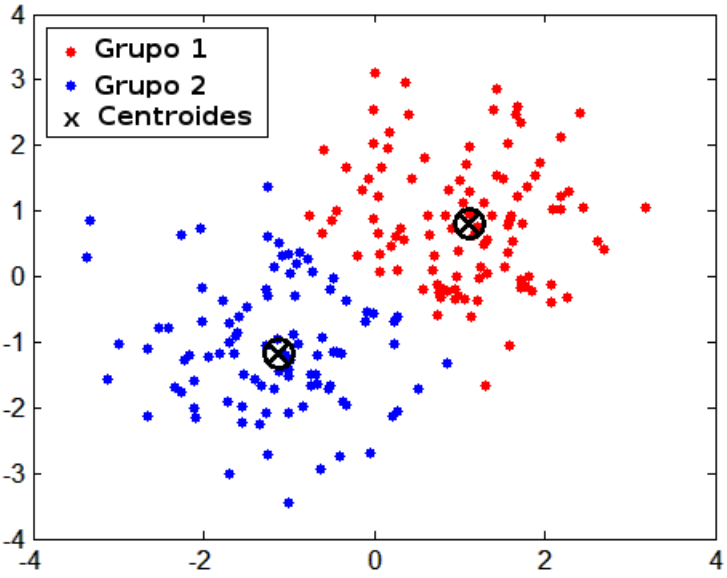


Figura 3.5: K-Means - $K = 2$.

3.5 Validação Cruzada e Grid Search

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. As técnicas de Validação Cruzada e Grid Search aparecem como alternativas bastante utilizadas no contexto citado anteriormente.

O grid search é a busca exaustiva pelos melhores parâmetros a partir da análise dos resultados obtidos com a execução do algoritmo de classificação para um intervalo de parâmetros e pode ser usado em conjunto com a validação cruzada. Neste caso, a base de dados é divida, de forma randômica, em dois conjuntos, treinamento e teste. A validação cruzada diminui o risco da base de dados ser dividida de forma não apropriada.

3.5.1 Técnicas de Validação Cruzada

Para avaliar os classificadores descritos anteriormente, foram empregadas as estratégias de validação cruzada K-Fold (KF), Stratified K-Fold (SKF) (Kohavi *et al.*, 1995).

3.5.1.1 K-Fold (KF)

No 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 ela possa generalizar sobre as futuras entradas. A Figura 3.6 mostra um esquema exemplificando o K-Fold.

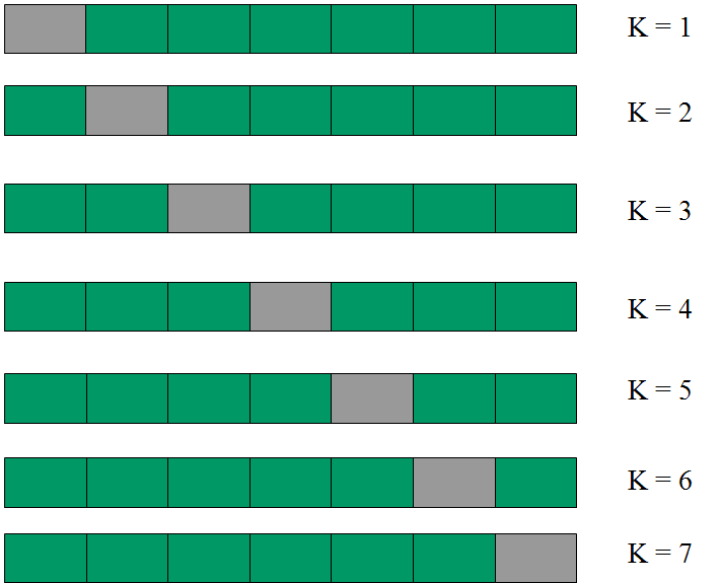


Figura 3.6: K-Fold - $K = 7$. Conjunto de treinamento, 6 amostras (quadros verdes) e conjunto de teste, 1 amostra (quadro cinza)

3.5.1.2 Stratified K-Fold (SKF)

Esta técnica é semelhante ao K-Fold, a diferença está na seleção dos conjuntos de treinamento e de teste. Os conjuntos são selecionados de modo que o subconjunto de treinamento contém aproximadamente as mesmas proporções dos tipos de classe.

3.6 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, Kappa e Erro Quadrático Médio (MSE) (Powers, 2007).

3.6.1 Acurácia

A Acurácia mede a proporção das amostras que foram classificadas corretamente, como pode ser visto na Eq. (3.3),

$$\mathrm{AC} = \frac{1}{N}\sum_{i=1}^{N} I(f(x_i) = y_i) \tag{3.3}$$

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 $I(falso) = 0$.

3.6.2 RECALL

O RECALL pode ser referido como a taxa de verdadeiros positivos ou sensibilidade, como definido na Eq. (3.4)

$$\mathrm{RECALL}(k) = \frac{TP}{TP + FN} \tag{3.4}$$

onde TP é taxa de verdadeiro positivo e FN é a taxa de falso negativo.

3.6.3 F1

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

$$\mathrm{F1} = \frac{2TP}{(2TP + FP + FN)} \tag{3.5}$$

onde TP é taxa de verdadeiro positivo, FN é a taxa de falso negativo e FP é a taxa de falsos positivos.

3.6.4 Kappa

O Teste de Kappa é uma medida de concordância interclassificador 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. (3.6):

$$Kappa = \frac{P_o - P_E}{1 - P_E} \tag{3.6}$$

onde

$$P_o = \frac{n^o \quad de \quad concordâncias}{n^o \quad de \quad concordâncias \quad + \quad n^o \quad de \quad discordâncias} \tag{3.7}$$

e

$$P_E = \sum_{i=1}^{N} (p_{i1} \times p_{i2}) \tag{3.8}$$

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 se o nível de concordância é razoável, Landis e Koch (1977) sugerem a interpretação mostrada na Tabela 3.3.

Tabela 3.3: 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

3.6.5 Erro Quadrático Médio

O Erro Quadrático Médio (Mean Squared Error - MSE), definido em na Eq.(3.9), é a média das diferenças entre o valor previsto e o valor real dos dados.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2 \tag{3.9}$$

onde $\hat{Y}$ é o vetor das classes verdadeiras e Y é o vetor das classes previstas.

3.6.6 Teste de Wilcoxon

O teste de Wilcoxon (Wilcoxon, 1945) é usado para comparar dois tratamentos quando os dados são obtidos através do esquema de pareamento. Os seguintes passos devem ser

seguidos na sua construção:

1. Calcular a diferença entre as observações para cada par.
2. Ignorar os sinais das diferenças e atribuir postos a elas.
3. Calcular a soma dos postos (S) de todas as diferenças negativas (ou positivas).

Para amostras pequenas (até 25 pares), o valor-p deve ser obtido através de uma tabela apropriada. "Valores críticos da Tabela - Prova de Wilcoxon" Para amostras grandes, a estatística do teste (S) tem aproximadamente distribuição gaussiana com média

$$\mu_s = \frac{n(n+1)}{4} \tag{3.10}$$

e desvio-padrão

$$\sigma_s = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{3.11}$$

Assim, o valor de

$$Z = \frac{S - \mu_s}{\sigma_s} \tag{3.12}$$

deve ser comparado ao valor do percentil da distribuição gaussiana.

3.6.7 Critérios de Validação - Agrupamento

Critérios de validação são medidas adotadas para avaliar a qualidade de um agrupamento. A escolha do critério depende do problema a ser tratado. Existem algumas publicações tratando desse assunto e comparando critérios, como em Vendramin *et al.* (2010). Nesta dissertação foi utilizado o coeficiente de silhueta (silhuette coefficient - SC) que será descrito a seguir.

3.6.7.1 Silhueta

A análise de silhueta (SC) é uma técnica proposta por Rousseeuw (1987); Vendramin *et al.* (2010). Trata-se de 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:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{3.13}$$

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

$$SC = \frac{\sum_{i=1}^{N} s_i}{N} \tag{3.14}$$

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 bem agrupados.

3.7 Análise de Componentes Principais

A Análise de Componentes Principais (ACP) é um método que tem como propósito, a análise dos dados usados objetivando sua redução, eliminação de sobreposições e a escolha dos modos mais representativos de dados através de combinações lineares das variáveis originais. Esta técnica é uma forma de identificar relação entre atributos retirados de dados. É muito útil quando os vetores de atributos possuem muitas dimensões, uma vez que é impossível uma representação gráfica.

3.7.1 Processo para uma Análise de Componentes Principais

O procedimento descrito a seguir foi baseado em Johnson e Wichern (2007) e em de Oliveira Araujo (2009). 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 n 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

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

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, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, portanto λ_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 λ_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

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

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 λ_i sua variância.

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

4 EXPERIMENTOS COMPUTACIONAIS

4.1 Bases de Dados

4.1.1 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 4.1 apresenta a localização dos furos de sondagem.

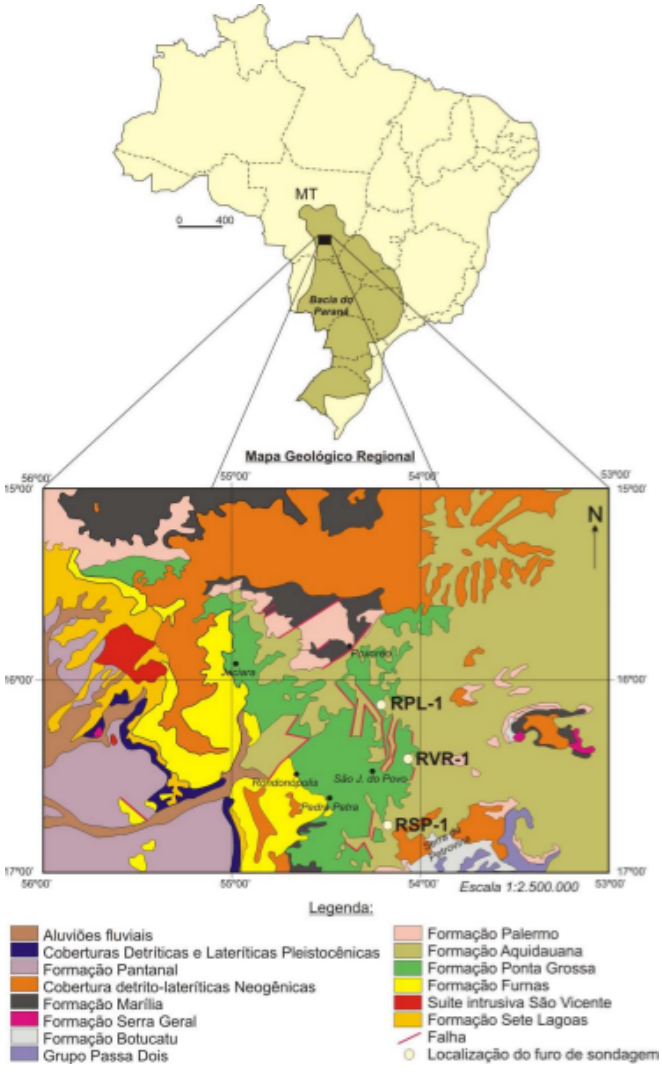


Figura 4.1: Localização dos Furos de Sondagem RSP-1, RVR-1 e RPL-1 (extraído de (Brazil, 2004)).

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 4.1). Detalhes do processo de obtenção dos dados podem ser encontrados em (Brazil, 2004).

Tabela 4.1: Petrográficos analisados - Base de Dados Paleosul

Qz Monocristalino	Siderite nodules
Qz Policristalino	Siderita
K-Feldspato Detrítico	Pirita
Plagioclase Detrítica	Caolinita
Mica	Super Cresc. Qtz
Minerais Pesados	Albitização do Feldspato
Por bioturbação	Carbonato
Glauconita	Outros (Ti-oxidos, super cres. F)
Bioclasto	Por. Intergranular
Ooide Goetita	Por. Intragranular
Ooide Berthierine	Oversized
Substituição de grão Berthierine (F, M)	Moldic
Argilomineral não-identificado	

De acordo com os resultados obtidos com a análise quantitativa dos dados petrográficos foi possível individualizar os constituintes (pela classificação convencional realizada pelo geólogo/petrólogo) em 3 distintas petrofácies, de acordo com (Brazil, 2004): PG-3, PG-2 e PG-1. A Tabela 4.2 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. A Figura 4.2 apresenta a distribuição das amostras. Nessa figura, foi empregado um procedimento de redução de dimensionalidade por meio do método de Análise de Componentes Principais (PCA) e são mostradas as coordenadas das duas primeiras componentes principais. Para a base de dados Paleosul a primeira componente explica 33.23% da variabilidade e a segunda componente 22.06%. Com essas duas componentes em torno de 55% da variabilidade dos dados está representada.

4.1.2 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 4.3 apresenta a localização dos furos de sondagem.

Tabela 4.2: 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%

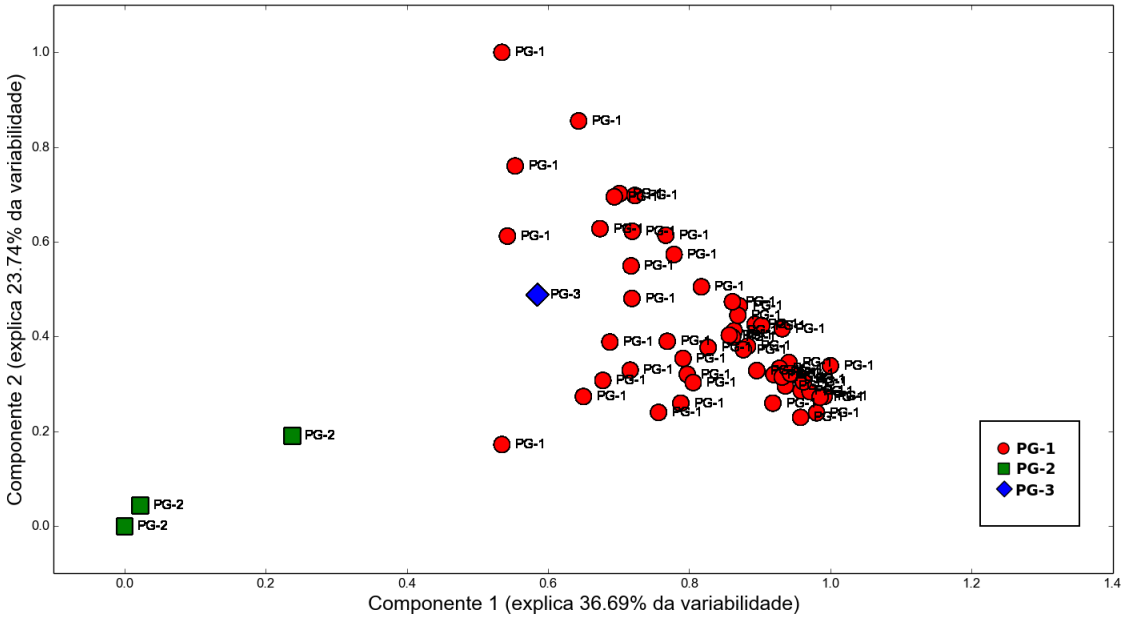


Figura 4.2: Distribuição das Amostras - Paleosul

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 (Oliveira e Pereira, 2009). Nas amostras provenientes destes poços, foram contabilizadas, para cada lâmina, as porcentagens de 22 constituintes (Tabela 4.3):

De acordo com os resultados obtidos com a análise quantitativa dos dados petrográficos foi possível individualizar os constituintes (pela classificação convencional realizada pelo geólogo/petrólogo) em 6 diferentes petrofácies, de acordo com (Oliveira, 2009): PT-1, PT-2, PT-3, PT-4, I-1 e I-2. A Tabela 4.4 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%).

A Figura 4.4 apresenta a distribuição das amostras. Nessa figura, foi empregado um procedimento de redução de dimensionalidade por meio do método de Análise de

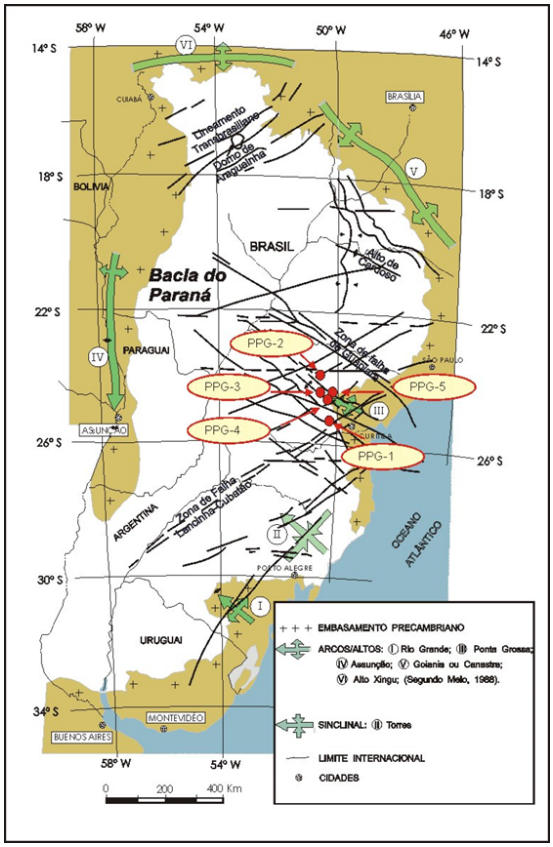


Figura 4.3: Localização dos Furos de Sondagem PPG-1, PPG-2, PPG-3, PPG-4 e PPG-5 (extraído de (Cevolani *et al.*, 2011)).

Tabela 4.3: Petrográficos analisados - Base de Dados Tibagi

Bioclasto	Quartzo
Cresc. Sec. Qtz	Feldspato
Caolinita	Muscovita
Ilita/Smectita	Opaco
Pirita	Turmalina
Siderita	Zircão
Cim. Carbonático	Rutilo
Cim. Silicoso	Glauconita
Cim. Ferruginoso	Clorita
Por. Intergranular	Pseudo Matriz
Por. Intragranular	Litoclasto

Componentes Principais (PCA) e são mostradas as coordenadas das duas primeiras componentes principais. Para a base de dados Tibagi a primeira componente explica 40.15% da variabilidade e a segunda componente 27.41%. Com essas duas componentes em torno de 67% da variabilidade dos dados está representada.

Tabela 4.4: Petrofácies Tibagi x Nº de amostras

Petrofácies	N^o de amostras	Percentual
PT-2	1	2.27%
I-2	1	2.27%
PT-3	2	4.55%
PT-4	5	11.37%
I-1	7	15.91%
PT-1	28	63.63%

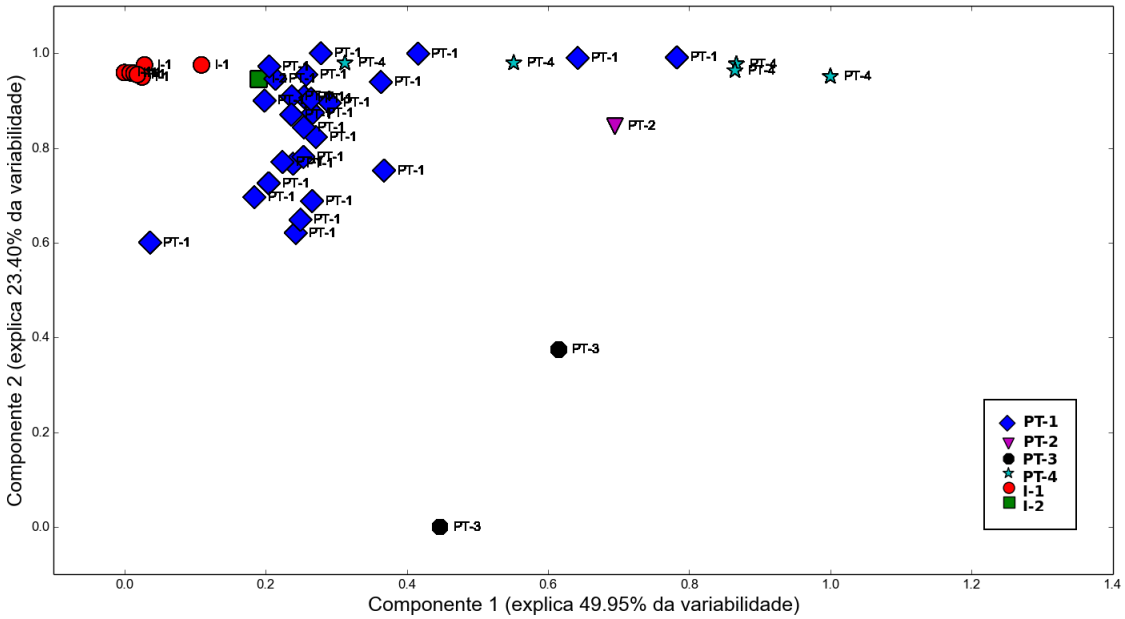


Figura 4.4: Distribuição das Amostras - Tibagi

4.1.3 Paraná+

Para criar esta base de dados Paraná+ realizou-se a junção das bases Paleosul e Tibagi, totalizando 104 amostras. Foram analisados 11 constituintes em comum às duas bases que foram reunidas em uma única.

