Universidade de São Paulo
Instituto de Astronomia, Geofísica e Ciências Atmosféricas
Departamento de Geofísica

Raphael Fernandes Prieto

Aplicação de Aprendizado Supervisionado em Dados de Perfilagem Geofísica para a Classificação Automatizada de Litotipos na Exploração de Minério de Ferro em Carajás

São Paulo
2021

Raphael Fernandes Prieto

Aplicação de Aprendizado Supervisionado em Dados de Perfilagem Geofísica para a Classificação Automatizada de Litotipos na Exploração de Minério de Ferro em Carajás

Dissertação apresentada ao Departamento Geofísica do Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo como requisito parcial para obtenção do título de Mestre em Ciências.

Área de Concentração: Geofísica
Orientador: Prof. Dr. Ricardo I. F da Trindade

São Paulo
2021

Aos meus pais e ao meu irmão.

Agradecimentos

Aos meus pais, Sandra e Antonio, por todo amor e dedicação. Por me ensinarem que o caminho é mais importante que o destino. E ao meu irmão Lucas, por sempre se fazer presente.

À Nana pelo amor, paciência e companheirismo, que foram extremamente necessários para trilhar essa jornada.

Ao meu orientador Prof. Dr. Ricardo Ivan Ferreira da Trindade por acreditar no trabalho e pela relação de confiança construída ao longo desses anos fundamental para todo o desenvolvimento da pesquisa.

Aos amigos Dionísio e Wanderson pelo incentivo, suporte, discussões, "puxões-de-orelha" e por toda a mentoria.

Aos colegas geofísicos da "Legião Estrangeira" pelas discussões profundas, e quase sempre bem humoradas, acerca dessa ciência fascinante.

Aos colegas de mestrado pelos conhecimentos e sofrimentos compartilhados e principalmente pelos momentos de "descompressão".

Aos professores do Programa de Pós-Graduação em Geofísica do IAG-USP por acreditarem na ciência e por criarem o ambiente favorável para o contínuo desenvolvimento da mesma.

À Lilian Grabellos, então Gerente de Geologia e Sondagem de Ferrosos, por ter acreditado no projeto e apoiado seu desenvolvimento.

A todos que de alguma forma contribuíram para a realização deste trabalho e à Vale pela disponibilização dos dados.

“Porque a vida é isso: um emaranhado de nós.”
- Autor desconhecido -

Resumo

Na indústria da mineração a principal forma de investigação geológica é através da sondagem exploratória, e das etapas subsequentes de descrição geológica de testemunhos de sondagem, preparação de amostras e análises químicas. Essas são atividades de mineração que possuem processos bem estabelecidos e elevado grau de confiabilidade. Entretanto os prazos envolvidos nessas etapas, no melhor dos cenários, variam na escala de semanas a meses. Nesse trabalho foi utilizado um conjunto de dados integrando as bases de dados de perfilagem geofísica convencional e de descrição geológica, na jazida de S11D, Província Mineral de Carajás, com o objetivo de desenvolver um modelo de classificação automatizada de litotipos a partir dos dados de perfilagem geofísica, sob a abordagem de aprendizado supervisionado, acelerando o processo de descrição e modelagem geológica a escala de dias. Os procedimentos para a classificação de dados visaram a individualização de diferentes litotipos no contexto da exploração de minério de ferro em S11D, e também de acordo com seu valor econômico (Minério de Ferro e Não-minério). Os dados obtidos pela perfilagem geofísica convencional refletem, nessa jazida, diferenças geológicas que permitiram, a classificação e predição bem sucedidas dos litotipos (F1 = 0.6079), pelo modelo *Decision Tree*, e a identificação dos intervalos contendo minério de ferro pelo modelo *Naïve Bayes* (F1 = 0.8246).

Palavras – Chaves: Aprendizado de Máquina; Perfilagem Geofísica; Minério de Ferro; S11D; Carajás

Abstract

In the mining industry, the main subsurface investigation method is exploratory drilling, and the subsequent stages of geological logging, sample preparation and chemical assay that supports geological interpretations. These are exploratory activities that have well-established processes and a high degree of reliability. However, the deadlines involved in these steps, at best, vary from weeks to months. In this work, a dataset integrating geophysical well logging and of the geological logging from S11D deposit, in the Mineral Province of Carajás, was used with the objective of developing a model for automated lithotypes and iron ore classification from geophysical well logging data, under the supervised learning approach, decreasing the time expended in comparison with geological interpretations. The procedures aimed at the individualization of different lithotypes in the context of S11D, and also according to its economic value (Iron Ore and Non-ore). The data obtained by conventional geophysical logging reflect, in this deposit, geological differences that allowed the successful classification and prediction of lithotypes (F1 = 0.6079), by the Decision Tree model, and the identification of the intervals containing iron ore by Naïve Bayes model (F1 = 0.8246).

Keywords: Machine Learning; Geophysical Logging; Iron Ore; S11D; Carajás.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

O desenvolvimento eficiente de uma jazida depende, dentre outros fatores, de uma boa interpretação da geologia de subsuperfície e, consequentemente, da construção de um bom modelo geológico. Para que isto ocorra se faz necessária a coleta de uma série de dados espaciais envolvendo diversos ramos das geociências. Na indústria da mineração a principal forma de investigação geológica é através da sondagem exploratória e das etapas subsequentes de descrição geológica de testemunhos de sondagem, preparação de amostras e análises químicas. Essas são atividades rotineiras na mineração, que possuem processos estabelecidos e elevado grau de confiabilidade. Os prazos envolvidos nas etapas desses processos para a extração de bons dados e interpretação, no melhor dos cenários, variam desde semanas até meses.

Apesar das medidas petrofísicas obtidas pela perfilagem geofísica não mapearem diretamente as litologias em subsuperfície, em muito ambientes geológicos as propriedades físicas do maciço podem estar relacionadas a uma série de feições e características de interesse, como estilo de alteração e mineralização, textura, qualidade/resistência do maciço rochoso, indicação da presença/ausência de certos minerais e metais no maciço, etc. Em função de sua elevada resolução espacial e da velocidade de aquisição dos dados, essa tecnologia também permite aos geocientistas interpretações e análises que são limitadas para as demais metodologias de obtenção de dados diretos do maciço. Os prazos envolvidos no processo de perfilagem variam, na maioria dos casos, desde horas até dias.

No contexto da indústria mineral, são cada vez mais comuns algumas aplicações da perfilagem geofísica, como correlação furo-a-furo, delimitação de corpos de minério e estimativa de teor. Perfilagens de densidade, gama natural e suscetibilidade magnética foram utilizadas na delimitação de corpos de metais base em sulfetos (Wanstedt, 1992). Sistemas eletromagnéticos de alta frequência também

foram utilizados na delineação de corpos já conhecidos de níquel sulfetado em Sudbury (Fullagar et al., 2000). No contexto da mecânica das rochas também foi estudada a aplicação da perfilagem geofísica na determinação de parâmetros geomecânicos em maciços (Pereira, 2017), usando dados da perfilagem gama natural, densidade e *Full Wave Form* (FWS).

Nas últimas três décadas houve um aumento no número de estudos de aplicação de técnicas de estatística multivariada, *data analytics* e aprendizado de máquina nas geociências, com as mais diversas finalidades, tais como: identificação de padrões, seleção de alvos de exploração, determinação de propriedades dos materiais geológicos pela combinação de diferentes medidas. São exemplos: Determinação da litologia, porosidade e grau de faturamento utilizando análise fatorial regressão linear e multilinear em dados de perfilagem geofísica (Pechnig et al., 1997); Determinação de contatos geológicos a partir da aplicação de redes neurais em dados de densidade, porosidade neutrônica, e gama natural (Maiti et al., 2007); Aplicação de diversos algoritmos de aprendizado de máquina para a classificação de fácies e determinação de intervalos mineralizados em ouro (Blouin et al., 2017; Caté et al., 2017).