Tabela 4.5: Petrográficos analisados - Base de Dados Paraná+

Quartzo	Pirita
Feldspato	Siderita
Glauconita	Cim. Carbonático
Bioclasto	Por. Intergranular
Cresc. Sec. Qtz	Por. Intragranular
Caolinita	

As amostras foram classificadas em 9 petrofácies distintas pela classificação manual((Brazil, 2004), (Oliveira, 2009)): PT-1, PT-2, PT-3, PT-4, I-1, I-2, PG-1, PG-2 e PG-3. O número de amostras para cada petrofácies pode ser visto na Tabela 4.6. A Figura 4.5 apresenta a distribuição das amostras. Nessa figura, foi empregado um procedimento de redução de dimensionalidade por meio do método de Análise de Componentes Principais (PCA) e são mostradas as coordenadas das duas primeiras componentes principais. Para a base de dados Paraná+ a primeira componente explica 62.26% da variabilidade e a segunda componente 24.07%. Com essas duas componentes em torno de 86% da variabilidade dos dados está representada.

Tabela 4.6: Petrofácies Paraná+ x N° de amostras

Petrofácies	$N°$ de amostras	Percentual
I-2	1	1.0%
PG-3	1	1.0%
PT-2	1	1.0%
PT-3	2	1.9%
PG-2	3	2.9%
PT-4	5	4.8%
I-1	7	6.7%
PT-1	28	26.9%
PG-1	56	53.8%

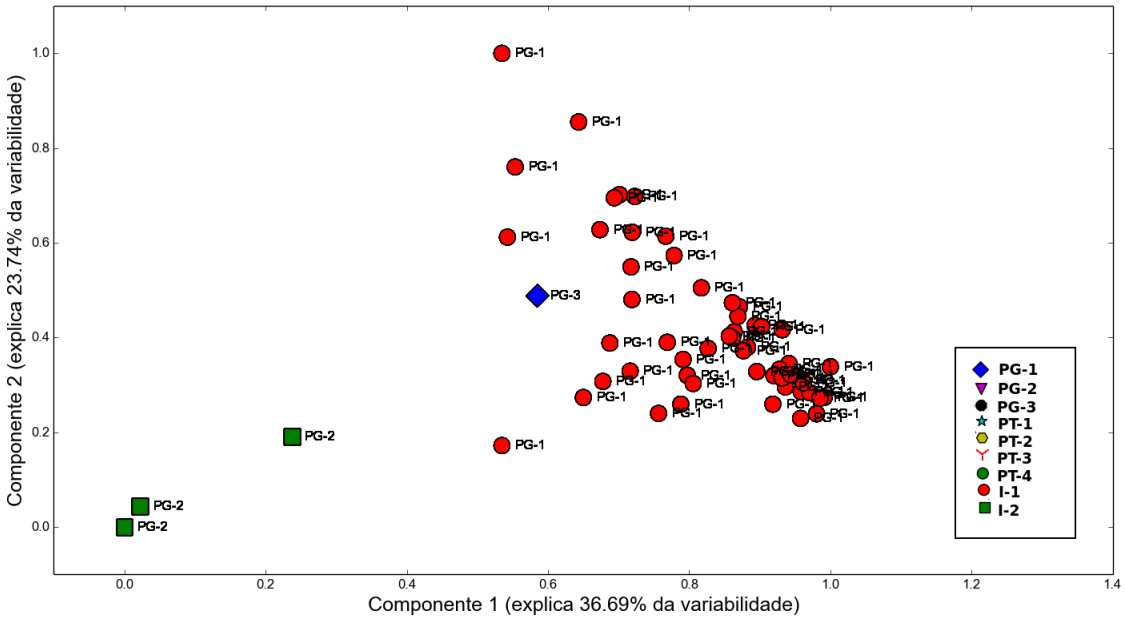


Figura 4.5: Distribuição das Amostras - Paraná+

4.1.4 Mucuri

As bacia do Espírito Santo e Mucuri localizam-se ao longo do litoral centro-norte do Estado do Espírito Santo e sul do Estado da Bahia. Seu limite sul é a feição geológica conhecida como Alto de Vitória, que a separa da Bacia de Campos, enquanto seu limite norte, com a Bacia de Cumuruxatiba, é apenas geográfico. A bacia possui uma área sedimentar total de 123.130 km^2 até a lâmina d'água de 3.000 m (17.900 km^2 em terra). A Figura 4.6 apresenta a localização da Bacia do Espírito Santo-Mucuri.

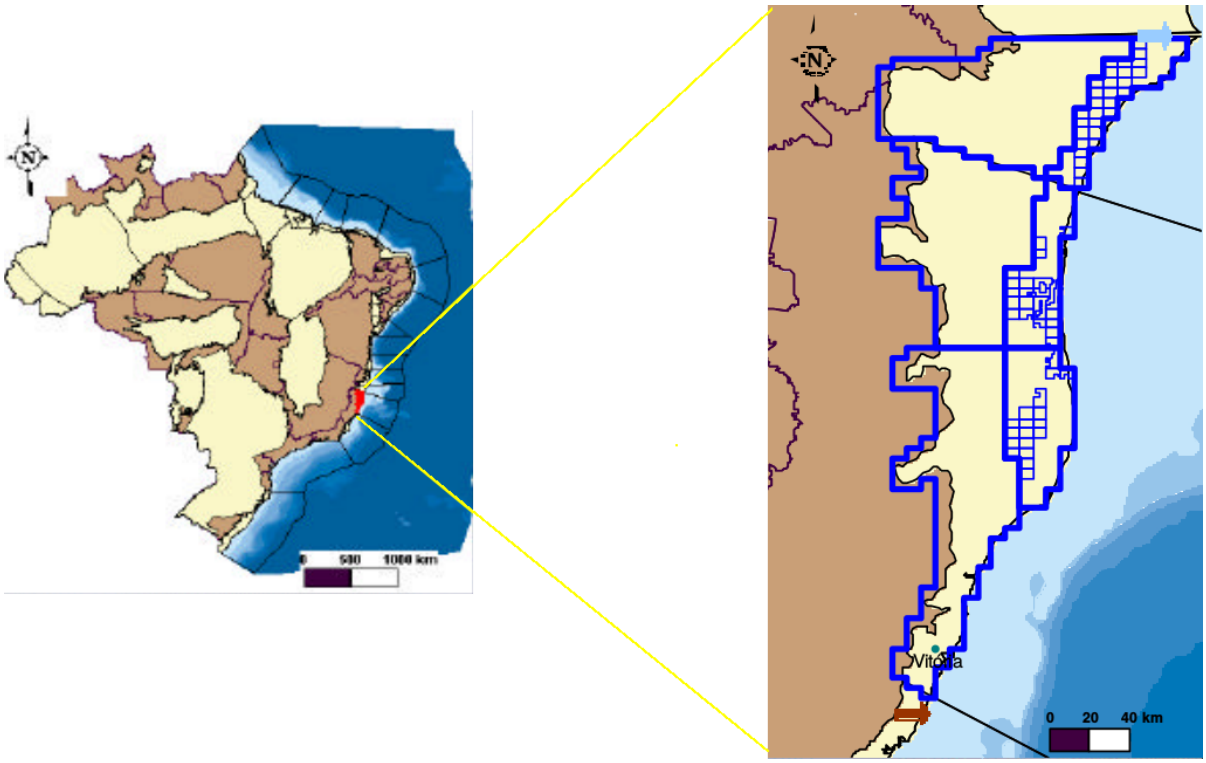


Figura 4.6: Localização da Bacia do Espírito Santo-Mucuri (extraído de (de Castro, 2003)).

Nas amostras provenientes desta bacia, foram contabilizadas, para cada lâmina, as porcentagens de 33 constituintes (Tabela 4.7):

A permeabilidade foi classificada de acordo com da Silva *et al.* (2015) em Baixa, Razoável, Boa e Excelente, de acordo com a faixa de permeabilidade que pode ser vista na Tabela 4.8. O número de amostras para cada classe pode ser visto na Tabela 4.9. A Figura 4.7 apresenta a distribuição das amostras. Nessa figura, foi empregado um procedimento de redução de dimensionalidade por meio do método de Análise de Componentes Principais (PCA) e são mostradas as coordenadas das duas primeiras componentes principais. A primeira componente explica 99.90% da variabilidade e a segunda componente 0.03% da variabilidade dos dados está representada.

Tabela 4.7: Petrográficos analisados - Base de Dados Mucuri

Índice de Embalagem	Caolinita
Qtz monocristalino	Clorita
Qtz policristalino	Calcita
K-feldspato	Calcita Ferruginosa
Plagioclásio	Dolomita
Biotita	Dolomita Ferruginosa
Muscovita	Qtz Autigênico
Granada	Cres. Sec. Feldspato
Fragmentos Liticos	Pirita
Argilas de grãos substituídos	Outras
Acessórios para Grãos	Por. Intergranular
Matriz Deposicional	Por. Intragranular
Pseudo Matriz	Porosidade Moldic
Argila de enchimento dos poros	Fratura Porosidade
Argilas Substituídas	Grande demais
Cutículas de argila infiltradas	Encolhimento
Argila Autigênico	

Tabela 4.8: Classes Permeabilidade x Faixa de Permeabilidade

Classes Permeabilidade	Faixa de Permeabilidade
Baixa (Lo)	< 1
Razoável (Fa)	1 - 10
Boa (Go)	10 - 100
Excelente (Ex)	> 100

Tabela 4.9: Classes Permeabilidade Mucuri x N^o de amostras

Classes Permeabilidade	N^o de amostras	Percentual
Fa	21	15.79%
Ex	22	16.54%
Go	32	24.06%
Lo	58	43.61%

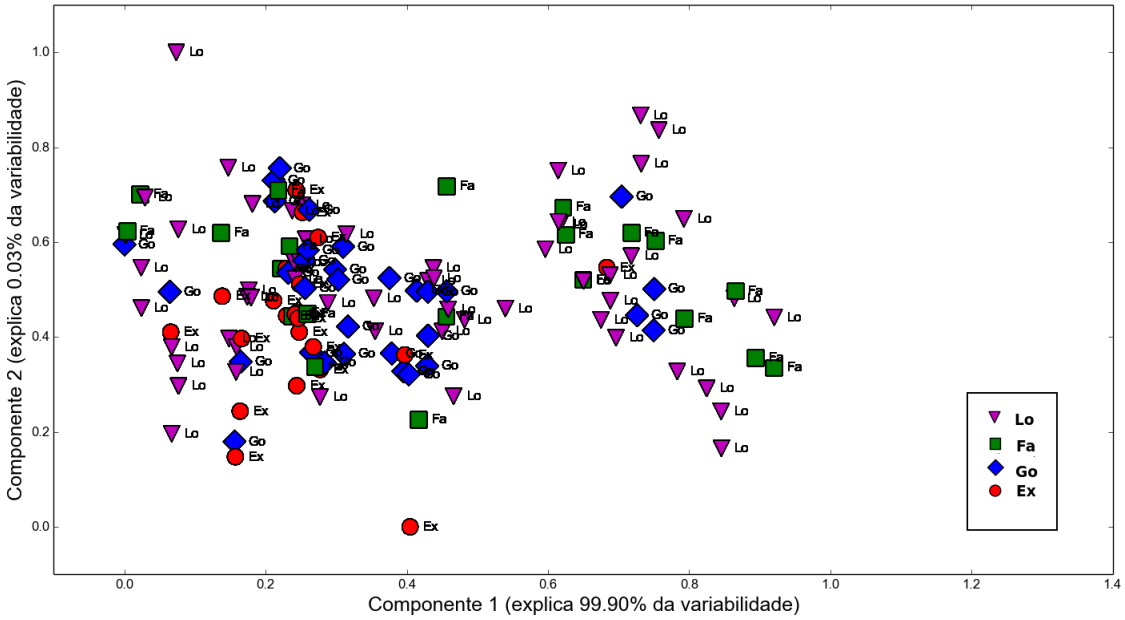


Figura 4.7: Distribuição das Amostras - Mucuri

4.2 Procedimentos Realizados

A metodologia empregada epara a classificação envolve o uso de normalização dos dados, validação cruzada, grid-search, seletores de características, classificadores, técnica de balanceamento dos dados e métricas para avaliação dos classificadores. Para o agrupamento envolve K-Means, Análise de Silhueta e Análise de Componentes Principais (ACP). Os métodos utilizados estão no pacote *Scikit-Learn* em Python (Pedregosa *et al.*, 2011). Inicialmente normalizou-se todas as bases de dados (Paleosul, Tibagi, Paraná+ e Mucuri). Comparou-se três estratégias:

1. Após a normalização, aplicar de técnicas de validação cruzada para dividir o conjunto de dados em conjunto de treinamento e conjunto de teste, grid-search e classificadores .

2. A utilização de seletores de características como pré-processamento dos dados e em seguida a aplicação da estratégia 1.

3. A realização do balanceamento dos dados antes de normalizar os dados e em seguida o emprego a estratégia 1.

Foram realizadas 30 execuções para todas as bases de dados utilizadas e K = 5 nas técnicas de validação cruzada, totalizando 150 execuções. O esquema apresentado na

Figura 4.8 ilustra os procedimentos. A faixa de variação dos parâmetros de cada modelo nos procedimentos pode ser visto na Tabela 4.10.

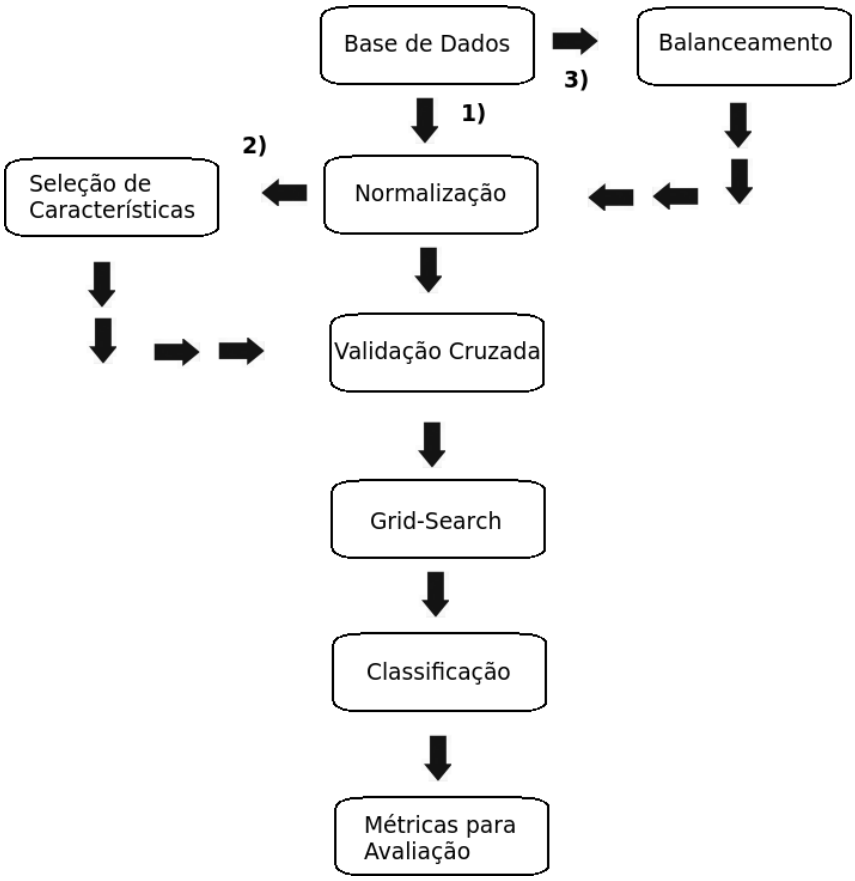


Figura 4.8: Esquema ilustrando o procedimento computacional

4.3 Resultados e Discussões

4.3.1 Paleosul

4.3.1.1 Dados Desbalanceados

A Tabela 4.11 apresenta a média e o desvio padrão da acurácia, F1, Kappa, Erro Quadrático Médio (MSE) e a média do tempo de processamento para validação cruzada KF. Realizou-se uma análise estatística através do teste não-paramétrico de Wilcoxon e foi constatado que, usando o KF, os métodos KNN e LinSVM apresentam desempenhos semelhantes ao RBFSVM. Observa-se que para a Acurácia o RBFSVM obteve melhor desempenho, por esse motivo foi escolhido como base para o teste de Wilcoxon. Houve uma diferença de 1.5% para o DT, de 0% para o KNN, de 0% para o LinSVM, 0.7%

Tabela 4.10: Parâmetros dos modelos utilizados no grid search com validação cruzada. A função dos parâmetros pode ser explicado em (Pedregosa *et al.*, 2011) e (Nissen, 2005)

Modelo	Parâmetros	Variação
KNN	nº de vizinhos	1, 2, 3, 4, 5, 6, 10
	pesos	uniforme, distância
DT	profundidade máxima	Nenhum, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 50
LinSVM	C	10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}, 10^{7}, 10^{8}
RBFSVM	C	10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}, 10^{7}, 10^{8}
	γ	10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}
RF	no. min. de amostras em folhas recém-criadas	1, 3, 5, 9, 17
MLP0	número de camadas ocultas e neurônios por camada	[5], [5, 5], [5, 5, 5], [5, 5, 5, 5], [10], [10, 10], [10, 10, 10], [50]
	algoritmo	tnc, l-bfgs, sgd, rprop, genetic
	conectividade	conectada,totalmente conectada
	renormalização	Verdadeira, Falsa

para a MLP0 e 0.3% para o RF. Estas diferenças, com exceção do KNN, do LinSVM e do MLP0, são consideradas estatisticamente significantes pelo teste de Wilcoxon. Resultados similares foram obtidos para o F1, RECALL e MSE. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado dos métodos RBFSVM, KNN e LinSVM foi considerado Quase Perfeito, para os demais métodos foi considerado Substancial.

Os resultados para a validação cruzada SKF também são mostrados na Tabela 4.11. Nota-se que o RBFSVM foi o método com maior acurácia, F1 e RECALL. As diferenças na acurácia foram de 9.6% para o DT, 0% para o KNN, 0% para o LinSVM, 0.8% para o MLP0 e 0% para o RF, com o mesmo comportamento observado no RECALL e F1. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado do MLP0 foi considerado Substancial, para os demais métodos foi considerado Quase Perfeito.

Para esta base de dados todos os classificadores obtiveram resultados muito bons,

Tabela 4.11: Média e Desvio Padrão da Acurácia, F1, RECALL, Kappa e Erro Quadrático Médio para os métodos KF e SKF – Base de Dados Paleosul. A coluna TEMPO indica o tempo médio de processamento em segundos do classificador para uma execução. 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.