Os procedimentos de classificação de dados apresentados nessa dissertação visaram a individualização de diferentes litotipos no contexto da exploração de minério de ferro (classificação visual - CLV), e também de acordo com seu valor econômico (Minério de Ferro e Não-minério). As variáveis (atributos) da perfilagem geofísica convencional refletem diferenças que permitiram, sob a abordagem de aprendizado supervisionado, a classificação e predição bem sucedidas dos litotipos e a identificação dos intervalos contendo minério de ferro. O conjunto de dados utilizado nessa tarefa foi fruto da integração das bases de dados de perfilagem geofísica convencional e da descrição geológica na jazida de S11D, Província Mineral de Carajás.

1.1 Contexto Geológico e Aspectos Operacionais do Depósito S11D

As jazidas de minério de ferro das Serras de Carajás foram descobertas em agosto de 1967 pelos geólogos da Companhia Meridional de Mineração, quando efetuavam programa sistemático de exploração a procura de jazidas minerais na região norte do Brasil.

A principal jazida de minério de ferro está situada na porção sudoeste da Província Mineral de Carajás – PMC – (Figura 1), umas das mais proeminentes província minerais do mundo, situada a 90 quilômetros ao sul das minas de ferro que compõem o Sistema Norte (N4 e N5), na região denominada de Serra Sul alvo 11, e corresponde aos corpos S11C e S11D. Essa jazida é constituída por rochas pré-cambrianas recobertas em grande parte por cangas, que são formações superficiais derivadas da alteração supergênica destas rochas (VALE, 2016). Os principais depósitos de minério de ferro estão associados a clareiras localizadas nos topos das serras (platôs) da região.

Os depósitos de minério de ferro de Carajás pertencem a unidades metavulcanosedimentares do Supergrupo Grão Pará (*sensu* Macambira, 2003). A mineralização é gerada pelo intemperismo da unidade da formação ferrífera bandada (BIF) denominada Formação Carajás. O BIF é composto por jaspelitos (JPC e JPF), enriquecidos durante processos de intemperismo supergênico, responsáveis pela formação de corpos de hematititos. Esses corpos de hematititos (HF) têm grandes extensões, alto teor de ferro (66%) e são tipicamente friáveis. A relação entre os contatos das unidades de alto teor de ferro e da unidade de baixo teor de ferro não alterada (jaspelitos) é nítida e irregular, apresentando pontões (picos) de jaspelitos compactos no meio do minério de hematita friável. Essas unidades de ferro são encontradas intercaladas em rochas metavulcânicas das formações Parauapebas e Igarapé Cigarra e são cobertas por unidades de solo e laterita detríticas, denominadas canga (CG), geradas durante o processo de evolução pedológica.

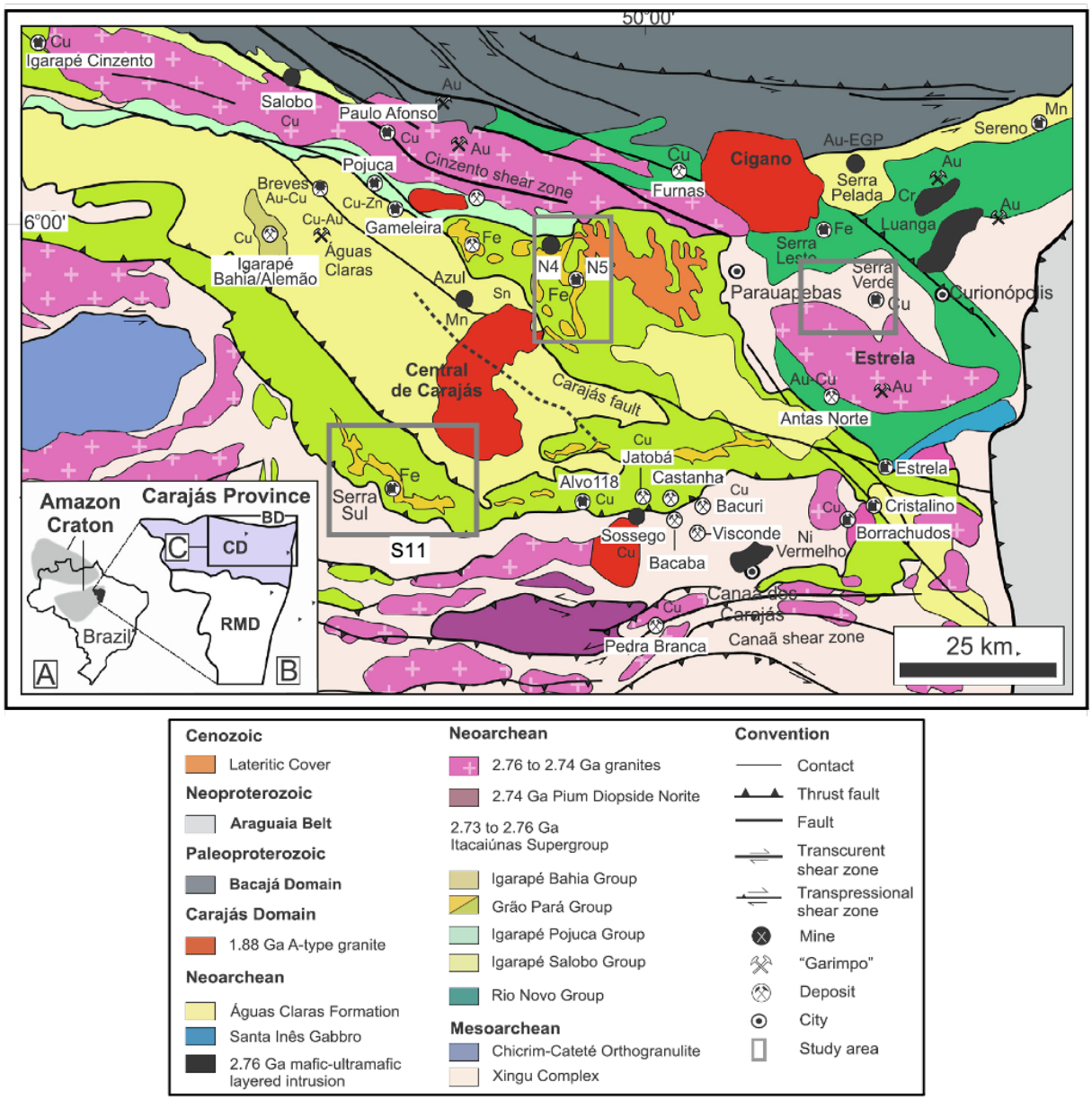


Figura 1 - Província Mineral de Carajás e identificação do Alvos para a exploração de Mfe (polígonos cinza). Modificado de (Figueiredo e Silva et al., 2020).

Devido à importância e relevância estratégica do Projeto S11D (produção estimada 90Mt/ano de minério de ferro), um método inovador de lavra conhecido como *Truckless* foi proposto. Essa abordagem faz uso de carregadoras, britadores e correias transportadoras de minérios e rejeitos modulares, ou seja, que acompanham o avanço da frente de lavra (Roldão et al., 2012).

A operação de mina no projeto S11D é composta por quatro frentes de lavra com sete sistemas de britagem. O uso de cada sistema depende do tipo de material que será minerado. Inicialmente os sistemas de britagem foram dimensionados

para materiais friáveis (Hematitito Friável - HF). Cada um desses sistemas atua em três bancos de lavra de 15 metros de altura cada. O equipamento projetado para trabalhar com materiais friáveis (HF) tem capacidade de produção de 8500 toneladas por hora, entretanto possui capacidade máxima de 24 horas de trabalho contínuo em materiais compactos (Jaspelito Compacto - JPC e Jaspelito Friável - JPF). Após esse período a abrasão do equipamento é muito elevada.

As vantagens apresentadas por esse método de lavra são diversas, quando comparado com o convencional. Destaque dado para a redução de emissão de material particulado, redução na infraestrutura de mina necessária, redução do impacto ambiental, processos mais automatizados e com maior controle instrumental.

A grande desvantagem é a menor flexibilidade operacional, ou seja, necessita de grande previsibilidade de sequenciamento de lavra para o dimensionamento adequado dos equipamentos (britadores principalmente). Portanto, variações de geologia na frente de lavra não contempladas pelo modelo geológico afetam de forma significativa o planejamento adequado da lavra e impactam diretamente toda a cadeia de processos para a produção de minério de ferro.

1.2 Objetivos

No contexto da exploração de minério de ferro do corpo S11D, esse trabalho tem os objetivos de:

i. Contribuir para o aumento do conhecimento geológico do depósito à luz da caracterização petrofísica do minério e dos Jaspelitos;
ii. Dinamizar o processo de descrição geológica através da criação de um modelo para a classificação automatizada de litotipos com base em dados de perfilagem geofísica, utilizando aprendizado supervisionado:
    a. Classificação multiclasse (ou multi-rótulos) – Classificação de Litotipos;
    b. Classificação binária – Detecção de Minério de Ferro (MFe).

2 Metodologia

As subseções que seguem apresentam as bases teóricas que suportaram a obtenção dos dados e os conceitos fundamentais de aprendizado supervisionado,

2.1 Descrição Geológica de Testemunhos de Sondagem Rotativa

A sondagem geológica consiste na extração de amostras de material geológico em subsuperfície a partir da superfície empregando-se uma unidade de perfuração que possui uma ponta rotativa de recorte chamada sonda (daí o nome sondagem). A sonda possui diâmetro de perfuração da ordem de poucos centímetros e pode atingir centenas, ou mesmo milhares de metros de profundidade de perfuração. Os testemunhos recuperados podem variar de centímetros a três metros, no máximo.

A descrição geológica dos testemunhos de sondagem é uma das atividades realizadas pelos geólogos na exploração de recursos minerais, com o intuito de reconhecer e descrever a geologia em subsuperfície. Essa atividade consiste na identificação tátil e visual das características mineralógicas, faciológicas, texturais e estruturais dos materiais geológicos amostrados na forma de testemunhos de sondagem, como pode ser observado na Figura 2.

Figura 2 - Caixas de testemunhos de rocha utilizados para descrição geológica, com indicação da denominação do furo de sondagem e do intervalo de profundidades correspondente.

Os dados adquiridos na descrição geológica suportam a análise qualitativa e quantitativa do corpo de minério, tornando-se uma ferramenta fundamental nos estudos de viabilidade econômica de um projeto, considerando-se as seguintes definições:

i. Intervalos geológicos: são unidades geológicas, fisicamente distintas, cujas representações nos modelos geológicos e geotécnicos são possíveis, face as características como: contatos definidos, identificação inequívoca, correlação e continuidade espacial e, seletividade da lavra.
ii. Tipos litológicos ou litotipos: minérios e rochas estéreis “in situ”, isto é, antes de terem sofrido movimentação de lavra. Os tipos litológicos devem corresponder a horizontes mapeáveis na jazida.
iii. Formação ferrífera: grupo de rochas que possuem como componentes mais representativos os minerais de ferro (hematitas e magnetita).
iv. Minério de ferro: rocha composta basicamente de minerais de ferro (hematitas e magnetita) e quartzo secundariamente. Dividido entre minério limpo e minério contaminado.
v. Minério limpo: rocha composta predominantemente de minerais de hematita e magnetita.
vi. Minério contaminado: rocha composta por minerais de ferro e porcentagens consideráveis de minerais de ganga ou pela presença de rochas encaixantes ou fragmentos de rocha estéril.
vii. Rocha estéril: rocha que não possui valor econômico para a indústria.
viii. Materiais superficiais: Caracterizados por materiais encontrados nas camadas mais superficiais como Solos, Cangas, Lateritas, Psolitos, Colúvios, Aterros e Pilhas de Materiais.
ix. SR: sigla para intervalo com recuperação igual a zero (R=0).

x. DT: sigla para intervalo de rocha destruída.

xi. Classificação visual (CLV): classificação dos tipos litológicos, conforme propriedades geológicas visuais, definidas durante a descrição geológica dos furos de sonda.

xii. Compacidade de rocha: classificação geotécnica definida pela identificação táctil-visual das características de resistência ao impacto, risco, compressão uniaxial e trabalho do material.

As observações feitas durante a descrição geológica são documentas em tabelas, por intervalos de descrição ao longo do testemunho, sendo um dos principais dados de entrada nos demais processos de prospecção e exploração mineral (confecção de seções e modelos geológicos, estimativa de recursos, planejamento de lavra e etc.).

O tempo estimado entre os términos do furo e da descrição geológica é de 3 semanas para furos prioritários, podendo chegar a poucos meses para o fluxo normal de trabalho.

A descrição geológica dos testemunhos de sondagem rotativa, utilizada nesse trabalho, baseou-se na identificação dos contatos geológicos definidos através da CLV (Classificação Visual), sendo a nomenclatura dos litotipos representada por siglas. As siglas seguiram critérios padronizados pela companhia Vale S.A. sendo compostas de dois a quatro caracteres a depender dos litotipos e de suas variáveis, sendo representadas por letras maiúsculas (6.1 - Códigos Gerais para Litotipos), segundo seguintes os critérios e exceções:

i. Minérios limpos: caracterizados por duas variáveis litologia (1 caractere) e compacidade (1 caractere), totalizando dois caracteres.

ii. Minérios contaminados: caracterizados por três variáveis, litologia (1 caractere), contaminação (2 caracteres) e compacidade (1 caractere), totalizando quatro caracteres.

iii. Rochas estéreis: caracterizados pelas variáveis: litologia, composição e compacidade, sendo representados por três ou quatro caracteres. As exceções dos estéreis - Veio de quartzo (VQ), zona de cisalhamento (ZC) e zona de transição (ZT) – que não apresentam compacidade portanto serão definidos com os dois caracteres da litologia.

iv. Coberturas: não apresentam compacidade ou contaminação, e seus códigos são descritos por suas siglas sendo representadas por dois caracteres. Com exceção do Pisólito que tem 3 caracteres.

v. Outros códigos: Caracterizam-se por códigos usados na sondagem para representar intervalos sem recuperação ou litologias destruídas, esses são descritos por suas siglas e representados por dois caracteres.

Os intervalos litológicos utilizados no processo de descrição foram definidos pelas áreas operacionais da Vale S.A., de acordo com as características das jazidas em cada área. Estes variaram de 1 à 7.5 metros.

Quando o intervalo da Formação Ferrífera correspondeu ao contato de topo e de base com rochas encaixantes, este foi diferenciado, independente do comprimento do intervalo. Quando o intervalo da Formação Ferrífera foi menor que o determinado para área e não estando este em contato com rocha encaixante, o intervalo foi englobado na litologia de maior afinidade imediatamente acima ou abaixo. Neste caso, a descrição foi colocada no campo das observações, juntamente com a metragem do início e fim dessa passagem.