MODELO	CV	ACURÁCIA	F1	RECALL	KAPPA	MSE	TEMPO
DT	KF	0.967 (0.016)	0.962 (0.014)	0.967 (0.016)	0.725 (0.138)	0.071 (0.042)	0.001
KNN	KF	**0.982 (0.007)***	**0.974 (0.011)***	**0.982 (0.007)***	**0.837 (0.121)**	0.053 (0.024)	0.003
LinSVM	KF	**0.982 (0.007)***	**0.974 (0.010)***	**0.982 (0.007)***	0.833 (0.119)	0.068 (0.007)	0.003
MLP0	KF	0.975 (0.013)	0.968 (0.014)	0.975 (0.013)	0.771 (0.156)	0.083 (0.030)	0.041
RBFSVM	KF	**0.982 (0.007)**	**0.974 (0.010)**	**0.982 (0.007)**	0.833 (0.119)	0.068 (0.007)	0.003
RF	KF	0.979 (0.011)	0.970 (0.015)	0.979 (0.011)	0.789 (0.165)	0.063 (0.022)	2.971
DT	SKF	0.976 (0.011)	0.970 (0.009)	0.976 (0.011)	0.807 (0.076)	0.053 (0.030)	0.001
KNN	SKF	**0.983 (1.192e-08)***	**0.975 (0.0003875)***	**0.983 (1.192e-08)***	**0.852 (0.006)**	0.060 (0.017)	0.002
LinSVM	SKF	**0.983 (1.204e-08)***	**0.975 (1.204e-08)***	**0.983 (1.204e-08)***	0.850 (1.204e-08)	0.067 (7.526e-10)	0.003
MLP0	SKF	0.975 (0.013)	0.970 (0.009)	0.975 (0.013)	0.799 (0.083)	0.076 (0.040)	0.051
RBFSVM	SKF	**0.983 (1.192e-08)***	**0.975 (1.192e-08)***	**0.983 (1.192e-08)***	0.850 (1.192e-08)	0.067 (7.451e-10)	0.003
RF	SKF	**0.983 (1.192e-08)**	**0.975 (0.0004094)**	**0.983 (1.192e-08)**	**0.852 (0.006)**	0.059 (0.018)	2.973

acima de 96% de acerto. Apesar da base de dados estar desbalanceada, o DT não criou árvores tendenciosas conseguindo um bom desempenho. A rede neural de múltiplas camadas MLP0 não teve a convergência dificultada pela quantidade de conexões na arquitetura, convergindo para a solução no processo de minimização do funcional do erro médio quadrático. Como os dados foram previamente normalizados, o KNN foi um método eficiente. Os SVMs originaram apreciável acurácia, por serem classificadores eficazes, na maioria dos problemas testados na literatura os métodos apresentam bons resultados e por utilizarem técnica de relaxamento minimizando o risco de overfitting, que é um problema que aparece em dados com grande dimensionalidade e esparsos, como é o caso da base de dados utilizada. A Floresta Aleatória (RF) geralmente obtém bons resultados, uma vez que gera uma gama de árvores de decisão e opta pela classe mais votada. Pelo fato do RF selecionar atributos para a construção de árvores, ele conseguiu criar árvores menos tendenciosas otimizando o resultado.

A Figura 4.9 apresenta os resultados dos dados sem seletores de características e sem o balanceamento, para os classificadores DT, KNN, LinSVM, RBFSVM, MLP0 e RF. Nota-se que o uso da validação cruzada SKF não fez muita diferença, uma vez que os resultados ficaram em torno de 97%. A partir da Tabela 4.11 nota-se que com a aplicação do SKF o coeficiente Kappa teve nível de concordância Quase Perfeito para todos os métodos com exceção do DT. Esse resultado era esperado, já que o SKF apresenta uma melhor distribuição entre as classes das amostras entre os subconjuntos de treinamento e de teste, permitindo que estes conjuntos tenham a mesma proporção de amostras de cada classe (petrofácies).

4.3.1.2 Dados Desbalanceados – Seleção de Características

A Figura 4.10 apresenta o gráfico de barras com a acurácia encontrada através da emprego dos métodos de seleção de características (SKB, SP, SFpr e Fwe) e sem o uso de nenhuma abordagem (NFS) além dos métodos de classificação combinados com o SKF. Para os classificadores KNN, LinSVM e o MLP0 o uso do Fwe resultou em uma acurácia mais alta. Para o RBFSVM, o método SKB gerou uma acurácia mais próxima de 1.0. No DT, o melhor resultado foi com o SFpr. Para o RF, o NFS obteve melhor desempenho.

Para os métodos KNN e o RBFSVM todos os métodos obtiveram o mesmo valor de acurácia. No caso do LinSVM todos os métodos, com exceção do SP, resultaram no mesma

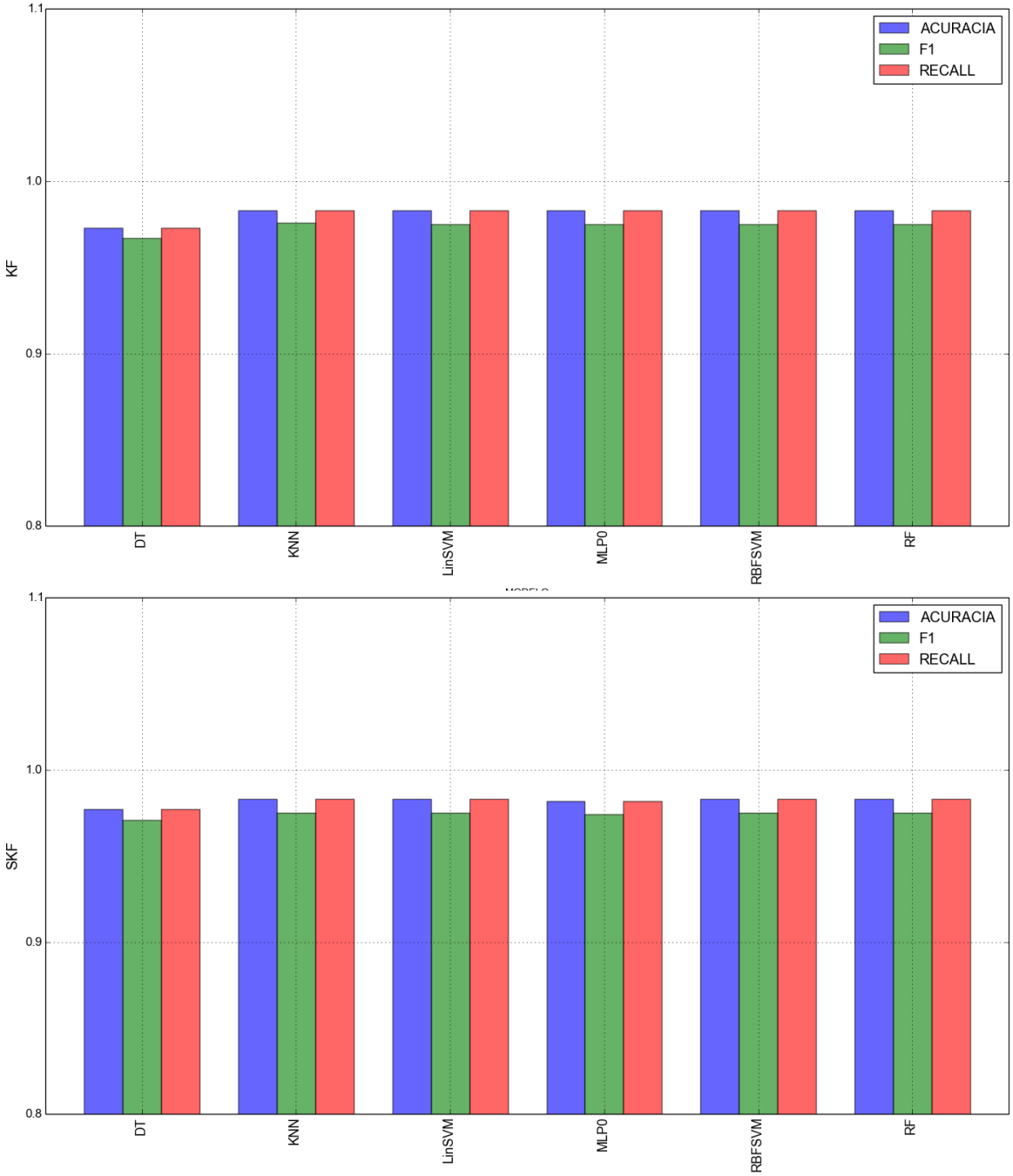


Figura 4.9: KF e SKF – Paleosul (30 iterações, K= 5) - Acurácia, F1 e RECALL

acurácia. No caso do MLP0 o SKB apresentou melhor acurácia. Foram selecionadas as 5 características com maior pontuação. Esse critério fez com que a árvore criada fosse menos tendenciosa em relação ao restante dos seletores. No caso do RF por ele já selecionar características para a construção das árvores de decisão o melhor desempenho foi obtido pelo NFS, sem nenhum seletor de característica ser aplicado. Os métodos SKB e Fwe também resultaram no mesmo desempenho. Nesse caso aplicar seletores de características não é uma boa alternativa, uma vez que só aumenta o tempo computacional. Para o

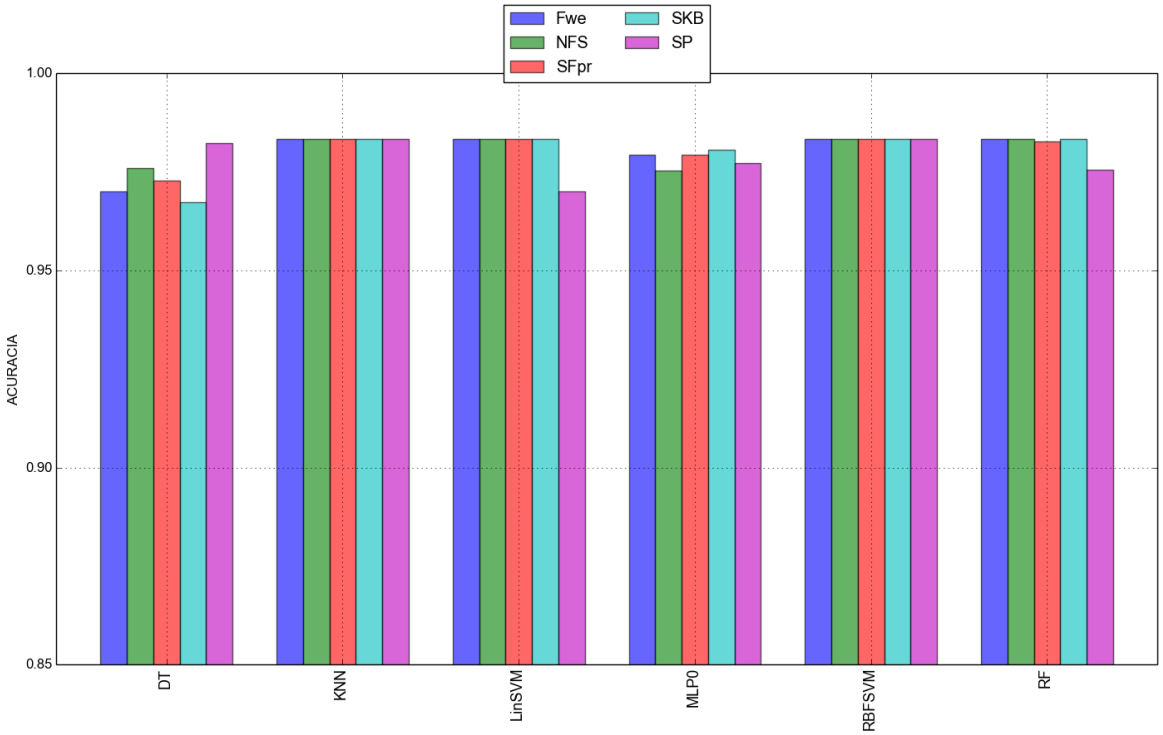


Figura 4.10: Seleção de Características – Paleosul (30 iterações, K= 5) - Acurácia

DT, selecionar 10% das características com maiores pontuações foi a melhor escolha em comparação com outros seletores e sem aplicação de seletores.

4.3.1.3 Dados Balanceados

A Tabela 4.12 exibe a média e o desvio padrão da acurácia, F1, Kappa, Erro Quadrático Médio (MSE) e a média do tempo de processamento para validação cruzada SKF para os dados balanceados. Efetuou-se uma análise estatística através do teste não-paramétrico de Wilcoxon e constatou-se que os métodos KNN, LinSVM e RBFSVM apresentam desempenhos significamente semelhantes ao RF.

Observa-se que para a Acurácia o RF obteve melhor desempenho, por esse motivo foi escolhido como base para o teste de Wilcoxon. Houve uma diferença de 0.8% para o DT, de 0% para o KNN, de 0% para o LinSVM, 1.0% para a MLP0 e 0% para o RBFSVM. Estas diferenças, com exceção do KNN, do LinSVM e do RBFSVM, são consideradas estatisticamente significantes pelo teste de Wilcoxon. Resultados similares foram obtidos para o F1, RECALL e MSE. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado dos métodos foi considerado Quase Perfeito. A Figura 4.11 ilustra os resultados com os dados balanceados, para os classificadores DT, KNN, LinSVM, RBFSVM, MLP0 e RF.

Tabela 4.12: Média e Desvio Padrão da Acurácia, F1, RECALL, Kappa e Erro Quadrático Médio para o método SKF - Dados Balanceados Paleosul. A coluna TEMPO indica o tempo médio de processamento em segundos do classificador para uma execução. 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.

MODELO	CV	ACURÁCIA	F1	RECALL	KAPPA	MSE	TEMPO
DT	SKF	0.992 (0.005)	0.992 (0.005)	0.992 (0.005)	0.988 (0.008)	0.014 (0.014)	0.003
KNN	SKF	**1.000 (0)***	**1.000 (0)***	**1.000 (0)***	**1.000 (0)***	0 (0)	0.005
LinSVM	SKF	**1.000 (0)***	**1.000 (0)***	**1.000 (0)***	**1.000 (0)***	0 (0)	0.010
MLP0	SKF	0.990 (0.022)	0.990 (0.022)	0.990 (0.022)	0.985 (0.033)	0.016 (0.024)	0.168
RBFSVM	SKF	**1.000 (0)***	**1.000 (0)***	**1.000 (0)***	**1.000 (0)***	0 (0)	0.016
RF	SKF	**1.000 (0)**	**1.000 (0)**	**1.000 (0)**	**1.000 (0)**	0 (0)	4.737

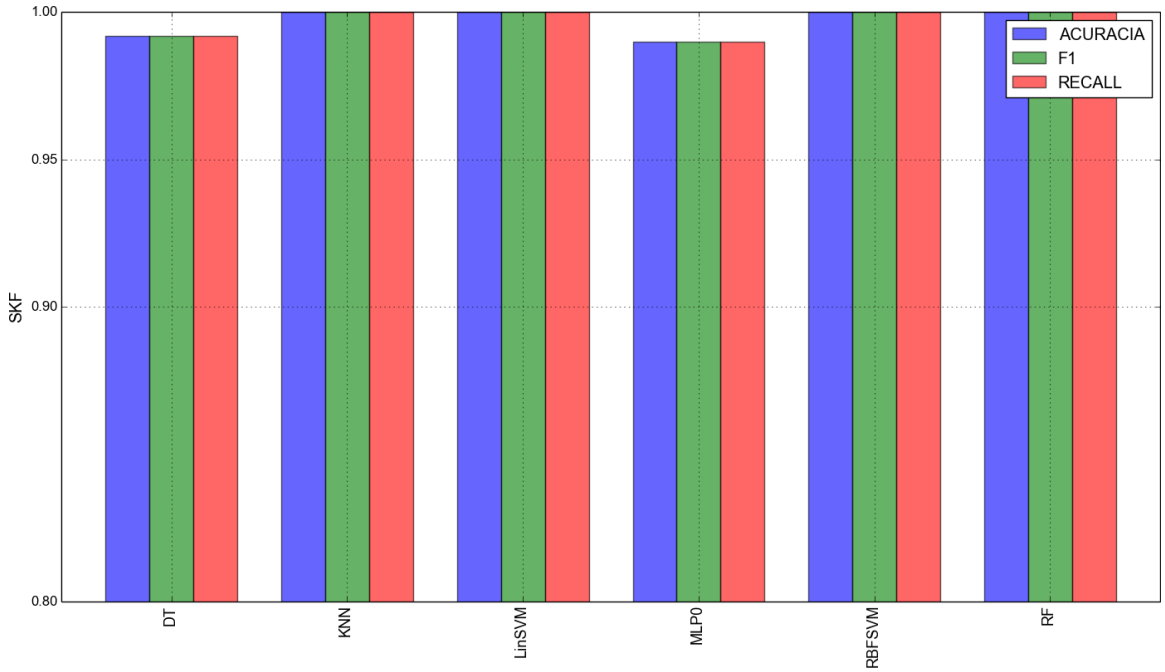


Figura 4.11: SKF – Paleosul (30 iterações, K= 5) - Acurácia, F1 e RECALL

Depois de realizado o balanceamento dos dados e após a análise dos resultados percebe-se que os métodos que geraram a melhor acurácia, F1 e RECALL foi o RF, RBFSVM, LinSVM e o KNN. Acertando 100% das classificações. Nesse caso todos os métodos se mostraram úteis para individualização desta base de dados, tanto desbalanceados como balanceados. No entanto, nota-se com base nos resultados apresentados que o balanceamento dos dados foi um fator que proporcionou ganho de desempenho em todos os métodos.

4.3.2 Tibagi

4.3.2.1 Dados Desbalanceados

A Tabela 4.13 apresenta a média e o desvio padrão da acurácia, F1, Kappa, Erro Quadrático Médio (MSE) e a média do tempo de processamento para validação cruzada KF. Realizou-se uma análise estatística através do teste não-paramétrico de Wilcoxon e foi constatado que, usando o KF, o método LinSVM apresenta desempenho significamente semelhante ao RBFSVM. Observa-se que para a Acurácia o RBFSVM obteve melhor desempenho, por esse motivo foi escolhido como base para o teste de Wilcoxon. Houve uma diferença de 7.2% para o DT, de 8.3% para o KNN, de 0.2% para o LinSVM, 9.5% para a MLP0 e 2.3% para o RF. Estas diferenças, com exceção do LinSVM, são

consideradas estatisticamente significantes pelo teste de Wilcoxon. Resultados similares foram obtidos para o F1, RECALL e MSE. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado dos métodos RBFSVM, e LinSVM foi considerado Substancial, para os demais métodos foi considerado Moderado.

Os resultados para a validação cruzada SKF podem ser vistos na Tabela 4.13. Nota-se que o RBFSVM foi o método com maior acurácia, F1 e RECALL. As diferenças na acurácia foram de 7.4% para o DT, 7.7% para o KNN, 0.2% para o LinSVM, 8.2% para o MLP0 e 2.0% para o RF, com o mesmo comportamento observado no RECALL e F1. Estas diferenças, com exceção do LinSVM, são consideradas estatisticamente significantes pelo teste de Wilcoxon. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado do RBFSVM e do LinSVM foi considerado Substancial, para os demais métodos foi considerado Moderado.

O DT não obteve bons resultados comparado com os outros métodos pois as árvores de decisão criam árvores tendenciosas se algumas classes possuem mais amostras que outras. O MLP0 é uma rede neural de múltiplas camadas, o que faz com que a quantidade de conexões na arquitetura e na rede seja grande o que dificulta a convergência para a solução no processo de minimização do funcional do erro médio quadrático. O KNN não obteve bom resultado por ser um método simples. A normalização dos dados não foi suficiente para que seu desempenho melhorasse. Os SVMs obtiveram apreciável acurácia, por serem classificadores eficientes, na maioria dos problemas testados na literatura os métodos apresentam bons resultados, e por utilizarem técnica de relaxamento minimizando o risco de overfitting, que é um problema que aparece em dados com grande dimensionalidade esparsos, como é o caso da base de dados utilizada. A diferença é o que o RBFSVM permite resolver problemas, originalmente, não linearmente separáveis, através do mapeamento para um espaço de maior dimensão. Um árvore de decisão sozinha não obteve bons resultados, mas o RF por realizar a seleção de atributos para a construção de árvores, criou árvores menos tendenciosas ocasionando no bom desempenho.