Os intervalos das rochas estéreis devem ter comprimento mínimo de 1 metro. Quando o intervalo rocha estéril foi menor e não sendo esta uma rocha encaixante, o intervalo foi englobado na litologia de maior afinidade imediatamente acima ou abaixo. Neste caso, a descrição foi colocada no campo das observações, juntamente com a metragem do início e fim dessa passagem.

Os intervalos de materiais superficiais (Cangas, Lateritas, Psolitos, Colúvios, Aterros e Pilhas Estéril) e de rocha destruída (DT) foram diferenciados independente do comprimento. Intervalos sem recuperação (SR) foram indicados quando o comprimento é de mínimo de 1,5 metros

No processo de descrição geológica foram majoritariamente individualizadas as tipologias de minério – canga estrutural (CE), hematita friável (HF), hematita compacta (HC) e hematita manganesífera (HMN); e, de estéril: canga química (CQ), jaspelito (JP), Máfica decomposta (MD), Máfica semidecomposta (MSD) e Máfica sã (MS). Os demais litotipos, ou mesmo as variações específicas desses já apresentados, que ocorrem tipicamente em menor proporção, são descritos e definidos ao longo do texto conforme a necessidade, uma vez que não é objetivo deste trabalho realizar a revisão dos litotipos encontrados no contexto geológico do corpo S11D.

2.2 Perfilagem Geofísica

A perfilagem geofísica nasceu no final da década de 1920 em Pechelbronn, pelos esforços dos irmãos Schlumberger, quando eles amostraram, de forma semi-contínua, a resistividade elétrica ao longo de um furo/poço de exploração em um campo maduro de óleo na região da Alsácia, França.

Essa metodologia basicamente pode ser descrita como a inserção de uma sonda de perfilagem, ou seja, uma haste metálica com sensores acoplados (ferramentas), no interior de um furo de sondagem (Pereira, 2017). O processo de medição é dado por meio de um guincho ligado à sonda por um cabo útil, pelo qual transita a informação proveniente dos sensores contidos na sonda. Para a investigação das informações ao longo do furo a sonda de perfilagem é içada à medida que registra as variações de sinal nas ferramentas (Figura 3).

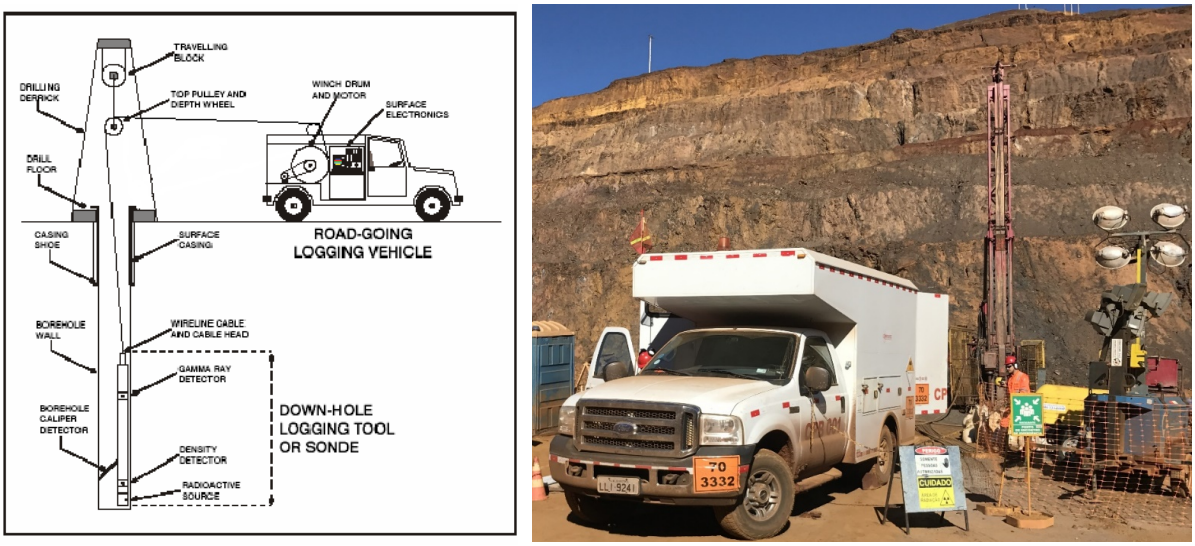


Figura 3 - Elementos do processo de perfilagem geofísica: Esquemático (esquerda) e unidade móvel de perfilagem em uma praça de sondagem geológica para mineração (direita) (imagens cedidas pela Comprobe/Vale S.A.)

Uma variedade ampla de propriedades físicas do maciço/formação com alta resolução espacial é obtida pela perfilagem, também chamada de petrofísica, principalmente quando comparada com a resolução obtida com as demais técnicas convencionais de análise química. O tempo estimado para a conclusão do processo de perfilagem de um furo é inferior à 1 (um) dia útil de trabalho.

No contexto na exploração de minério de ferro, a perfilagem pode ser subdividida em duas categorias:

a. Perfilagem Convencional – Composta pelas ferramentas Caliper (diâmetro do furo), Temperatura, Gama Natural (atividade radioativa) e Gama-Gama (densidade *in situ*);

b. Perfilagem Multiferramentas – Composta pelas ferramentas da perfilagem convencional adicionadas outras ferramentas como Susceptibilidade Magnética, Resistividade, Gama Espectral, *Full Wave Sonic* (Velocidade de ondas P e S), Polarização Induzida, Imageamentos acústico e visual (ATV e OTV), dentre outros.

Nesse trabalho foram utilizados os dados provenientes da perfilagem geofísica convencional, disponíveis para a jazida de S11D. Na sequência é realizado um breve detalhamento das ferramentas Gama Natural e Gama-Gama, em função do caráter óbvio das ferramentas Caliper (variação do diâmetro do furo) e Temperatura.

Um aspecto que vale a pena destacar é o fato de que a perfilagem de medidas nucleares responder tanto pelos fluídos contidos nas formações quanto pelas características da sua matriz rochosa, ao contrário das ferramentas baseadas nas propriedades elétricas dos materiais, que respondem majoritariamente pelo conteúdo de fluidos contidos nas formações (Ellis & Singer, 2008). Tal fato, por si, justifica a aplicação das ferramentas Gama-Gama e Gama Natural na Perfilagem Geofísica Convencional no contexto da exploração de minério de ferro. Somado a isso temos que a densidade medida pela perfilagem Gama-Gama, apresenta papel de destaque na estimativa dos recursos e reservas, no qual a massa de minério é determinada pelo produto entre teor de metal, volume e densidade.

2.2.1 Gama Natural

A perfilagem Gama Natural mede a emissão de raios gamas naturais originados nos materiais geológicos. As medidas são feitas em contagens por segundo (cps) ou dadas em grau API (calibradas pelas referências no *American Petrolleum Institute*).

Os raios gama naturais são originados em três fontes principais nos materiais geológicos: no grupo dos elementos radiogênicos Urânio (235U e 238U) e Tório (232Th), que decaem para o isótopo estável de Chumbo, e no Potássio (40K), que decai para Argônio emitindo radiação gama. A Figura 4 mostra o esquema de decaimento do 40K para 40Ar, juntamente com os espectros de probabilidade de emissão dos elementos U, Th e K por faixa energética (Bateman, 2015). Nela podemos observar que os três elementos possuem picos distintos de maior probabilidade de emissão, individualizando-se as respectivas faixas energéticas.

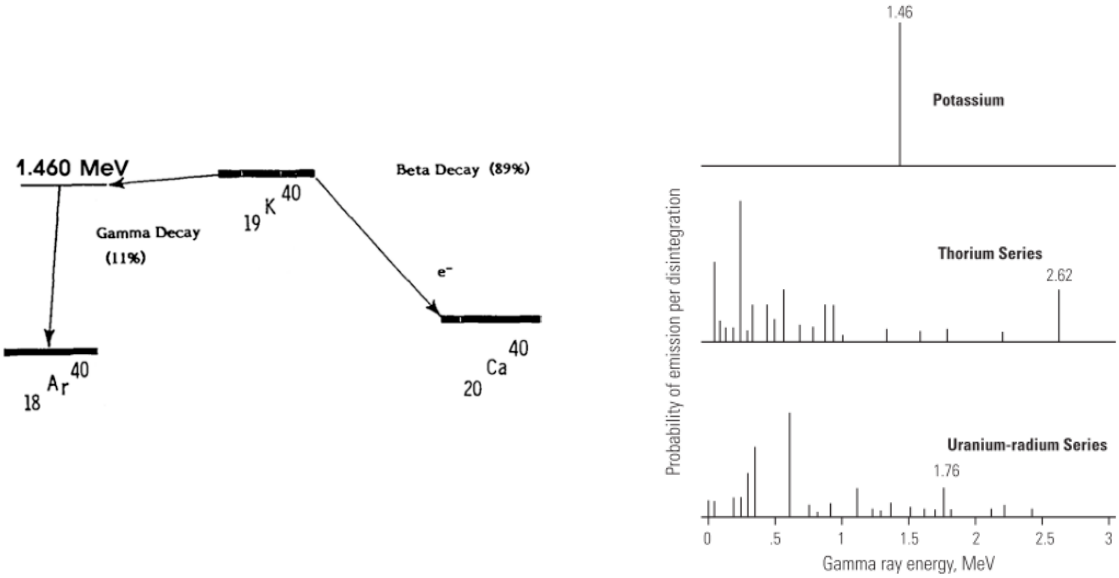


Figura 4 - Esquema de decaimento do isótopo ^{40}K (esquerda). Espectro de radiação gama dos minerais radiogênicos (direita). Extraído de (Bateman, 2015).

O processo de detecção da radiação gama possui duas etapas. Primeiramente existe a interação dos raios gama com o detector. Nesse contato há a conversão de toda a energia contida nos raios gama em radiação ionizante (elétrons energéticos). Já na segunda etapa ocorre a conversão desses elétrons em um sinal elétrico.

A ferramenta Gama Natural consiste em um detector de radiação gama, cintilador, contendo como elemento sensor tipicamente um cristal de Iodeto de Sódio dopado com Tálio ou Germaneto de Bário. Cada vez que esse cristal é atingido por um raio gama ele emite um fóton. Esse fóton é capturado pelo fotocátodo que libera uma série de elétrons. Esses por sua vez são acelerados e replicados na fotomultiplicadora gerando um pulso de tensão de saída no resistor de medição. Esse sinal passa por um amplificador antes de ser digitalizado pelo conversor analógico digital (Figura 5)

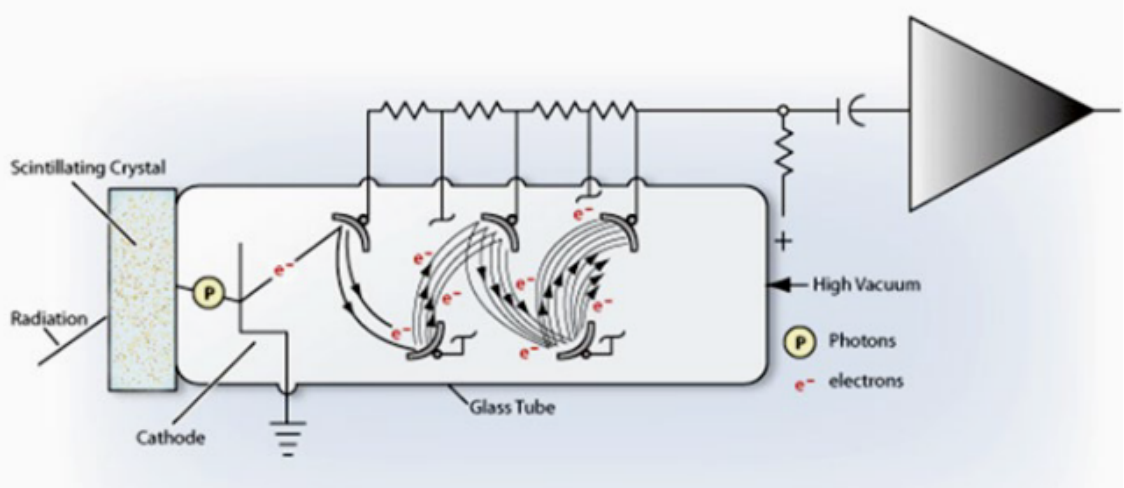


Figura 5 - Esquema ilustrativo de um Cintilador, ou detector de raios gama (Bateman, 2015).

Todas as ferramentas gama natural saem de fábrica após passar por um rigoroso protocolo de calibração, validado em campo de prova, para garantir a padronização das leituras, ou seja das contagens.

2.2.2 Gama-Gama (Densidade)

Os diferentes tipos existentes de interação entre a radiação gama e os materiais ocorrem em função das características do material (essencialmente número atômico e densidade de partículas, ou densidade) e da energia dessa radiação. A perfilagem Gama-Gama utiliza as interações do material geológico com a radiação gama gerada por uma fonte na ferramenta. Ela difere da ferramenta Gama Natural, que mede os raios gama naturais originados no decaimento dos isótopos radiogênicos dos elementos U, Th e K.

A fonte de raios gama mais largamente utilizada na indústria é a fonte de ^{137}Cs, que emite raios gama a uma energia de 662 keV, suficientemente abaixo da faixa de energia associada à produção de pares (Figura 6) e acima da faixa de efeito fotoelétrico (Ellis & Singer, 2008). Outra fonte padrão industrial, de uso mais restrito, é a fonte de ^{60}Co, que emite duas faixas de energia em torno de 1200 keV. Essa fonte permite a medição de densidades maiores que 4.0 g/cm^3 (Pereira, 2017). Sendo então a fonte de ^{60}Co utilizada nas ferramentas perfilagem geofísica Gama-Gama no S11D.

Dentre estes tipos de interação da radiação gama com a matéria, o Espalhamento Compton é a interação modelada para a obtenção da densidade dos materiais geológicos (Ellis & Singer, 2008). Esse é o principal processo de interação da radiação gama com a matéria. Nela a partícula gama colide com um elétron livre sofrendo espalhamento elástico. Nesse ponto assume-se que a energia da partícula gama incidente é significativamente superior à energia de ligação do elétron que compõe o material, podendo este ser considerado livre e em repouso.