A Figura 4.12 exibe os resultados dos dados sem seletores de características e sem o balanceamento, para os classificadores DT, KNN, LinSVM, RBFSVM, MLP0 e RF. Nota-se que o uso da validação cruzada SKF resultou em melhores acurácia, F1 e RECALL para todos os métodos de classificação estudados. Esse resultado era esperado, já que o SKF

Tabela 4.13: Média e Desvio Padrão da Acurácia, F1, RECALL, Kappa e Erro Quadrático Médio para os métodos KF e SKF – Base de Dados Tibagi. A coluna TEMPO indica o tempo médio de processamento em segundos do classificador para uma execução. 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.

MODELO	CV	ACURÁCIA	F1	RECALL	KAPPA	MSE	TEMPO
DT	KF	0.703 (0.054)	0.663 (0.066)	0.703 (0.054)	0.508 (0.097)	1.280 (0.322)	0.001
KNN	KF	0.692 (0.025)	0.620 (0.032)	0.692 (0.025)	0.467 (0.045)	1.554 (0.190)	0.003
LinSVM	KF	0.773 (0.032)*	0.726 (0.053)	0.773 (0.032)	0.613 (0.069)	0.894 (0.207)	0.007
MLP0	KF	0.680 (0.059)	0.653 (0.056)	0.680 (0.059)	0.485 (0.090)	1.345 (0.466)	0.545
RBFSVM	KF	**0.775 (0.034)**	**0.728 (0.054)**	**0.775 (0.034)**	**0.618 (0.069)**	0.880 (0.178)	0.009
RF	KF	0.752 (0.022)	0.677 (0.030)	0.752 (0.022)	0.554 (0.044)	0.921 (0.170)	4.000
DT	SKF	0.716 (0.043)	0.679 (0.054)	0.716 (0.043)	0.533 (0.073)	1.130 (0.254)	0.001
KNN	SKF	0.713 (0.016)	0.636 (0.026)	0.713 (0.016)	0.494 (0.034)	1.422 (0.184)	0.003
LinSVM	SKF	0.788 (0.028)*	0.744 (0.052)	0.788 (0.028)*	0.641 (0.062)	0.873 (0.195)	0.007
MLP0	SKF	0.708 (0.058)	0.687 (0.056)	0.708 (0.058)	0.539 (0.090)	1.122 (0.420)	0.545
RBFSVM	SKF	**0.790 (0.031)**	**0.752 (0.051)**	**0.790 (0.031)**	**0.648 (0.061)**	0.857 (0.185)	0.009
RF	SKF	0.770 (0.018)	0.703 (0.027)	0.770 (0.018)	0.593 (0.033)	0.824 (0.114)	3.989

apresenta uma melhor distribuição entre as classes das amostras entre os subconjuntos de treinamento e de teste, permitindo que estes conjuntos tenham a mesma proporção de amostras de cada classe (petrofácies).

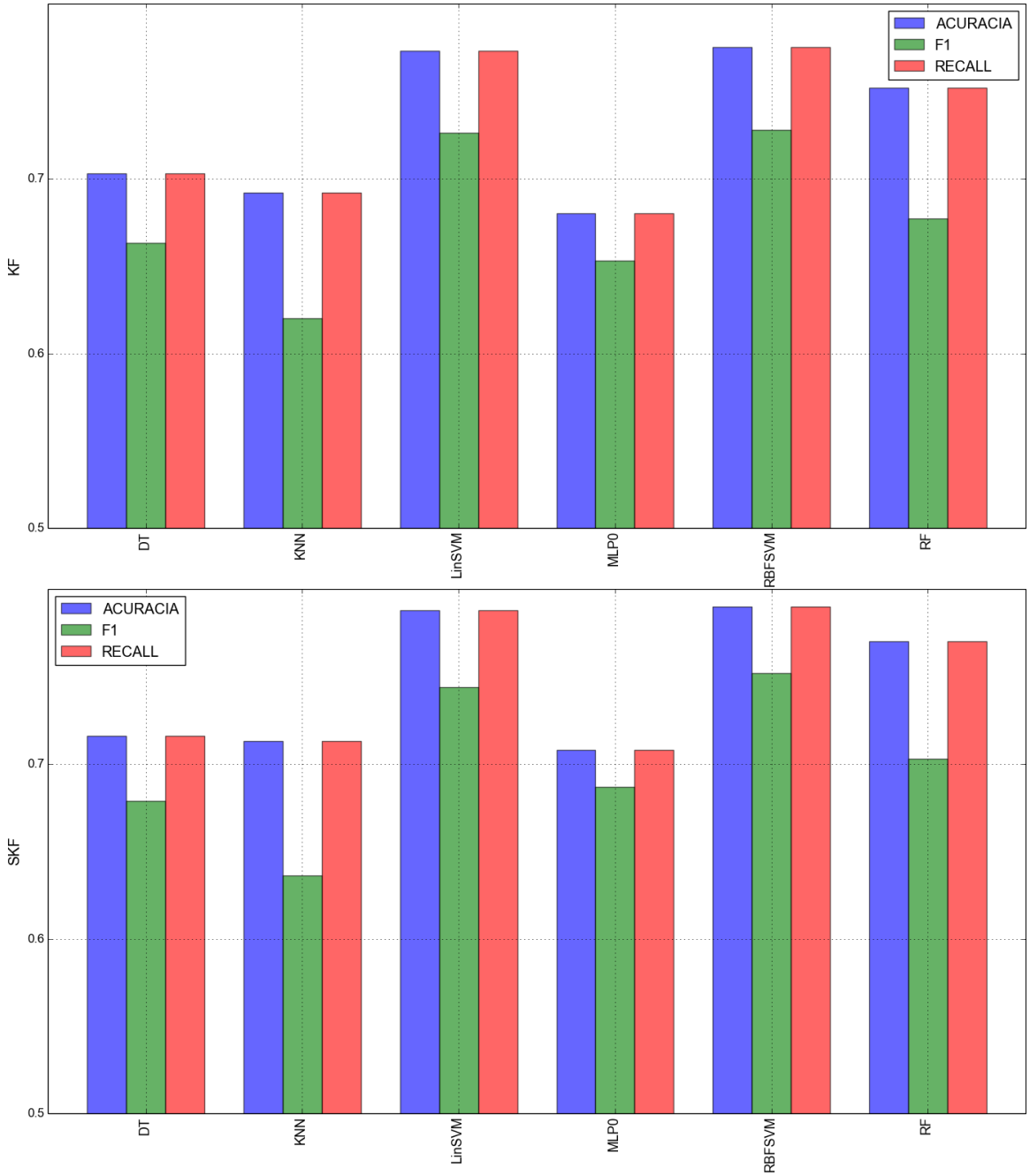


Figura 4.12: KF e SKF – Tibagi (30 iterações, K= 5) - Acurácia, F1 e RECALL

4.3.2.2 Dados Desbalanceados – Seleção de Características

A Figura 4.13 apresenta o gráfico de barras com a acurácia encontrada através da emprego dos métodos de seleção de características (SKB, SP, SFpr e Fwe) e sem o uso de

nenhuma abordagem (NFS) além dos métodos de classificação combinados com o SKF. Para os classificadores KNN, LinSVM e o MLP0 o uso do Fwe resultou em uma acurácia mais alta. Para o RBFSVM, o método SKB gerou uma acurácia mais próxima de 1.0. No DT, o melhor resultado foi com o SFpr. Para o RF, o NFS obteve melhor desempenho.

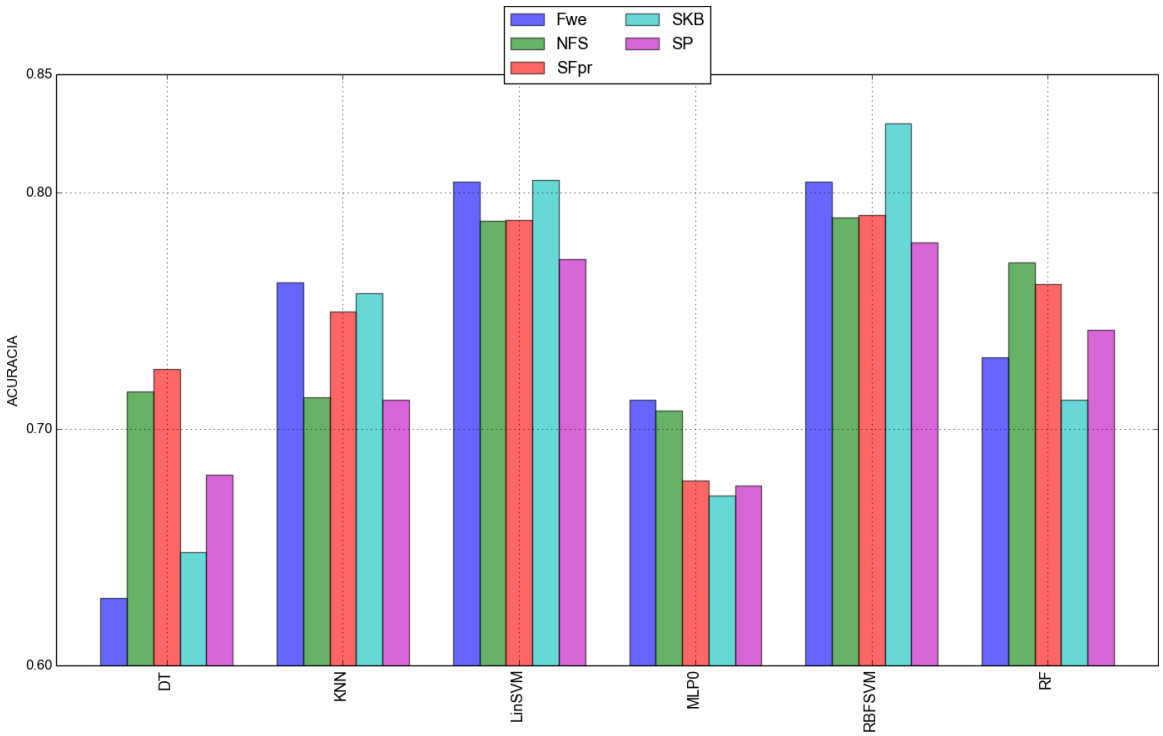


Figura 4.13: Seleção de Características – Tibagi (30 iterações, K= 5) - Acurácia

A minimização do erro Fwe foi uma boa medida para selecionar características nos métodos KNN, LinSVM e MLP, melhorando o desempenho em relação ao NFS. Para o RBFSVM o SKB apresentou melhor acurácia. Foram selecionadas as 5 características com maior pontuação. No DT encontrar o teste baseado em falsas detecções aumentou o valor da acurácia em relação ao resultado obtido sem nenhum emprego de seletores de características. No caso do RF por ele já selecionar características para a construção das árvores de decisão o melhor desempenho foi obtido pelo NFS, sem nenhum seletor de característica ser aplicado.

4.3.2.3 Dados Balanceados

A Tabela 4.14 mostra a média e o desvio padrão da acurácia, F1, Kappa, Erro Quadrático Médio (MSE) e a média do tempo de processamento para validação cruzada SKF para os dados balanceados. Efetuou-se uma análise estatística através do teste não-paramétrico de Wilcoxon e constatou-se que o método RBFSVM apresentam desempenhos

significamente semelhantes ao RF.

Observa-se que para a Acurácia o RF obteve melhor desempenho, por esse motivo foi escolhido como base para o teste de Wilcoxon. Houve uma diferença de 3.0% para o DT, de 4.7% para o KNN, de 0.9% para o LinSVM, 1.9% para a MLP0 e 0.4% para o RBFSVM. Estas diferenças, com exceção do RBFSVM, são consideradas estatisticamente significantes pelo teste de Wilcoxon. Resultados semelhantes foram obtidos para o F1, RECALL e MSE. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado dos métodos foi considerado Quase Perfeito. A Figura 4.14 ilustra os resultados dos dados com o balanceamento, para os classificadores DT, KNN, LinSVM, RBFSVM, MLP0 e RF.

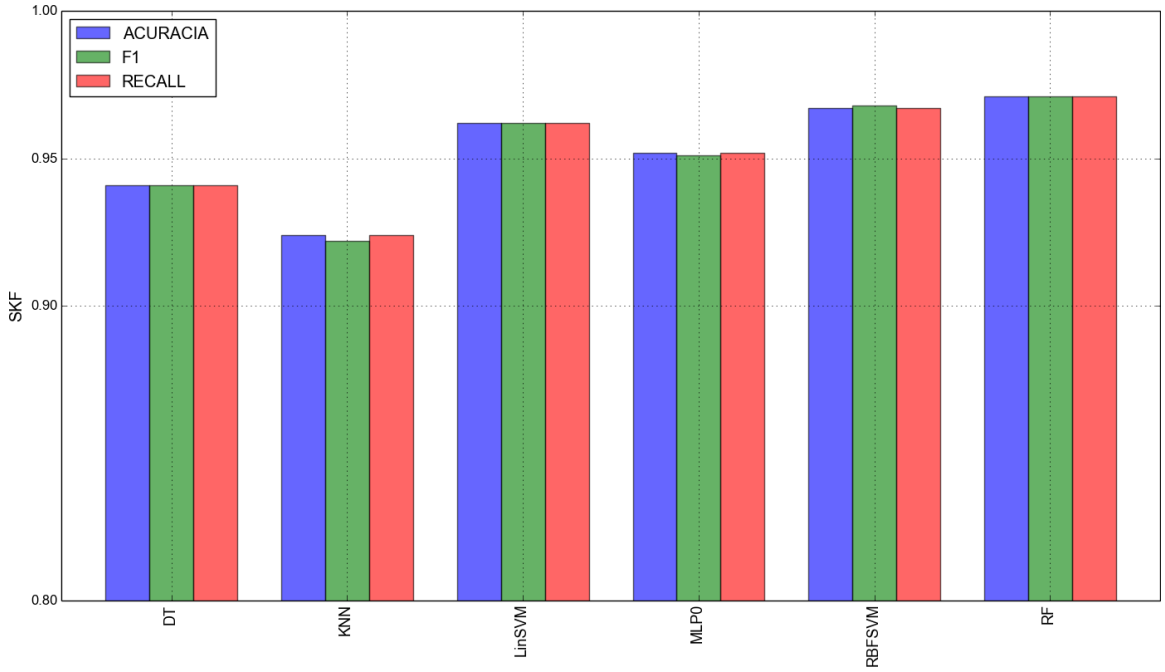


Figura 4.14: SKF – Tibagi (30 iterações, K= 5) - Acurácia, F1 e RECALL

Após o balanceamento dos dados e de uma posterior análise dos resultados observa-se que o método que ocasionou melhor acurácia, F1 e RECALL foi o RF, seguido do RBFSVM. Os dois classificadores foram úteis para a classificação desta base de dados, tanto desbalanceados como balanceados. Os classificadores MLP0, LinSVM, KNN e DT apresentaram bons resultados e mostraram ser uma alternativa para bases de dados balanceadas. Todavia, percebe-se com base nos resultados exibidos que o balanceamento dos dados proporcionou um ganho de desempenho em todos os métodos.

Tabela 4.14: Média e Desvio Padrão da Acurácia, F1, RECALL, Kappa e Erro Quadrático Médio para o método SKF - Dados Balanceados Tibagi. A coluna TEMPO indica o tempo médio de processamento em segundos do classificador para uma execução. 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.

MODELO	CV	ACURÁCIA	F1	RECALL	KAPPA	MSE	TEMPO
DT	SKF	0.941 (0.010)	0.941 (0.010)	0.941 (0.010)	0.929 (0.012)	0.218 (0.053)	0.003
KNN	SKF	0.924 (0.006)	0.922 (0.006)	0.924 (0.006)	0.909 (0.007)	0.215 (0.025)	0.003
LinSVM	SKF	0.962 (0.010)	0.962 (0.010)	0.962 (0.010)	0.954 (0.012)	0.142 (0.030)	0.015
MLP0	SKF	0.952 (0.018)	0.951 (0.020)	0.952 (0.018)	0.942 (0.022)	0.173 (0.054)	1.157
RBFSVM	SKF	0.967 (0.008)*	0.968 (0.007)8	0.967 (0.008)*	0.960 (0.009)	0.124 (0.017)	0.019
RF	SKF	**0.971 (0.005)**	**0.971 (0.005)**	**0.971 (0.005)**	**0.965 (0.006)**	0.109 (0.005)	4.019

4.3.3 Paraná+

4.3.3.1 Dados Desbalanceados

A Tabela 4.15 apresenta a média e o desvio padrão da acurácia, F1, Kappa, Erro Quadrático Médio (MSE) e a média do tempo de processamento para validação cruzada KF. Realizou-se uma análise estatística através do teste não-paramétrico de Wilcoxon e foi constatado que, usando o KF, os métodos apresentam desempenhos distintos. Observa-se que para a Acurácia o RF obteve melhor desempenho, por esse motivo foi escolhido como base para o teste de Wilcoxon. Houve uma diferença de 9.2% para o DT, de 1.2% para o KNN, de 2.4% para o LinSVM, 3.6% para a MLP0 e 0.9% para o RBFSVM. Estas diferenças são consideradas estatisticamente significantes pelo teste de Wilcoxon. Resultados similares foram obtidos para o F1, RECALL e MSE. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado do RF foi considerado Quase Perfeito, para os demais métodos foi considerado Substancial.

Os resultados para a validação cruzada SKF estão dispostos na Tabela 4.15. Nota-se que o RF foi o método com maior acurácia, F1 e RECALL. As diferenças na acurácia foram de 9.6% para o DT, 1.4% para o KNN, 3.2% para o LinSVM, 4.1% para o MLP0 e 1.7% para o RBFSVM, com o mesmo comportamento observado no RECALL e F1. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado do DT foi considerado Substancial, para os demais métodos foi considerado Quase Perfeito.

O DT não obteve bons resultados comparado com os outros métodos, as árvores de decisão criam árvores tendenciosas se algumas classes possuem mais amostras que outras. O MLP0 é uma rede neural de múltiplas camadas, o faz com que a quantidade de conexões na arquitetura e na rede sejam grandes o que dificulta a convergência para a solução no processo de minimização do funcional do erro médio quadrático. O bom desempenho do KNN se deve ao fato que os valores de cada atributo foram normalizados, com intuito de que todos caiam num mesmo intervalo de variação, não havendo muita discrepância entre os valores dos diferentes atributos, que poderia influir tendenciosamente no cálculo da distância. Os SVMs obtiveram apreciável acurácia, por serem classificadores eficientes, na maioria dos problemas testados na literatura os métodos apresentam bons resultados, e por utilizarem técnica de relaxamento minimizando o risco de overfitting,

Tabela 4.15: Média e Desvio Padrão da Acurácia, F1, RECALL, Kappa e Erro Quadrático Médio para os métodos KF e SKF – Base de Dados Paraná+. A coluna TEMPO indica o tempo médio de processamento em segundos do classificador para uma execução. 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.