A função de atenuação passa a ser baseada nas *N* ocorrências detectadas em comparação com as N_0 emissões de uma determinada fonte de raios gama localizada a uma distância *x*, de acordo com a relação de atenuação:

$$N = N_0 e^{-\mu_a \rho x}$$

Onde:

μ_a é o coeficiente de absorção mássica (cm^2/g)

ρ é a densidade do material (g/cm^3)

x é a distância fonte-receptor (cm)

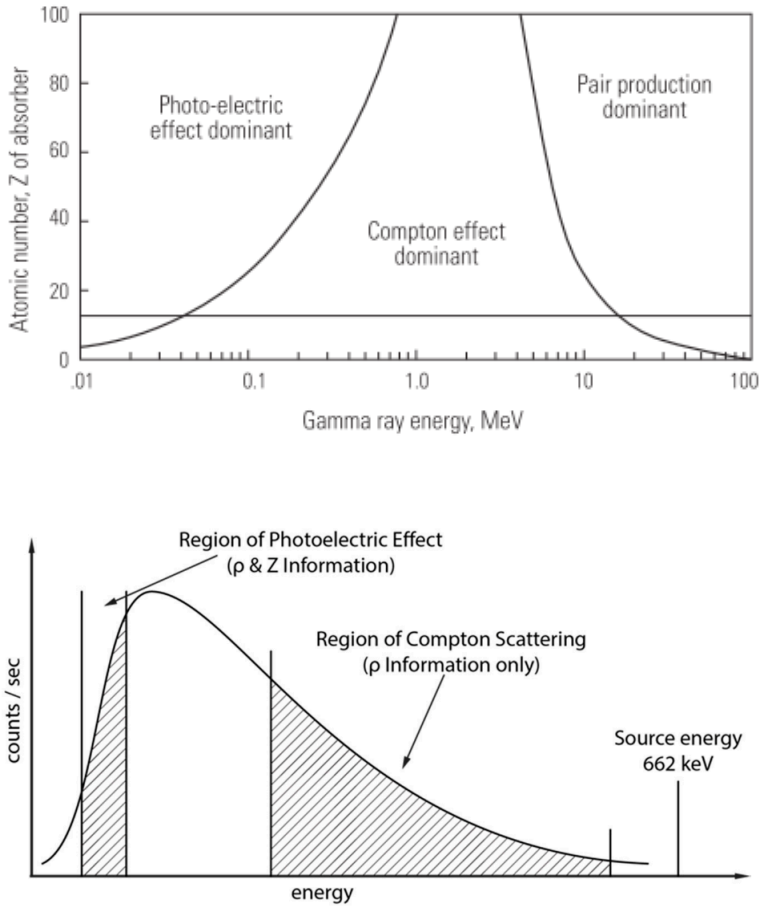


Figura 6 - Regiões de predomínio dos principais mecanismos de espalhamento de raios gama em função da energia e do número atômico do material de interação (Ellis & Singer, 2008) e respectiva curva de atenuação do fluxo de radiação gama em função da energia.

A Figura 7 mostra a comparação da curva de atenuação da contagem gama para uma fonte de ^{137}Cs com a curva de atenuação de uma fonte de ^{60}Co (Pereira, 2017). Para materiais geológicos de maior densidade (caso esperado no contexto da exploração mineral de metais), a escolha da fonte adequada tem reflexo direto na qualidade da densidade recuperada. Como pode ser observado na Figura 7, para densidades superiores a $4\ g/cm^3$, o decaimento da fonte ^{137}Cs para uma grande variação de densidades reflete uma pequena variação de contagem gama.

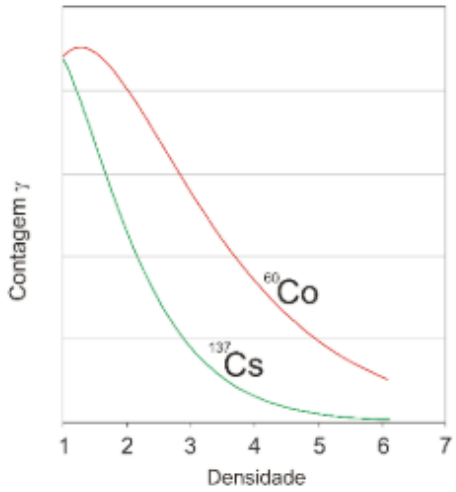


Figura 7 - Variação das contagens em função da densidade para uma fonte de 137 Cs (verde) e uma fonte de 60 Co (vermelho)

A ferramenta Gama-Gama utilizada neste trabalho para a obtenção dos valores de densidade dos materiais geológicos compreende uma fonte de ^{60}Co (Cobalto) com energia de 1200 keV, juntamente a dois detectores (longo e curto), compondo o arranjo fonte-detector. Os detectores são blindados com relação a fonte, registrando-se então apenas os raios gama provenientes do espalhamento, fruto da interação entre os raios gama com o material geológico na vizinhança do ponto de medida ao longo do furo (Figura 8). A profundidade de investigação e a resolução vertical dependem da geometria fonte-detectores, que no caso dos dados para S11D estão espaçados a 16 cm e 29 cm da fonte.

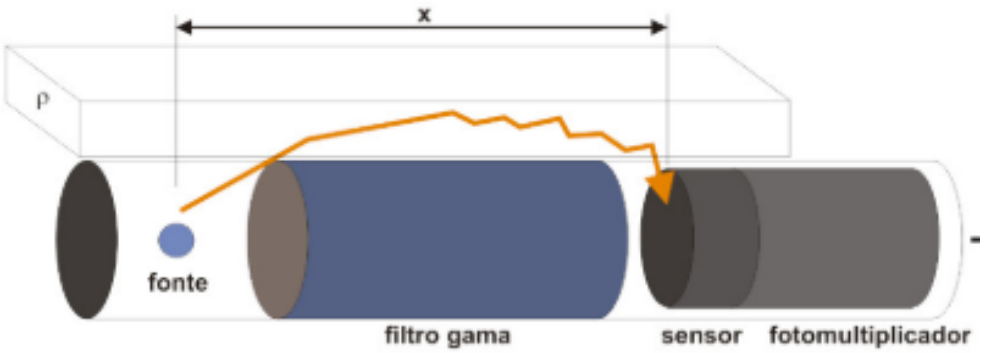
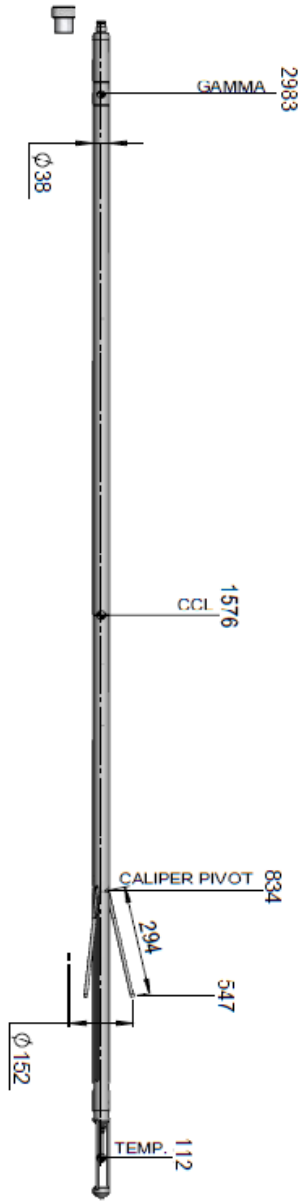


Figura 8 - Arranjo simplificado fonte-sensor utilizado no processo de perfilagem geofísica Gama-Gama (Pereira, 2017).

As sondas de perfilagem podem frequentemente conter mais do que uma ferramenta em seu corpo e ao longo de uma mesma campanha de perfilagem sondas distintas podem ser utilizadas. Como estas portam algumas ferramentas comuns às outras sondas (normalmente *caliper* e *gama natural*), isso permite, além da validação cruzada, a aferição da posição em profundidade das leituras.

A Figura 9 mostra as ferramentas utilizadas nas campanhas de perfilagem convencional no contexto da exploração de minério de ferro em S11D. Vale a pena destacar que o uso indistinto dos termos sondas e ferramentas de perfilagem ocorre frequentemente, mesmo na literatura técnica.