MODELO	CV	ACURÁCIA	F1	RECALL	KAPPA	MSE	TEMPO
DT	KF	0.791 (0.024)	0.779 (0.024)	0.791 (0.024)	0.660 (0.037)	3.429 (0.591)	0.002
KNN	KF	0.871 (0.014)	0.845 (0.016)	0.871 (0.014)	0.786 (0.024)	2.783 (0.375)	0.004
LinSVM	KF	0.859 (0.016)	0.846 (0.013)	0.859 (0.016)	0.772 (0.024)	2.429 (0.372)	0.010
MLP0	KF	0.847 (0.025)	0.836 (0.025)	0.847 (0.025)	0.756 (0.041)	2.585 (0.675)	0.962
RBFSVM	KF	0.874 (0.014)	**0.859 (0.019)***	0.874 (0.014)	0.794 (0.025)	2.012 (0.407)	0.012
RF	KF	**0.883 (0.012)**	**0.859 (0.013)**	**0.883 (0.012)**	**0.806 (0.020)**	1.742 (0.304)	3.248
DT	SKF	0.798 (0.017)	0.782 (0.021)	0.798 (0.017)	0.667 (0.028)	3.504 (0.656)	0.002
KNN	SKF	0.880 (0.008)	0.855 (0.009)	0.880 (0.008)	0.802 (0.014)	2.702 (0.313)	0.004
LinSVM	SKF	0.862 (0.015)	0.850 (0.013)	0.862 (0.015)	0.778 (0.023)	2.368 (0.304)	0.010
MLP0	SKF	0.853 (0.024)	0.847 (0.023)	0.853 (0.024)	0.767 (0.037)	2.346 (0.621)	0.873
RBFSVM	SKF	0.877 (0.011)	0.863 (0.016)	0.877 (0.011)	0.799 (0.020)	2.074 (0.315)	0.012
RF	SKF	**0.894 (0.010)**	**0.872 (0.009)**	**0.894 (0.010)**	**0.825 (0.016)**	1.437 (0.259)	3.250

que é um problema que aparece em dados com grande dimensionalidade esparsos, como é o caso da base de dados utilizada. A diferença é o que o RBFSVM permite resolver problemas, originalmente, não linearmente separáveis, através do mapeamento para um espaço de maior dimensão. A Floresta Aleatória geralmente obtém bons resultados, uma vez que gera uma gama de árvores de decisão e opta pela classe mais votada. Um árvore de decisão sozinha não obteve bons resultados, mas a floresta aleatória pelo fato de selecionar atributos para a construção de árvores, conseguiu criar árvores menos tendenciosas ocasionando no bom desempenho.

A Figura 4.15 apresenta os resultados dos dados sem seletores de características e sem o balanceamento, para os classificadores DT, KNN, LinSVM, RBFSVM, MLP0 e RF. Nota-se que o uso da validação cruzada SKF resultou em melhores acurácia, F1 e RECALL para todos os métodos de classificação estudados. Pode-se observar na Tabela 4.15 que com a aplicação do SKF o coeficiente Kappa teve nível de concordância Quase Perfeito para todos os métodos com exceção do DT. O SKF realiza uma melhor distribuição entre as classes das amostras entre os subconjuntos de treinamento e de teste, possibilitando que estes conjuntos tenham a mesma fração de amostras de cada classe (petrofácies).

4.3.3.2 Dados Desbalanceados – Seleção de Características

Para esta base de dados em específico criou-se uma assembleia de constituintes, uma divisão dos constituintes em Detríticos e Diagenéticos. Detríticos são os grãos depositados, provenientes da desagregação da litologias da área fonte e/ou de altos internos da bacia. Diagenéticos são os constituintes precipitados quimicamente depois da deposição dos sedimentos, durante o processo diagenético. Assim, esses constituintes visam representar nas amostras os processos de formação do arcabouço e de consolidação. A Tabela 4.16 mostra a descrição dos componentes Detríticos e Diagenéticos. Efetuou-

Tabela 4.16: Descrição dos componentes Detríticos e Diagenéticos usados na estratégia de assembleia de constituintes.

Descrição do tipo de componente	Constituintes
Componentes Detríticos (arcabouço)	quartzo, feldspato e bioclasto
Componentes Diagenéticos (consolidação)	glauconita, crescimento secundário de quartzo, caulinita, pirita, siderita, cimento carbonático, porosidade intergranular e porosidade intragranular

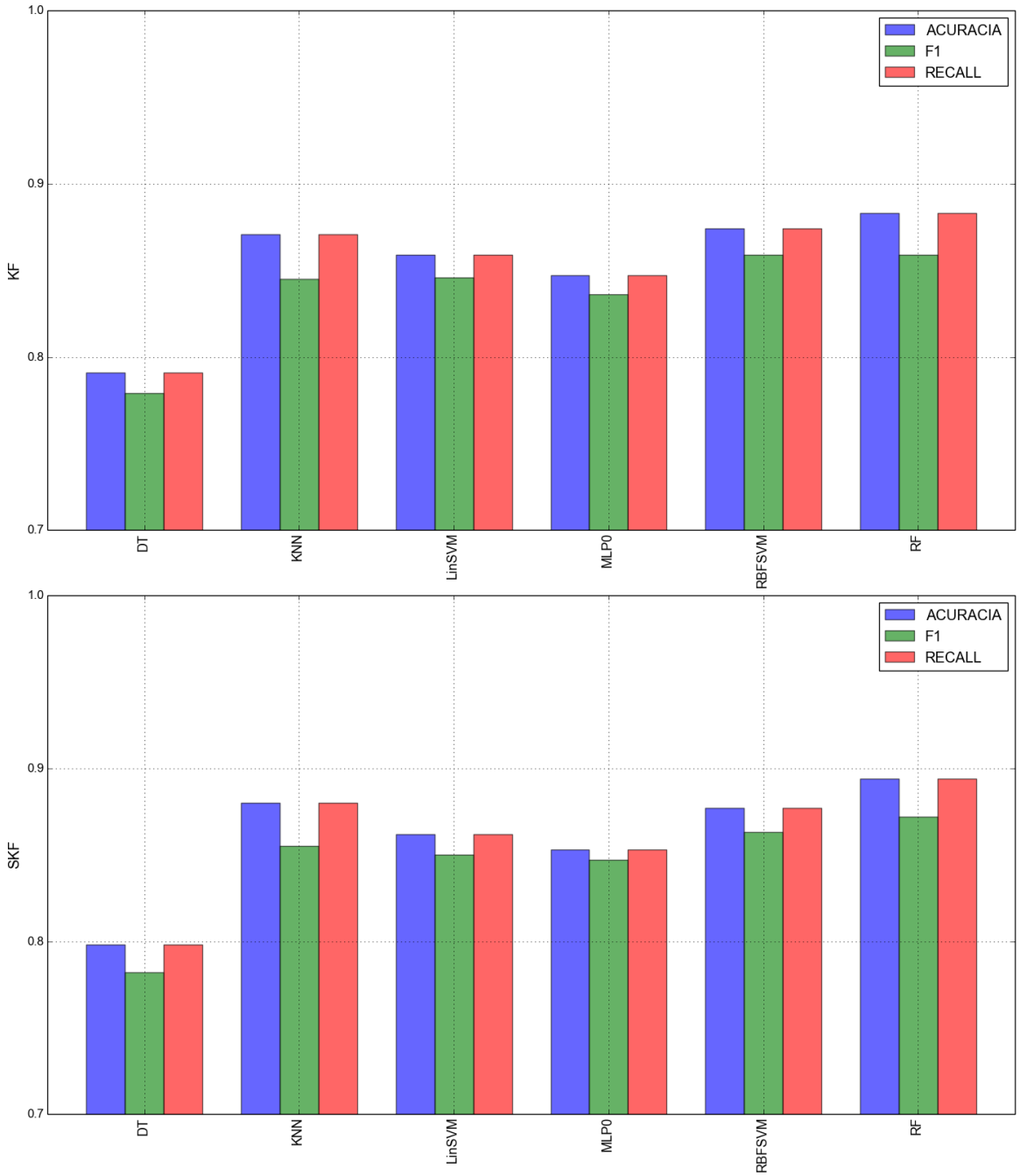


Figura 4.15: KF e SKF – Paraná+ (30 iterações, K= 5) - Acurácia, F1 e RECALL

se a normalização desses dados separadamente, reduzindo a dimensionalidade para duas dimensões, utilizando Análise de Componentes Principais (PCA) (Friedman *et al.*, 2001). Em seguida agrupou-se os mesmos, formando uma base de dados. O intuito dessa abordagem é verificar se essa separação dos constituintes influencia na classificação dos dados.

A Figura 4.16 apresenta o gráfico de barras com a acurácia encontrada através da emprego dos métodos de seleção de características (SKB, SP, SFpr e Fwe), da assembleia

de constituintes (ASBLY) e sem o uso de nenhuma abordagem (NFS) além dos métodos de classificação combinados com o SKF. Para os classificadores MLP0 e KNN o uso do ASBLY resultou em uma acurácia mais próxima de 1.0. Nenhum dos outros seletores de características obtiveram bom desempenho para estes dois classificadores, o segundo melhor resultado foi para o NFS. Para o LinSVM, o método SKB gerou uma acurácia mais alta. No DT, o melhor resultado foi com o Fwe e o SP, que obtiveram o mesmo valor da acurácia. Para o RF, o NFS obteve melhor desempenho seguido do ASBLY.

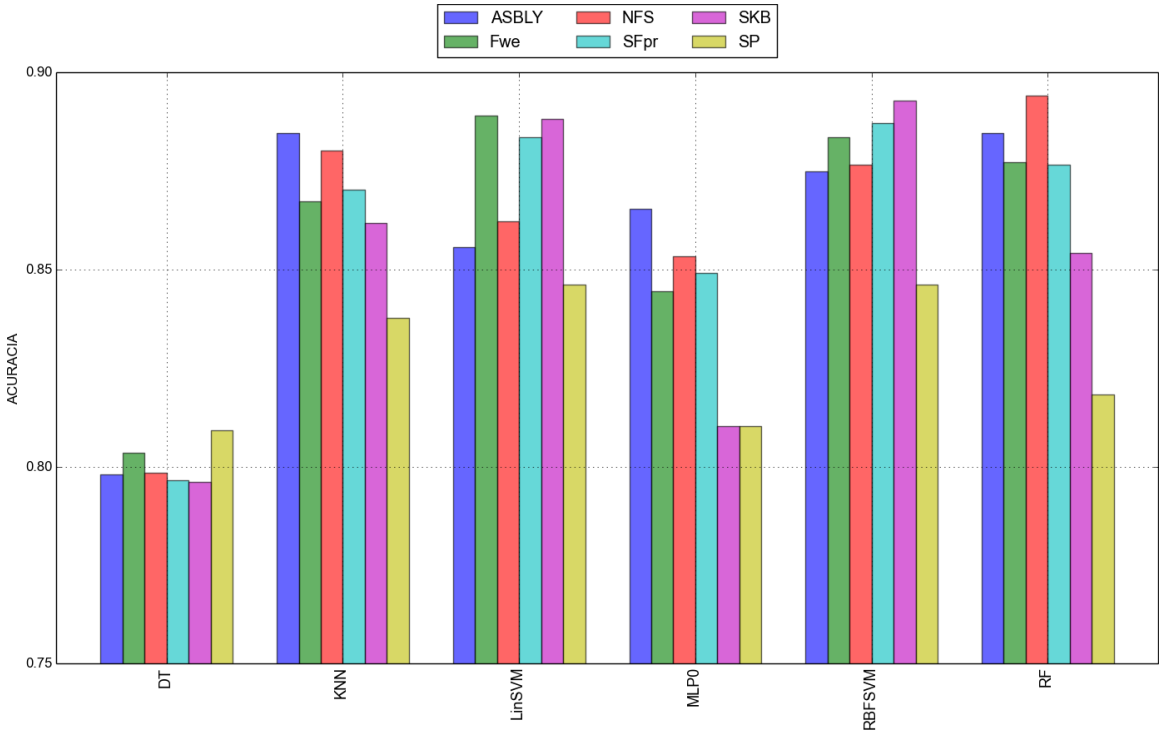


Figura 4.16: Seleção de Características – Paraná+ (30 iterações, K= 5) - Acurácia

O bom desempenho do método MLP0 utilizando a assembleia de constituintes pode ser explicada pelo fato da complexidade da rede ter sido reduzida, uma vez que foram utilizadas quatro características, sendo duas detríticas e duas diagenéticas. O MLP0 é uma rede neural de múltiplas camadas, e a redução de dimensionalidade proporciona a redução da quantidade de conexões na arquitetura e na rede e consequentemente facilita a convergência para a solução no processo de minimização do funcional do erro médio quadrático. Neste caso, a perda de informações devido a redução de dimensões do problema é compensada com a simplificação do problema de otimização envolvido no treinamento da rede neural. O seletor nesse caso melhorou o desempenho do classificador em 1.2% em relação ao NFS.

No caso do LinSVM o SKB apresentou melhor acurácia. Foram selecionadas as 5

características com maior pontuação. O seletor nesse caso melhorou o desempenho do classificador em 0.1% em relação ao SFpr e em 2% em relação ao NFS, essas abordagens tiveram desempenho semelhantes. Nota-se que a assembleia de constituintes não se destacou, pois nesse classificador os dados sofrem uma transformação não linear, dessa maneira, houve duas transformações, uma realizada pela assembleia de constituintes e em seguida outra realizada pela formulação matemática do classificador LinSVM, o que ocasionou uma perda de informação e resultou no baixo desempenho do método.

Para o KNN com 4 vizinhos utilizando a assembleia de constituintes obteve melhor resultado. A abordagem nesse caso melhorou o desempenho do classificador em 0.8% em relação ao NFS, segunda abordagem com melhor desempenho. Um dos fatores que influenciou os bons resultados da abordagem utilizada foi a normalização dos dados que não ocorre nos outros métodos (SKB, SP, SFpr e Fwe). Por outro lado, O DT não obteve bons resultados comparado com os outros métodos independente da técnica de seleção de características empregada. As árvores de decisão criam árvores tendenciosas se algumas classes possuem mais amostras que outras. A base de dados em questão está desbalanceada, e o uso do SKF não foi suficiente para que o classificador gerasse bons resultados.

No caso do RF, por ele já selecionar características para a construção das árvores de decisão, o melhor desempenho foi obtido pelo NFS, sem nenhum seletor de característica ser aplicado. A diferença para o ASBLY foi de 0.9%, uma diferença considerada pequena, o que mostra que a assembleia de constituintes foi uma técnica que levou a bons resultados.

4.3.3.3 Dados Balanceados

A Tabela 4.17 mostra a média e o desvio padrão da acurácia, F1, Kappa, Erro Quadrático Médio (MSE) e a média do tempo de processamento para validação cruzada SKF para os dados balanceados. Efetuou-se uma análise estatística através do teste não-paramétrico de Wilcoxon e constatou-se que os métodos KNN e RBFSVM apresentam desempenhos significamente semelhantes ao RF.

Observa-se que para a Acurácia o RF obteve melhor desempenho, por esse motivo foi escolhido como base para o teste de Wilcoxon. Houve uma diferença de 1.8% para o DT, de 0.3% para o KNN, de 1.0% para o LinSVM, 0.9% para a MLP0 e 0% para o RBFSVM. Estas diferenças, com exceção do KNN e do RBFSVM, são consideradas estatisticamente

significantes pelo teste de Wilcoxon. Resultados semelhantes foram obtidos para o F1, RECALL e MSE. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado dos métodos foi considerado Quase Perfeito. A Figura 4.17 ilustra os resultados dos dados com o balanceamento, para os classificadores DT, KNN, LinSVM, RBFSVM, MLP0 e RF.

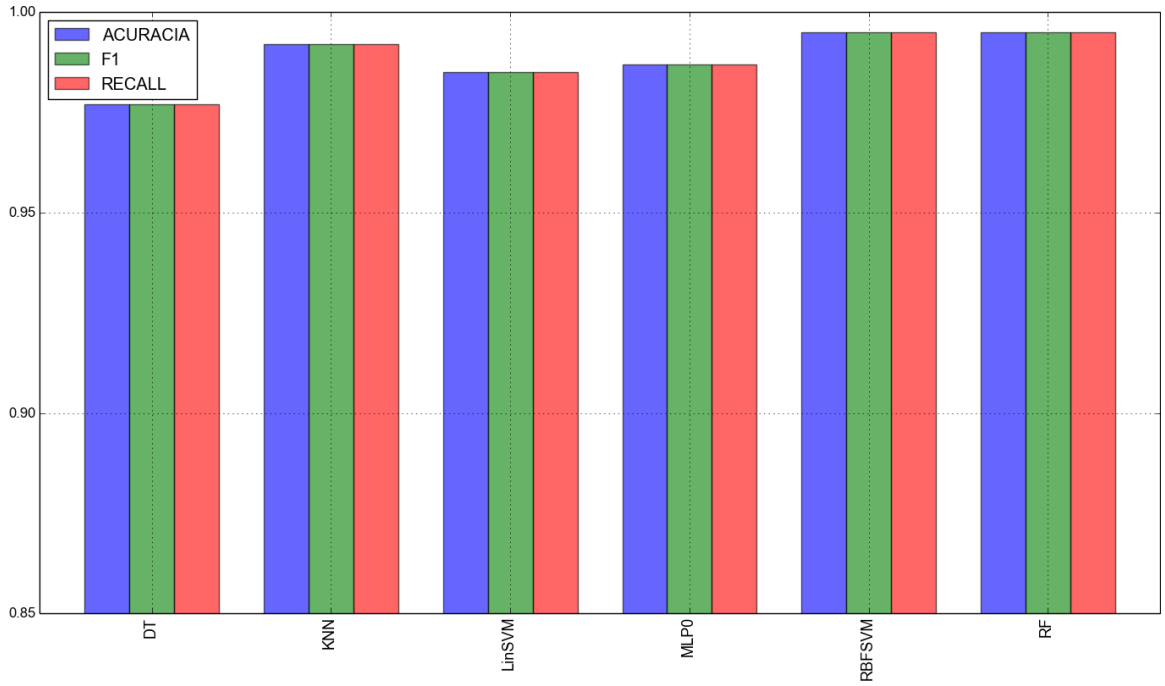


Figura 4.17: SKF – Paraná+ (30 iterações, K= 5) - Acurácia, F1 e RECALL

Depois de realizado o balanceamento dos dados e após a análise dos resultados percebe-se que o método que ocasionou melhor acurácia, F1 e RECALL foi o RF juntamente com o RBFSVM, seguido do KNN. Estes três métodos se mostraram úteis para a classificação desta base de dados, tanto desbalanceados como balanceados. Os classificadores MLP0 , LinSVM e DT apresentaram bons resultados e mostraram ser uma alternativa para bases de dados balanceadas. No entanto, nota-se com base nos resultados apresentados que o balanceamento dos dados foi um fator que proporcionou ganho de desempenho em todos os métodos.

4.3.4 Mucuri

4.3.4.1 Dados Desbalanceados

A Tabela 4.18 apresenta a média e o desvio padrão da acurácia, F1, Kappa, Erro Quadrático Médio (MSE) e a média do tempo de processamento para validação cruzada

Tabela 4.17: Média e Desvio Padrão da Acurácia, F1, RECALL, Kappa e Erro Quadrático Médio para os métodos KF e SKF - Dados Balanceados Paraná+. A coluna TEMPO indica o tempo médio de processamento em segundos do classificador para uma execução. 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.

MODELO	CV	ACURÁCIA	F1	RECALL	KAPPA	MSE	TEMPO
DT	SKF	0.977 (0.002)	0.977 (0.002)	0.977 (0.002)	0.988 (0.008)	0.026 (0.006)	0.004
KNN	SKF	0.992 (0.001)*	0.992 (0.001)*	0.992 (0.001)*	0.991 (0.001)	0.014 (0.014)	0.007
LinSVM	SKF	0.985 (0.003)	0.985 (0.003)	0.985 (0.003)	0.990 (0.002)	0.011 (0.006)	0.045
MLP0	SKF	0.986 (0.004)	0.986 (0.004)	0.986 (0.004)	0.985 (0.033)	0.010 (0.005)	2.975
RBFSVM	SKF	**0.995 (0.003)***	**0.995 (0.003)***	**0.995 (0.003)***	**1.000 (0)**	0.005 (0.002)	0.055
RF	SKF	**0.995 (0.003)**	**0.995 (0.003)**	**0.995 (0.003)**	**1.000 (0)**	0.005 (0.002)	1.043

KF. Realizou-se uma análise estatística através do teste não-paramétrico de Wilcoxon e foi constatado que, usando o KF, os métodos apresentam desempenhos distintos. Observa-se que para a Acurácia o RF obteve melhor desempenho, por esse motivo foi escolhido como base para o teste de Wilcoxon. Houve uma diferença de 7.3% para o DT, de 9.5% para o KNN, de 3.4% para o LinSVM, 11.2% para a MLP0 e 2.8% para o RBFSVM. Estas diferenças são consideradas estatisticamente significantes pelo teste de Wilcoxon. Resultados similares foram obtidos para o F1, RECALL e MSE. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado dos métodos foi considerado Pobre.