GTC (Gamma-Temperatura-Caliper)



DD6 (Gamma–Densidade–Caliper)

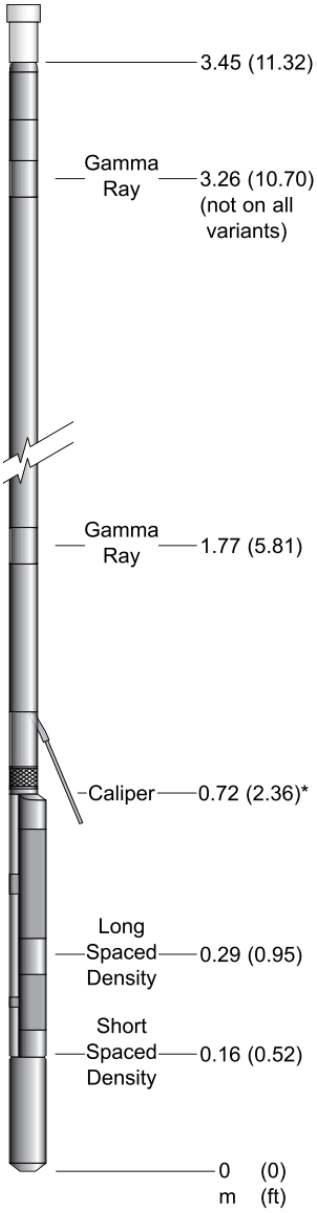


Figura 9 - Esquema das ferramentas utilizadas nas campanhas de perfilagem convencional em S11D. (Desenho cedido por Comprobe)

2.3 Aprendizado de Máquina Supervisionado

As definições e conceitos apresentados aqui basearam-se na documentação das bibliotecas *python Scikit-Learn* (Pedregosa et al., 2011) e *Orange* (Demsar et al., 2013), e nos livros "*Hands-On machine Learning with Scikit-Learn & TensorFlow*" (Géron, 2017) e "*Data Mining And Knowledge Discovery For Geoscientists*" (Shi, 2014).

O conjunto de técnicas e métodos de programação, álgebra linear, otimização, estatística, cálculo multivariado, dentre outros, que possibilitam que os computadores aprendam a partir dos dados, ou seja, estes aprendam sem que as regras/instruções sejam explicitamente programadas é dado de Aprendizado de Máquina.

Além da capacidade de aprender a partir dos dados, essa estrutura e abordagem tem o efeito positivo de contribuir para um melhor entendimento dos dados, e consequentemente do problema, uma vez que, nessa abordagem, não é utilizado nenhum tipo de conhecimento que modele o fenômeno/problema, o foco está em entender o problema a partir dos dados, capturar essência e pro vezes relações complexas entre os atributos (variáveis) que compõe a base de dados.

O aprendizado de máquina costuma ter bom desempenho em problemas que: necessitam de muitos ajustes manuais para ser modelado; muitas regras para em uma cadeia de análise.; volumes de dados muito grandes para permitirem visualização ou mesmo que sejam analisados segundo algum modelo.

Os sistemas de aprendizado de máquina podem ser agrupados em categorias baseadas no tipo e quantidade de informação utilizada na aprendizagem; se a aprendizagem é incremental ou tempo real, e se os dados novos são comparados diretamente com dados conhecidos ou se durante a aprendizagem são mapeadas características padrões do conjunto de dados e, a partir disso, construído um modelo preditivo (Géron, 2017).

O processo de aprendizado é composto por pelo menos cinco etapas (Figura 10). Cada uma dessas etapas é apresentada na seção de processamento dos dados.

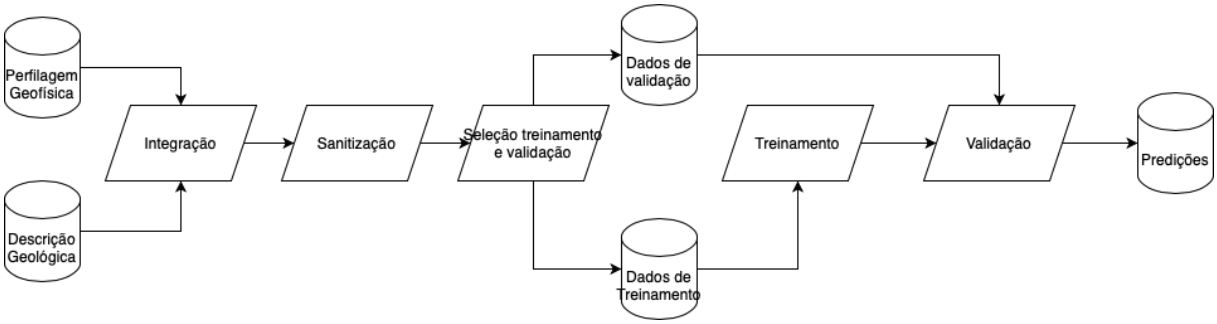


Figura 10 – Processo de aprendizado de máquina adotado neste trabalho.

Nesse trabalho foram utilizados algoritmos de aprendizado supervisionado, incremental e baseado em modelo. Na etapa de treinamento (Figura 11), o algoritmo escolhido recebe como entrada a base de dados de treinamento, composta pelos atributos e rótulo da variável-alvo para cada amostra (por isso o nome supervisionado), e tem como saída uma função não linear (modelo). Na etapa de predição o algoritmo recebe como entrada a base de dados de predição, composta apenas pelos atributos, junto do da função não-linear (modelo fruto da etapa de treinamento), e tem como saída o rótulo, predito pelo modelo, em cada uma das amostras da base de dados da predição.

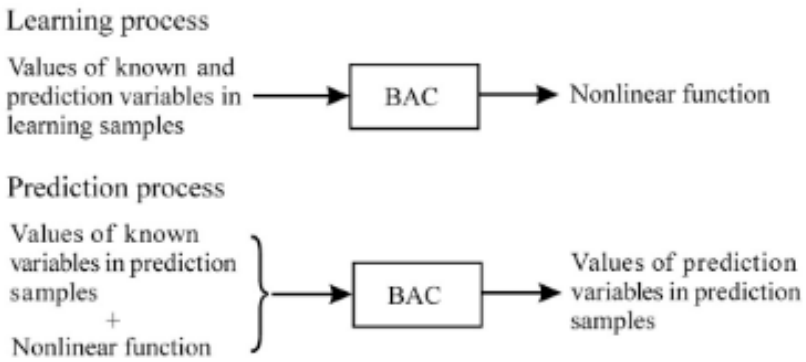


Figura 11 - Exemplo de processos de treinamento e predição. Extraído de (Shi, 2014).

O aprendizado supervisionado pode ser do tipo multiclasse (multi-rótulos), quando o objetivo é predizer uma saída composta de mais de um rótulo para a variável-alvo. Ou pode ser do tipo binário, quando a variável-alvo apresenta apenas dois rótulos possíveis como saída. Os problemas abordados neste trabalho são de

amos os tipos, sendo a classificação de litotipos um problema multiclasse e a determinação dos intervalos mineralizados um problema do tipo binário.

Como o objetivo desse trabalho foi dinamizar o processo de descrição geológica a partir dos dados da perfilagem geofísica convencional, os algoritmos utilizados foram selecionados, dentre a extensa variedade de algoritmos disponíveis, em função da sua simplicidade (poucos parâmetros ajustáveis), e baixa complexidade computacional (possibilidade de serem embarcados em campo). Foram eles os algoritmos *Naïve Bayes* e *Árvore de Decisão (Decision Tree)*.