Os resultados para a validação cruzada SKF também são mostrados na Tabela 4.18. Nota-se que o RF foi o método com maior acurácia, F1 e RECALL. As diferenças na acurácia foram de 7.8% para o DT, 9.5% para o KNN, 3.6% para o LinSVM, 11.1% para o MLP0 e 2.7% para o RBFSVM, com o mesmo comportamento observado no RECALL e F1. Em relação ao Kappa, comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado dos métodos foi considerado Pobre.

Nenhum dos métodos empregados obteve um resultado satisfatório. O DT não obteve bons resultados comparado com os outros métodos, as árvores de decisão criam árvores tendenciosas se algumas classes possuem mais amostras que outras. O MLP0 é uma rede neural de múltiplas camadas, o faz com que a quantidade de conexões na arquitetura e na rede sejam grandes o que dificulta a convergência para a solução no processo de minimização do funcional do erro médio quadrático. A normalização dos dados não foi suficiente para que o KNN obtivesse bom desempenho. Os SVMs apesar de na maioria dos problemas testados na literatura os métodos apresentam bons resultados, nesse caso o desempenho não foi apreciável. As árvores de decisão criadas pelo RF não conseguiram ser suficientes para que o desempenho não fosse comprometido. A seleção de característica realizada pelo método não ocasionou em grandes acertos.

A Figura 4.18 apresenta os resultados dos dados sem seletores de características e sem o balanceamento, para os classificadores DT, KNN, LinSVM, RBFSVM, MLP0 e RF. Observa-se que o uso da validação cruzada SKF resultou em melhores acurácia, F1 e RECALL para todos os métodos de classificação estudados. O SKF realiza uma melhor distribuição entre as classes das amostras entre os subconjuntos de treinamento e de teste, possibilitando que estes conjuntos tenham a mesma fração de amostras de cada

Tabela 4.18: Média e Desvio Padrão da Acurácia, F1, RECALL, Kappa e Erro Quadrático Médio para os métodos KF e SKF – Base de Dados Mucuri. A coluna TEMPO indica o tempo médio de processamento em segundos do classificador para uma execução. 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.

MODELO	CV	ACURÁCIA	F1	RECALL	KAPPA	MSE	TEMPO
DT	KF	0.403 (0.026)	0.368 (0.035)	0.403 (0.026)	0.101 (0.049)	2.132 (0.232)	0.002
KNN	KF	0.381 (0.022)	0.378 (0.021)	0.381 (0.022)	0.140 (0.030)	2.087 (0.151)	0.003
LinSVM	KF	0.442 (0.009)	0.310 (0.063)	0.442 (0.009)	0.048 (0.068)	2.276 (0.134)	0.014
MLP0	KF	0.364 (0.067)	0.358 (0.079)	0.364 (0.067)	0.105 (0.070)	2.225 (0.384)	1.839
RBFSVM	KF	0.448 (0.012)	0.343 (0.074)	0.448 (0.012)	0.087 (0.084)	2.125 (0.243)	0.021
RF	KF	**0.476 (0.021)**	**0.407 (0.034)**	**0.476 (0.021)**	**0.157 (0.044)**	2.004 (0.131)	3.623
DT	SKF	0.410 (0.019)	0.364 (0.029)	0.410 (0.019)	0.102 (0.039)	2.022 (0.225)	0.002
KNN	SKF	0.393 (0.020)	0.390 (0.021)	0.393 (0.020)	0.156 (0.027)	2.053 (0.169)	0.003
LinSVM	SKF	0.452 (0.015)	0.361 (0.065)	0.452 (0.015)	0.103 (0.075)	2.182 (0.149)	0.017
MLP0	SKF	0.377 (0.056)	0.373 (0.064)	0.377 (0.056)	0.119 (0.062)	2.216 (0.387)	1.983
RBFSVM	SKF	0.461 (0.017)	0.390 (0.055)	0.461 (0.017)	0.145 (0.067)	1.958 (0.185)	0.021
RF	SKF	**0.488 (0.021)**	**0.423 (0.032)**	**0.488 (0.021)**	**0.180 (0.043)**	1.971 (0.134)	3.653

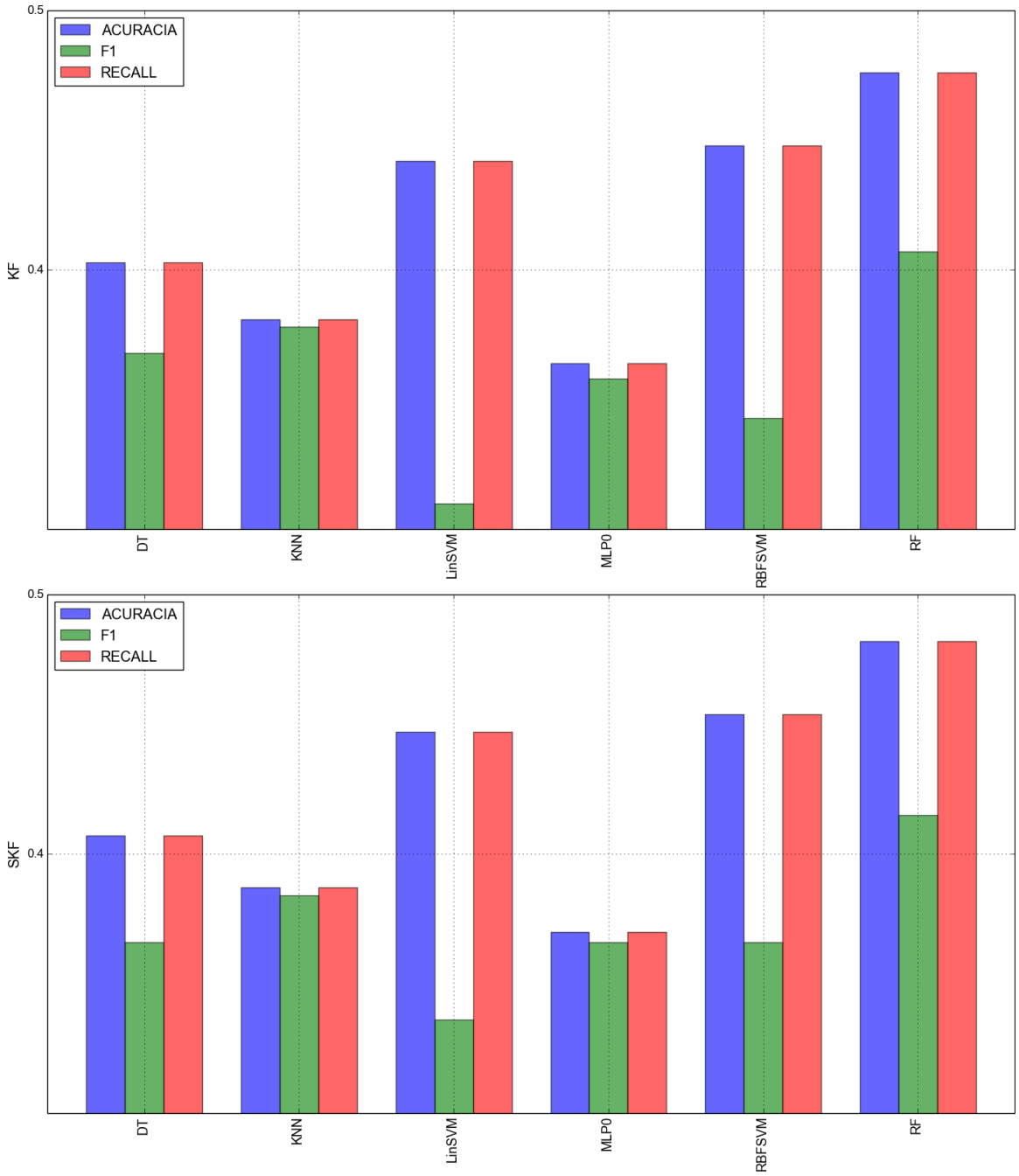


Figura 4.18: KF e SKF – Mucuri (30 iterações, K= 5) - Acurácia, F1 e RECALL

classe (permeabilidade).

4.3.4.2 Dados Desbalanceados – Seleção de Características

A Figura 4.19 apresenta o gráfico de barras com a acurácia encontrada através da emprego dos métodos de seleção de características (SKB, SP, SFpr e Fwe) e sem o uso de nenhuma abordagem (NFS) além dos métodos de classificação combinados com o SKF. Para os classificadores KNN, LinSVM e o RBFSVM o uso do SKB resultou em uma

acurácia mais alta, foram selecionadas as 5 características com maior pontuação. Para o MLP0, o método SP gerou uma acurácia mais próxima de 1.0, selecionar 10% das características com maiores pontuações foi a melhor opção neste caso. No DT e no RF, o melhor resultado foi com o SFpr e com o SP que obtiveram mesmo valor de acurácia. Selecionar 10% das características com maiores pontuações e realizar a seleção através de falsas detecções foram as melhores escolhas em comparação com outros seletores.

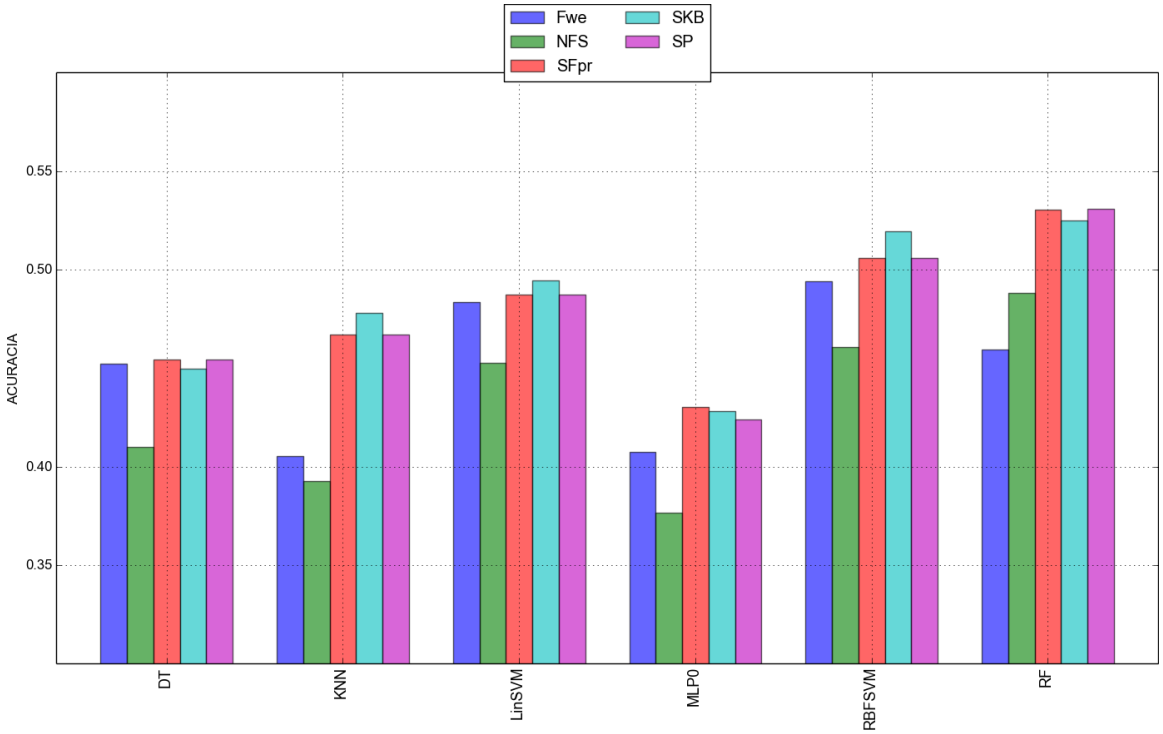


Figura 4.19: Seleção de Características – Mucuri (30 iterações, K= 5) - Acurácia

4.3.4.3 Dados Balanceados

A Tabela 4.19 mostra a média e o desvio padrão da acurácia, F1, Kappa, Erro Quadrático Médio (MSE) e a média do tempo de processamento para validação cruzada SKF para os dados balanceados. Efetuou-se uma análise estatística através do teste não-paramétrico de Wilcoxon e constatou-se que os métodos KNN e RBFSVM apresentam desempenhos significamente semelhantes ao RF.

Observa-se que para a Acurácia o RF obteve melhor desempenho, por esse motivo foi escolhido como base para o teste de Wilcoxon. Houve uma diferença de 3.4% para o DT, de 6.5% para o KNN, de 3.5% para o LinSVM, 4.3% para a MLP0 e 0.9% para o RBFSVM. Estas diferenças são consideradas estatisticamente significantes pelo teste de Wilcoxon. Resultados semelhantes foram obtidos para o F1, RECALL e MSE. Em relação ao Kappa,

comparando com a classificação manual, pode-se afirmar que o nível de concordância obtido pelo resultado dos métodos DT, KNN, LinSVM e MLP0 foi considerado Moderado, para os demais métodos foi considerado Substancial. A Figura 4.20 ilustra os resultados dos dados com o balanceamento, para os classificadores DT, KNN, LinSVM, RBFSVM, MLP0 e RF.

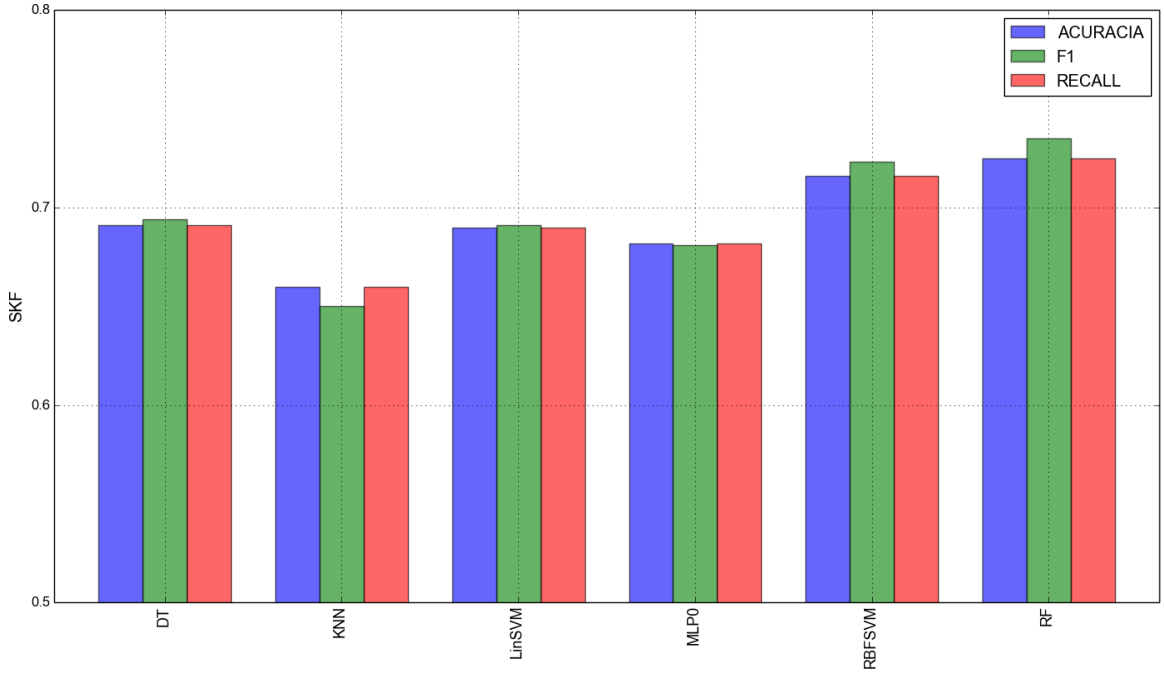


Figura 4.20: SKF – Mucuri (30 iterações, K= 5) - Acurácia, F1 e RECALL

Após a realização do balanceamento dos dados e depois da análise dos resultados nota-se que o método que ocasionou melhor acurácia, F1 e RECALL foi o RF, seguido do RBFSVM. Estes métodos conseguiram em torno de 70% de acurácia, resultado geralmente obtido com dados desbalanceados, comparando com as outras bases de dados testadas. Isso mostra que classificar a permeabilidade é uma tarefa complexa. Entretanto, observa-se com base nos resultados apresentados que o balanceamento dos dados foi um fator que proporcionou ganho de desempenho em todos os métodos.

Tabela 4.19: Média e Desvio Padrão da Acurácia, F1, RECALL, Kappa e Erro Quadrático Médio para o método SKF- Dados Balanceados Mucuri. A coluna TEMPO indica o tempo médio de processamento em segundos do classificador para uma execução. 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.

MODELO	CV	ACURÁCIA	F1	RECALL	KAPPA	MSE	TEMPO
DT	SKF	0.691 (0.015)	0.694 (0.015)	0.691 (0.015)	0.588 (0.020)	0.971 (0.096)	0.005
KNN	SKF	0.660 (0.012)	0.650 (0.013)	0.660 (0.012)	0.547 (0.016)	1.112 (0.042)	0.005
LinSVM	SKF	0.690 (0.015)	0.691 (0.014)	0.690 (0.015)	0.587 (0.020)	1.022 (0.054)	0.038
MLP0	SKF	0.682 (0.018)	0.681 (0.018)	0.682 (0.018)	0.576 (0.024)	1.063 (0.065)	5.553
RBFSVM	SKF	0.716 (0.007)	0.723 (0.008)	0.716 (0.007)	0.621 (0.010)	0.975 (0.052)	0.083
RF	SKF	**0.725 (0.012)**	**0.735 (0.012)**	**0.725 (0.012)**	**0.633 (0.017)**	1.008 (0.035)	5.147

4.3.5 Agrupamento - Tibagi

Com intuito de introduzir trabalhos futuros foi realizado um teste para técnica não-supervisionada. Foi utilizada somente a base de dados Tibagi para esse procedimento. Nese caso, a base de dados foi separada e cada poço foi analisado separadamente. São 5 poços: PPG1, PPG2, PPG3, PPG4 e PPG5. Para o Poço 1 (PPG1) foram averiguadas 12 lâminas classificadas de PPG1-1 até PPG1-12. Segundo a análise manual, a lâmina PPG1-1 foi identificada como pertencente à petrofácies denominada I-2; as amostras PPG1-2 a PPG1-8 à petrofácies I-1; as demais foram agrupadas à petrofácies chamada PT-1. No Poço 2 (PPG2) analisou-se 11 lâminas, PPG2-1 a PPG2-11, onde a lâmina PPG2-1 foi classificada na petrofácies PT-2 e o restante na petrofácies PT-1. Para Poço 3 (PPG3) análise foi realizada em 12 lâminas, PPG3-1 a PPG3-12. A lâmina PPG3-11 foi classificada como pertencente à PT-3 e as 11 amostras restantes foram agrupadas à PT-1. O Poço 4 (PPG4) teve 5 lâminas analisadas, PPG4-1 a PPG4-5, onde todas foram classificadas como referente à PT-4. Foram examinadas 4 lâminas do Poço 5 (PPG5), PPG5-1 a PPG5-4, no qual a lâmina PPG5-3 foi identificada como relativa à PT-3 e o restante à PT-1. As Figuras 4.21, 4.22, 4.23, 4.24 e 4.25 exibem um resultado preliminar através da aplicação do K-Means.

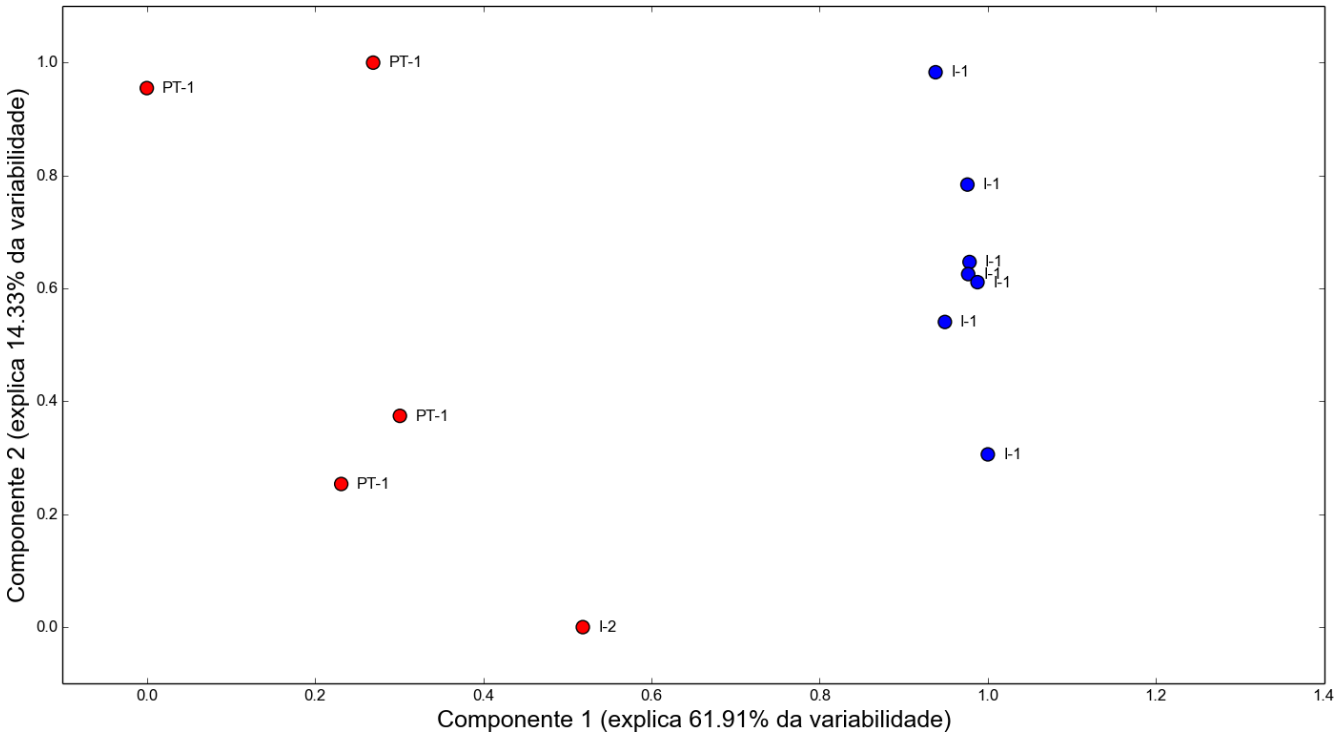


Figura 4.21: Visualização dos agrupamentos obtidos com amostras para o primeiro furo de sondagem do membro Tibagi. PPG1–$SC = 0.456$

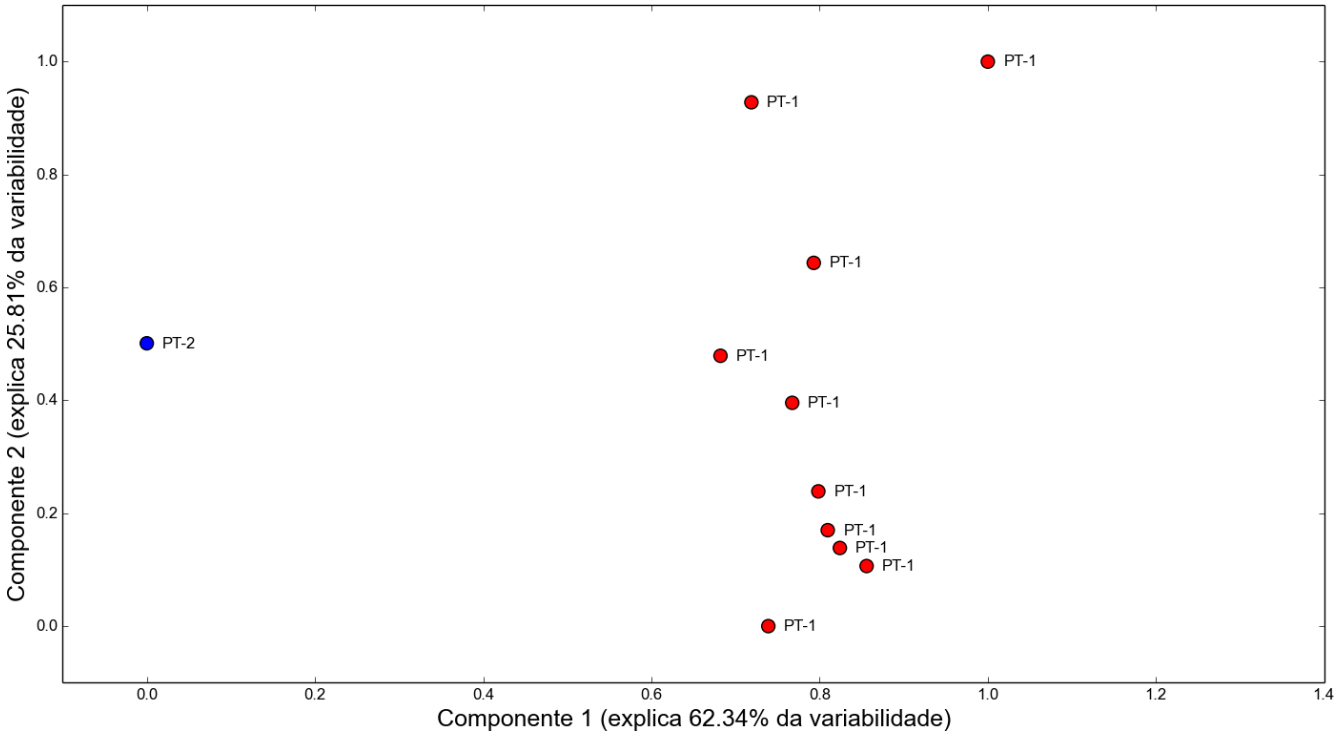


Figura 4.22: Visualização dos agrupamentos obtidos com amostras para o segundo furo de sondagem do membro Tibagi. PPG2–$SC = 0.586$

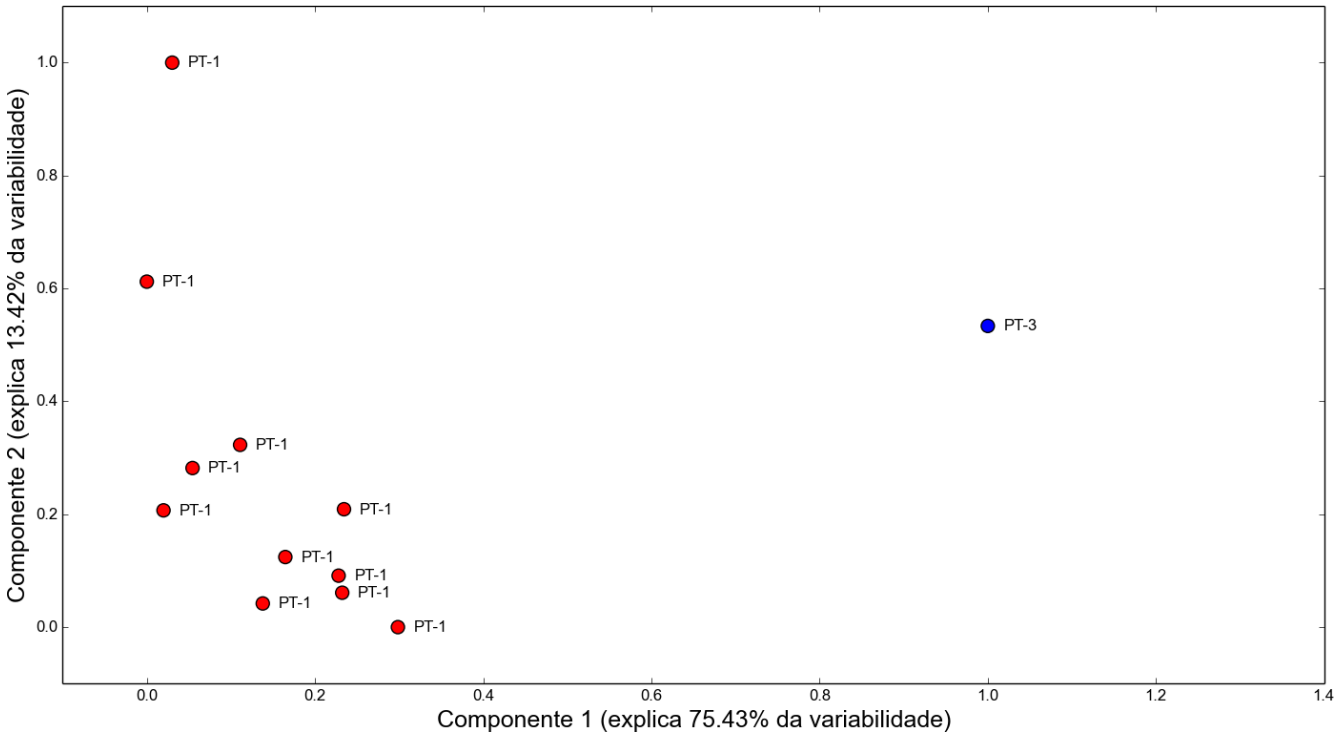


Figura 4.23: Visualização dos agrupamentos obtidos com amostras para o terceiro furo de sondagem do membro Tibagi. PPG3–$SC = 0.657$

O método de agrupamento utilizado é o K-Means que foi avaliado através do coeficiente de silhueta (SC), descrito na Seção 3.6.7.1. No PPG1 o K-Means identificou dois

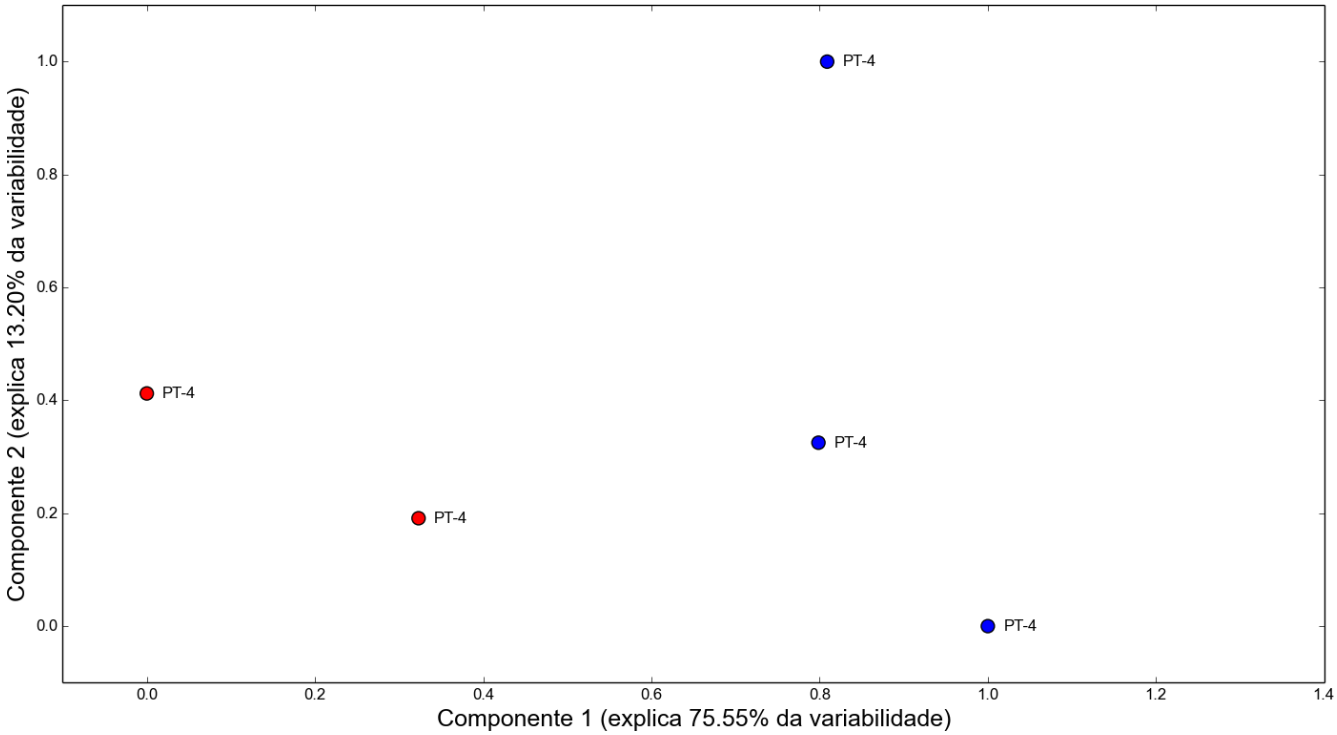


Figura 4.24: Visualização dos agrupamentos obtidos com amostras para o quarto furo de sondagem do membro Tibagi. PPG4–$SC = 0.426$

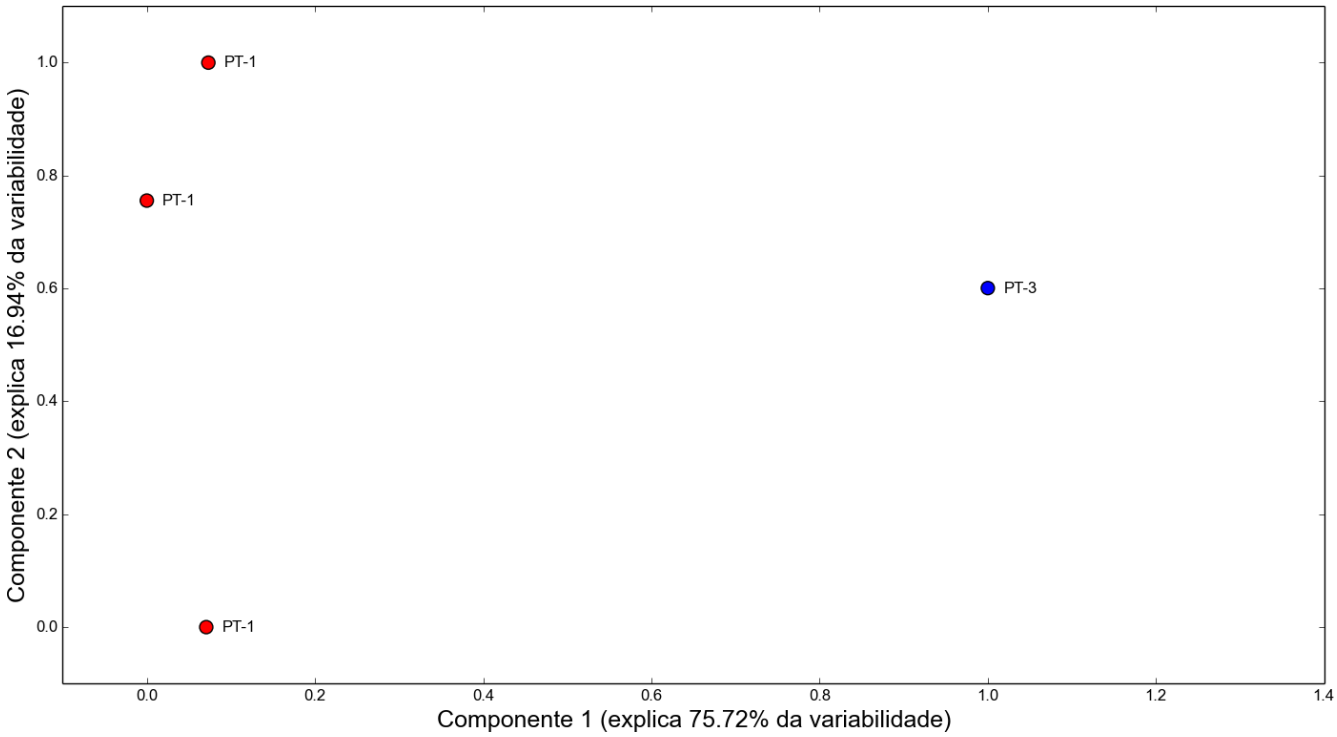


Figura 4.25: Visualização dos agrupamentos obtidos com amostras para os três furos de sondagens restantes do membro Tibagi. PPG5–$SC = 0.396$

agrupamentos, coincidindo com o método manual na classificação da petrofácies I-1. Em relação a PT-1 e I-2, o algoritmo sugeriu que amostras pertencentes a essas petrofácies

são semelhantes e agrupou 92% do total de lâminas de forma correta. Para o PPG2 o método encontrou dois agrupamentos e o resultado coincidiu com a classificação manual ao identificar corretamente as petrofácies associadas. Em relação ao PPG3, o K-Means obteve dois agrupamentos e o resultado coincidiu com a classificação manual. No PPG4 conseguiu agrupar as amostras em 2 agrupamentos; porém pelo método convencional foi identificada uma única petrofácies, PT-4. O método não classificou as amostras PPG4-1 e PPG4-5 como pertencentes a PT-4, pois não encontrou semelhanças entre essas duas amostras em relação as outras três. Isso mostra a necessidade de integrar dados provenientes de imagens petrográficas para ter mais informações a respeito das lâminas. O método conseguiu distribuir 60% das lâminas de forma correta. Para o PPG5, o K-Means obteve resultado que coincidiu com a classificação manual. O método mostrou-se útil para individualizar petrofácies, obtendo grandes índices de acerto em comparação com o resultado obtido através da classificação (divisão) manual.

5 CONCLUSÕES

5.1 Conclusões

5.1.1 Do desenvolvimento da dissertação

A metodologia utilizada para classificação de petrofácies sedimentares e de permeabilidade baseou-se na aplicação de métodos de classificação, técnicas de validação cruzada, seletores de características. Três testes foram realizados: comparação de classificadores com duas técnicas de validação cruzada, comparação de seletores de características com a técnica SKF de validação cruzada e por último realizou-se o balanceamento dos dados com intuito de verificar se o desbalanceamento dos dados influenciou no desempenho da metodologia. A metodologia foi aplicada em quatro bases de dados. Além disso, foi proposta uma técnica de seleção de características baseada na ideia de assembleia de constituintes, dividindo-os em detríticos e diagenéticos, com o objetivo de melhor classificar as amostras nas bases de dados.

Para a base de dados Paleosul, os classificadores apresentaram um ótimo desempenho. O KNN, o LinSVM , o RBFSVM e o RF obtiveram resultados semelhantes e o nível de concordância da classificação foi considerado Quase Perfeito pelo coeficiente Kappa. A acurácia desses quatro métodos foi de 98.3% utilizando o SKF como técnica de validação cruzada. Por meio do uso da seleção de características foi constatado que a acurácia dos métodos não sofreu alteração significativa. De modo geral, o seletor que obteve melhor resultado foi o SKB e NFS no caso do RF. Realizando o balanceamento dos dados pode-se observar que o desbalanceamento influenciou no desempenho dos métodos. Aplicando o balanceamento chegou-se a 100% de acerto na classificação.

Na base de dados Tibagi, os classificadores conseguiram ter bom desempenho. O LinSVM e o RBFSVM obtiveram resultados significamente semelhantes e o nível de concordância da classificação foi considerado Substancial pelo coeficiente Kappa. A acurácia desses dois métodos foi de 77.3% utilizando o SKF como técnica de validação cruzada. Utilizando seleção de características foi constatado que a acurácia dos métodos aumentou em 1%. Os seletores que obtiveram acurácia mais alta foram o SKB e o

Fwe. Realizando o balanceamento dos dados pode-se observar que o desbalanceamento influenciou no desempenho dos métodos. Aplicando o balanceamento chegou-se a 97% de acerto na classificação.

Para a base de dados Paraná+, os classificadores conseguiram ter bom desempenho. O RF teve o nível de concordância da classificação considerado Quase Perfeito pelo coeficiente Kappa. Nenhum método teve o comportamento considerado estatisticamente semelhante em relação ao RF. A acurácia desse métodos foi de 89.4% utilizando o SKF como técnica de validação cruzada. Utilizando seleção de características foi constatado que a acurácia dos métodos não sofreu alteração significativa. Os seletores que obtiveram melhor resultado foram o SKB, a assembleia de constituintes e o NFS. Ao realizar-se o balanceamento dos dados pode-se observar que o desbalanceamento influenciou no desempenho dos métodos. Aplicando o balanceamento chegou-se a 99.5% de acerto na classificação.

Na base de dados Mucuri, os classificadores não apresentaram bom desempenho. O RF teve o nível de concordância da classificação considerado Pobre pelo coeficiente Kappa. A acurácia desses três métodos foi de 48.8% utilizando o SKF como técnica de validação cruzada. Utilizando seleção de características foi constatado que a acurácia dos métodos não se alterou. O seletor que obteve acurácia mais alta foi o SKB e o NFS. Realizando o balanceamento dos dados pode-se observar que o desbalanceamento influenciou no desempenho dos métodos. Aplicando o balanceamento chegou-se a 72.5% de acerto na classificação.

Analisado os resultados acima, conclui-se que o melhor classificador para as tarefas propostas foi o RF, sem utilizar seleção de características. Nota-se que o desbalanceamento dos dados influencia na classificação, mas outras técnicas de balanceamento devem ser aplicadas. Em relação ao agrupamento realizado na base de dados Tibagi, o K-Means mostrou-se útil para identificar petrofácies, obtendo grandes índices de acerto em comparação com as petrofácies obtidas a partir da individualização realizada manualmente.

5.1.2 Dos objetivos

Através da metodologia adotada conseguiu-se avaliar o desempenho de técnicas de inteligência computacional para classificação de dados petrográficos a partir da

comparação realizada entre as mesmas. Por meio do procedimento adotado foi permitido o estudo e a implementação de classificadores, seletores de características e métodos de validação cruzada. Foi implementado uma busca exaustiva e dessa maneira o desempenho dos classificadores foi avaliado com os parâmetros ótimos. Métricas para examinar as técnicas de inteligência computacional foram executadas e estudadas. A metodologia foi aplicada em diferentes bases de dados, o que possibilitou a avaliação do seu comportamento em diferentes aplicações importantes para determinar o potencial de um reservatório de petróleo. Desse modo, todos os objetivos específicos foram alcançados.

5.1.3 Da conclusão da dissertação

De forma geral, a metodologia proposta mostrou-se eficiente para a classificação de petrofácies, contribuindo para a rápida determinação dos atributos do reservatório. Para a tarefa de classificar a permeabilidade, a metodologia deve ser melhorada para produzir melhores resultados. Através da classificação pode-se concluir que se há homogeneidade deposicional então tem capacidade de classificar utilizando o método computacional. Se não é homogênea a rocha não foi uniformemente depositaada, como é o caso da base de dados Mucuri, e o método apresenta dificuldade para realizar a classificação. Portanto, pode-se ter uma informação indireta a respeito da qualidade do reservatório. Assim, se há homogeneidade na área.

O agrupamento serve para validar a individualização realizada pelo especialista, indicando possíveis pontos que precisam de uma revisão. Além de ser aplicável em outras bases de dados semelhantes as usadas nessa dissertação, o método computacional resultante pode assistir ao geólogo/petrólogo na tarefa de identificação das petrofácies, reduzindo o tempo de análise em comparação com a classificação manual.

5.2 Trabalhos Futuros

Como trabalhos futuros pode-se empregar comitês de modelos (*ensembles*), da forma

$$F(x) = \sum_{i=1}^{N} \alpha_i M_i(x, \beta_{i1}, \cdots, \beta_{iL}), \qquad \sum_{i=1}^{N} \alpha_i = 1$$

onde $F(\cdot)$ é o classificador resultante, x é a amostra avaliada, N é o número de modelos envolvidos no comitê, $M_i(\cdot)$ é o i-ésimo classificador (por exemplo, Árvores de Decisão ou Máquinas de Vetores Suporte), $\beta_{i1}, \cdots, \beta_{iL}$ é o conjunto de L (hiper)parâmetros associados ao modelo M_i, e α_i são os coeficientes do ensemble associados a cada modelo. Uma outra alternativa é o uso de algoritmos evolucionistas para otimizar (a) os coeficientes do ensemble e também (b) os (hiper)parâmetros dos modelos envolvidos no ensemble, o que pode melhorar a precisão dos modelo resultante e possivelmente o de tempo de processamento quando há uma grande quantidade de (hiper)parâmetros a serem ajustados, pois o grid-search trabalha com uma busca exaustiva.

Um outro aspecto que pode ser aplicado é a integração de dados petrográficos com imagens e dados de perfis de poços (well logs). Através das imagens das lâminas pode-se obter informações a respeito das "texturas" a partir do uso de técnicas de processamento gráfico. A nova base de dados armazenaria informações petrográficas, petrofísicas (retiradas de perfis de poços) e texturais. O uso de técnicas de agrupamento tornam-se interessantes para individualizar as informações petrográficas sem a necessidade de um conjunto de dados previamente classificado. Reduzir a dependência de dados petrográficos para caracterizar reservatórios é de grande importância, visto que a geração desses dados é custosa e demanda muito tempo de análise de um conjunto de especialistas.

Posteriormente pode-se fazer uso de técnicas de regras de associação (Tan *et al.*, 2005) que é uma técnica de mineração aproveitável para a descoberta de relações significativas ocultas em grandes bases de dados, e dessa forma obter alguma informação a respeito da relação entre os constituintes. As relações formadas podem ser exibidas na forma de regras de associação. Além disso pode-se empregar métricas específicas para avaliar as associações previamente formadas. Um algoritmo bastante utilizado é o Apriori (Agrawal *et al.*, 1993) que minera itens contínuos em bases de dados para a identificação de regras de associação entre os itens.

Outra linha de trabalhos futuros envolve a pesquisa em métodos filogenéticos, empregados com sucesso em Biologia e Linguística Computacional (Gray e Atkinson, 2003; Gray *et al.*, 2011; Bouckaert *et al.*, 2012). Esses métodos poder ser úteis para representar a diagênese por meio da similaridade entre os constituintes: uma vez que cada amostra foi retirada de profundidades diferentes, com diferentes processos de deposição e diferentes composições, o emprego de técnicas filogenéticas pode ajudar a construir relações sobre a

história deposicional da rochas sedimentares.

Referências Bibliográficas

Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases. ACM SIGMOD Record 22 (2), 207–216.

Al-Anazi, A., Gates, I., 2010. Support vector regression for porosity prediction in a heterogeneous reservoir: A comparative study. Computers and Geosciences 36 (12), 1494 – 1503.

Asfahani, J., Abdul Ghani, B., Ahmad, Z., 2015. Basalt identification by interpreting nuclear and electrical well logging measurements using fuzzy technique (case study from southern Syria). Applied Radiation and Isotopes 105, 92–97. Disponivel em: `http://linkinghub.elsevier.com/retrieve/pii/S0969804315301342`

Azevedo, F. d. S., 2005. Estudo experimental da influência de tensões na permeabilidade de rochas produtoras de petróleo. Dissertação de mestrado, Programa de Pósgraduação em Engenharia Civil da PUC-Rio, Geotecnia.

Bartlett, A. A., 2000. An analysis of US and world oil production patterns using Hubbert-style curves. Mathematical Geology 32 (1), 1–17.

Berner, R. A., Holdren, G. R., 1977. Mechanism of feldspar weathering: Some observational evidence. Geology 5 (6), 369–372.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., Atkinson, Q. u. D., ago. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337 (6097), 957–960. Disponivel em: `http://science.sciencemag.org/content/337/6097/957`

Brazil, F. A. F., 2004. Estratigrafia de Sequências e Processo Diagenético: Exemplo dos Arenitos Marinho-Rasos da Formação Ponta Grossa, Noroeste da Bacia do Paraná.

Dissertação de mestrado, Dissertação de Mestrado – Programa de pós-graduação em Análise de Bacias e Faixas Móveis. Universidade do Estado do Rio de Janeiro.

Breiman, L., 2001. Random forests. Machine learning, 5–32. Disponivel em: http://link.springer.com/article/10.1023/A:1010933404324

Cevolani, J. T., Oliveira, L. C., Goliatt, L., Pereira, E., Visualização e Classificação Automática de Petrofácies Sedimentares. In: 6^o Congresso Brasileiro de Pesquisa e Desenvolvimento em Petróleo e Gás, 2011.

Chen, J., Li, Z., Bian, B., 2010. Application of Data Mining in Multi-Geological-Factor Analysis. In: Cai, Z., Hu, C., Kang, Z., Liu, Y. (Eds.), Advances in Computation and Intelligence. Vol. 6382 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 402–411. Disponivel em: http://dx.doi.org/10.1007/978-3-642-16493-4_41

Contreras, R. J., 2002. Técnicas de Seleção de Características aplicadas a Modelos Neuro-Fuzzy Hierárquicos BSP. Dissertação de mestrado, Programa de Pós-graduação em Engenharia Elétrica da PUC-Rio.

Crozier, M. J., 1987. Landslides: Causes, Consequences and Environment. Géographie physique et Quaternaire 41 (3), 409–410. Disponivel em: http://id.erudit.org/iderudit/032702ar

da Silva, P. N., Gonçalves, E. C., Rios, E. H., Muhammad, A., Moss, A., Pritchard, T., Glassborow, B., Plastino, A., de Vasconcellos Azeredo, R. B., 2015. Automatic classification of carbonate rocks permeability from 1H {NMR} relaxation data. Expert Systems with Applications 42 (9), 4299 – 4309. Disponivel em: http://www.sciencedirect.com/science/article/pii/S0957417415000494

de Castro, M. R., 2003. BACIA DO ESPÍRITO SANTO/MUCURI - TERRA. Rel. tec., Brasil Round5.

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

de Oliveira, V. A. A., 2014. Caracterização de Reservatórios Não Convencionais/ Tight Gas. Projeto de conclusão do curso de graduação em Geofísica - Universidade Federal Fluminense.

de Oliveira Araujo, W., 2009. Análise de Componentes Principais (ACP). Rel. tec., Relatório Técnico - Mestrado Sociedade, Tecnologia e Meio Ambiente. Centro Universitário de Anápolis.

Egging, R., Franziska, F. H., Gabriel, S. A., 2010. The World Gas Model: A multi-period mixed complementarity model for the global natural gas market. Energy 35 (10), 4016–4029.

Folk, R. L., 1957. Petrology of sedimentary rocks. Hemphill Publishing Company.

França, A. B., Oliveira, C. G., Bacia Do Paraná: Rochas E Solos, 2010. Palestra Almerio Por Claudinei.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. Vol. 1. Springer series in statistics Springer, Berlin.

Fávera, J. C. D., 2001. Fundamentos de estratigrafia moderna. EdUERJ.

Gholami, R., Moradzadeh, A., Maleki, S., Amiri, S., Hanachi, J., 2014. Applications of artificial intelligence methods in prediction of permeability in hydrocarbon reservoirs. Journal of Petroleum Science and Engineering 122, 643 – 656.

Grahn, Y., 1999. Recent progress in the Silurian and Devonian biostratigraphy of the Paraná Basin in Brazil and Paraguay. Universidade do Estado do Rio de Janeiro, 147–163.

Gray, R. D., Atkinson, Q. D., nov. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426 (6965), 435–439. Disponivel em: http://dx.doi.org/10.1038/nature02029

Gray, R. D., Atkinson, Q. D., Greenhill, S. J., abr. 2011. Language evolution and human history: what a difference a date makes. Philosophical Transactions of the Royal Society B: Biological Sciences 366 (1567), 1090–1100. Disponivel em: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049109/

Gressly, A., 1938. Observations géologiques sur le Jura Soleurois. Neue Denkschr 2, 1–112.

Han, J., 2005. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Haykin, S., 2001. Redes Neurais, segunda Edição. Bookman.

Hyne, N., 2014. Dictionary of petroleum exploration, drilling & production. PennWell Corporation.

Imeson, A. C., 2005. Addressing soil erosion in Europe: proceedings of the SCAPE workshop in Alicante, Spain, June 2003. Land Degradation & Development 16 (6), 505–508. Disponivel em: http://dx.doi.org/10.1002/ldr.704

Iturraran-Viveros, U., Parra, J. O., 2014. Artificial Neural Networks applied to estimate permeability, porosity and intrinsic attenuation using seismic attributes and well-log data. Journal of Applied Geophysics 107, 45 – 54. Disponivel em: http://www.sciencedirect.com/science/article/pii/S092698511400144X

Jacomo, C. P., 2014. OS HIDROCARBONETOS NÃO CONVENCIONAIS: UMA ANÁLISE DA EXPLORAÇÃO DO GÁS DE FOLHELHO NA ARGENTINA À LUZ DA EXPERIÊNCIA NORTE AMERICANA. Tese de doutorado, Programa de Pós Grauduação em Planejamento Energético, COPPE, da Universidade Federal do Rio de Janeiro.

Johnson, R. A., Wichern, D. W., 2007. Applied Multivariate Statistical Analysis. 6.ed.ED. Person.

Jones, F. O., Owens, W., 1980. A laboratory study of low-permeability gas sands. Journal of Petroleum Technology 32 (09), 1–631.

Kastro, A., 2010. A geologia, os geólogos e seus métodos. Oficina de textos.

Kohavi, R., *et al.*, A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, 1995. Vol. 14. pp. 1137–1145.

Lacentre, P. E., C. P. M., 2013. A method to estimate permeability on uncored wells. World Oil 225.

Landis, J. R., Koch, G. G., 1977. The Measurement of Observer Agreement for Categorical Data 33, 159–174.

Lee, S. H., Datta-Gupta, A., Electrofacies Characterization and Permeability Predictions in Carbonate Reservoirs: Role of Multivariate Analysis and Nonparametric Regression. In: Society of Petroleum Engineers – Annual Technical Conference and Exhibiton, 1999.

Maraschin, A. J., Mizusaki, A. M., 2008. Datação De Processos Diagenéticos Em Arenitos-Reservatório De Hidrocarbonetos: Uma Revisão Conceitual 35(1), 27–41.

Milani, E. J., Brandão, J. A. S. L., Zalán, P. V., Gamboa, L. A. P., 00 2000. Petróleo na margem continental brasileira: geologia, exploração, resultados e perspectivas. Revista Brasileira de Geofísica 18, 352 – 396.

Milani, E. J., Zalan, P. V., 1999. An outline of the geology and petroleum systems of the Paleozoic interior basins of South America. Episodes 22, 199–205.

Motta, C. G. L., 2004. Sistema Inteligente para Avaliação de Riscos em Vias de Transporte Terrestre. Dissertação de mestrado, Universidade Federal do Rio de Janeiro.

Neto, J. M., Moita, G. C., 1998. Uma introdução à análise exploratória de dados multivariados. Química Nova 21 (4), 467–469.

Nissen, S., 2005. Neural Networks made simple. Software 2.0 2, 14–19.

Nocedal, J., Wright, S. J., 1999. Numerical Optimization. Springer-Verlag, New York.

Olatunji, S. O., 2011. Modeling the permeability of carbonate reservoir using type-2 fuzzy logic systems. Computers in Industry 62 (2), 147 – 163, fuzziness in Industry and Applications. Disponivel em: `http://www.sciencedirect.com/science/article/pii/S0166361510001569`

Oliveira, L. C., 2009. Estudos das relações entre o arcabouço estratigráfico e as alterações diagenéticas observadas na seção Devoniana da Bacia do Paraná. Dissertação de mestrado, Universidade do Estado do Rio de Janeiro.

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, 2009.

Ortiz Neto, J. B., Costa, A. J. D., 2007. A Petrobrás e a exploração de Petróleo Offshore no Brasil: um approach evolucionário. Revista Brasileira de Economia 61 (1), 95–109.

Pedregosa, F., Varoquax, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. D., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Petri, S., Fúlvaro, V. J., 1983. Geologia do Brasil. Editora Universidade de São Paulo.

Ponte, F. B., 2010. Fluxo de Trabalho para mapeamento de litofácies com aplicação no Campo de Namorado, Bacia de Campos, Rio de Janeiro. Monografia (Graduação em Geologia) – Universidade Federal Rural do Rio de Janeiro.

Poole, D., Mackworth, A., Goebel, R., 1997. Computational Intelligence: A Logical Approach. Oxford University Press, Oxford, UK.

Powers, D. M. W., Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation, 2007. Technical Report SIE-07-001.

Press, F., Menegat, R., 2006. Para entender a Terra. Vol. 656. Bookman Porto Alegre.

Quinlan, J. R., mar 1986. Induction of Decision Trees. Mach. Learn. 1 (1), 81–106. Disponivel em: http://dx.doi.org/10.1023/A:1022643204877

Raeesi, M., Moradzadeh, A., Ardejani, F. D., Rahimi, M., 2012. Classification and identification of hydrocarbon reservoir lithofacies and their heterogeneity using seismic attributes, logs data and artificial neural networks. Journal of Petroleum Science and Engineering 82, 151–165.

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, 2007. p. 1.

Roth, E. S., 1965. Temperature and Water Content as Factors in Desert Weathering. The Journal of Geology 73 (3), 454–468. Disponivel em: http://www.jstor.org/stable/30059267

Rousseeuw, P., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics.

Sharma, P., Mangain, G., Bahuguna, V. K., Lal, C., Improved Permeability Estimates in Carbonate Reservoirs Using Electrofacies Characterization: A Case Study of Mumbai High South. In: 2nd South Asian Geoscience Conference, 2012.

Silva, A. A., Neto, I. A. L., Misságia, R. M., Ceia, M. A., Carrasquilla, A. G., Archilha, N. L., 2015. Artificial neural networks to support petrographic classification of carbonate-siliciclastic rocks using well logs and textural information. Journal of Applied Geophysics 117, 118 – 125.

Suguio, K., 2003. Geologia Sedimentar. Blucher.

Tan, P.-N., Steinbach, M., Kumar, V., 2005. Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Technology, P. G., 2011. Geologia do Petróleo.

Teixeira, W., de Toledo, M. C. M., Fairchild, T. R., 2003. Decifrando a terra. Oficina de textos.

Toledo, M. C. M. e. a., 2000. Intemperismo e formação do solo. In: Decifrando a Terra. Decifrando a Terra, 140–166.

Tucker, M. E., Wright, V. P., 2009. Carbonate Sediments and Limestones: Constituents. Blackwell Publishing Ltd., pp. 1–27. Disponivel em: `http://dx.doi.org/10.1002/9781444314175.ch1`

Vapnik, V. N., Kotz, S., 1982. Estimation of dependences based on empirical data. Vol. 40. Springer-Verlag New York.

Vendramin, L., Campello, R. J. G. B., Hruschka, E. R., 2010. Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining 3(4), 209–235.

Walther, J., 1984. Einleitung in die Geologie als historische Wissenschaft. Gustav Fischer.

Wang, G., Carr, T. R., 2012. Marcellus Shale Lithofacies Prediction by Multiclass Neural Network Classification in the Appalachian Basin. Mathematical Geosciences 44 (8), 975–1004.

Wilcoxon, F., 12 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1 (6), 80–83.

Xie, J., 2008. Improved permeability prediction using multivariate analysis methods. Tese de doutorado, Texas A&M University.

Xiong, H., WU, J., Chen, J., 2009. K-means Clustering versus Validation Measures: A Data Distribution Perspective. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics 39.

Zhang, X., Collaboration of Information from Different Sources for Petroleum Reservoir Prediction. In: The Second International Conference on Information Fusion. Sunnyvale, USA: International Society of Information Fusion, 1999.

Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengaeuer, T., Muller, K. R., 2000. Engineering suppot vector machine kernels that recognize translation initiation sites in DNA. Bioinformatics (16), 906 – 914.

APÊNDICE A - Bases de Dados usadas na Dissertação

Figura A.1: QR-Code Bases de Dados. Endereço para consulta: `goo.gl/ocFYDZ`