2.3.1 *Naïve Bayes*

O modelo probabilístico *Naïve Bayes* se utiliza do teorema de *Bayes* considerando a premissa de independência condicional entre os atributos (variáveis), utilizados para predizer uma classe (atributo alvo) para uma dada instância (amostra) em um dado problema:

$$P\ (Y = C_k|\boldsymbol{X} = x) = \frac{P(Y = C_k)\ P(X = x|Y = C_k)}{P(X = x)}$$

Onde $x = (x_1, x_2, \dots, x_d)$ é a instância com *d* valores de um atributo (variável ou *feature*) *X* da base de dados. O objetivo do algoritmo é predizer um novo atributo *Y* (nesse caso também chamado de *target*), que pode assumir *K* valores possíveis, denotados $C_1,\ C_2, \dots, C_k)$. Sendo o problema de classificação binária quando *k=2* e multiclasse (ou multi-rótulos) quando *K>2*. Ou seja:

$$Posteriori = \frac{Priori\ \times\ Verossimilhança}{Evidência}$$

Como os atributos *X* são assumidos como aleatórios e condicionalmente independentes (daí o termo *Naïve* – ingênuo), para uma dada classe C_k, a verossimilhança $P(X = x|Y = C_k)$ pode ser reescrita, ficando a probabilidade posteriori:

$$P\ (Y = C_k|\boldsymbol{X} = x) = \frac{P(Y = C_k)\ \prod_{i=1}^{d} P(X_i = x_i|Y = C_k)}{P(X_1 = x_1, \dots, X_d = x_d)}$$

Para a classe C_ko denominador é constante, a probabilidade condicional pode ser escrita como proporcional ao numerador (na escala log para evitar *underflow*):

$$\log P\ (Y = C_k|\boldsymbol{X} = x) \propto \log P(Y = C_k) + \sum_{i=1}^{d} \log P(X_i = x_i|Y = C_k)$$

Logo a classe com a maior probabilidade posteriori é escolhida como a predição:

$$C = \arg \max_{k \in \{1,\dots,K\}} \left(\log P(Y = C_k) + \sum_{i=1}^{d} \log P(X_i = x_i|Y = C_k) \right)$$

O classificador *Naïve* Bayes possui poucas fragilidades, sendo as duas principais a premissa de independência condicional entre os atributos, e que estes atributos são categóricos/discretos. Na prática, conjuntos de dados multidimensionais apresentam atributos que possuem algum grau de correlação. A consequência disso é que variáveis correlacionadas acabam tendo o peso dobrado nos cálculos.

Com relação ao fato de por vezes os atributos serem contínuos é necessário modelar esses atributos por alguma função de probabilidade (assumir uma distribuição) e discretizar esses atributos. Ou seja, caso $X_i$seja um atributo contínuo, uma distribuição Gaussiana pode ser assumida:

$$P(X_i = x|Y = C_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left(-\frac{(x - \mu_{ik})^2}{2\sigma_{ik}^2} \right)$$

Onde μ_{ik} e σ_{ik}^2 são respectivamente a média e variância condicional da classe.

2.3.2 Árvore de Decisão (*Decision Tree*)

Árvores de Decisão são algoritmos de aprendizado supervisionado que utilizam, como o nome sugere, uma estrutura de árvore. Nessa estrutura existem dois tipos de *nodes:* decisão ou folha. A partir de um operador booleano o *node* do tipo decisão divide o conjunto de dados em dois ramos (*branches*) ao ser aplicado operador booleano a um dos atributos. O *node* tipo folha representa uma classe. O

processo de treinamento consiste em encontrar o melhor *split* em um dado atributo para um certo valor. O processo de predição consiste em atingir um *node* do tipo folha a partir da raiz (primeiro *node* do tipo decisão) aplicando-se os operadores booleanos em cada *node* do tipo decisão pelo caminho até o *node* tipo folha em questão.

Para avaliar a qualidade do melhor *split* que um *node* do tipo decisão pode executar é necessário definir o conceito (ou métrica) de ordenação/pureza, são elas *Gini* e *Entropia*, onde:

$$Gini = \ G_i = 1 - \sum_{k=1}^{n} p_{ik}^2$$

$$Entropia = \ H_i = -\sum_{k=1}^{n} p_{ik} \log\ (p_{ik})\ , p_{ik} \neq 0$$

Onde p_{ik} é a proporção de instâncias da classe *k* dentre todas as instâncias no i-ésimo *node*. Logo, melhor *split* significa que após a aplicação do operador booleano, os dois novos ramos apresentam menor impureza (ou maior ordenação).

A estrutura de uma árvore de decisão é muito semelhante a um fluxograma. As árvores de decisão são algoritmos de aprendizado por indução baseados em exemplos práticos, que conseguem deduzir as regras de classificação em forma de árvore a partir de um conjunto de exemplos desordenados que não seguem nenhuma regra.

Como não é considerada nenhuma premissa sobre o a estrutura da base de dados, ou mesmo as relações de dependência entre os atributos, as árvores de decisão podem se ajustar facilmente aos dados quando não são estabelecidos vínculos. Podendo ocasionar o *overfitting* e perda da capacidade de generalização. Esse problema é contornado tipicamente com o ajuste (ou regularização) dos hiperparâmetros do algoritmo durante a etapa de treinamento. Os hiperparâmetros frequentemente observados nas diferentes aplicações que possuem árvores de decisão

são: profundidade máxima da árvore, mínimo de amostras para split, mínimo de amostras na folha, máximo de nodes tipo folha, dentre outros.

3 Processamento

São apresentadas na sequência as etapas de tratamento e preparação dos dados, bem como as definições, os critérios, algoritmos e escolhas feitas em cada umas dessas etapas para a estimativa dos modelos de classificação de litotipos a partir de dados de perfilagem geofísica, sob a abordagem de aprendizado supervisionado.

3.1 Preparação da Base de Dados

Os dados utilizados nesse trabalho foram cedidos pela Vale S.A., e correspondem à perfilagem geofísica e descrição geológica de 5 furos de sondagem exploratória (SSD-FD00995, SSD-FD00998, SSD-FD01001, SSD-FD01006 e SSD-FD01038), totalizando pouco mais de 2000 m de investigação. Os arquivos contendo os dados de perfilagem foram disponibilizados em formato *.LAS (*Log ASCII Standard)*. Já os dados de descrição foram extraídos do banco de dados de Geologia em tabelas no formato *XLSX.

Em função da sensibilidade, bem como do caráter estratégico das informações contidas nesses dados, toda referência à localização dos mesmos foi removida.

Os dados da descrição geológica são mostrados na Tabela 1, onde pode ser observado um trecho dos primeiros intervalos da descrição geológica, os dados fornecidos estão sujeitos a cláusulas de confidencialidade, sendo as informações sensíveis suprimidas da tabela. Neste trabalho apenas o atributo Classificação Visual (CLV), descrito por intervalo, foi utilizado.

Tabela 1 – Exemplo de trecho da tabela de dados da descrição geológica.

Furo	De	Até	CLV	Litotipos
SSD-FD01038	0	4	AT	Aterro
SSD-FD01038	4	8	CG	Canga
SSD-FD01038	8	182.2	MD	Máfica Decomposta
SSD-FD01038	182.2	195.6	MSD	Máfica Semi Decomposta
SSD-FD01038	195.6	203.4	MS	Máfica Sã
SSD-FD01038	203.4	217.45	MSD	Máfica Semi Decomposta
...	...	...	...	...

A Tabela 2 mostra o resumo da classificação visual dos litotipos, realizada durando o processo de descrição geológica dos testemunhos de sondagem, consolidado para todos os 5 furos utilizados no estudo. Ao longo da dissertação serão utilizadas frequentemente menções as CLVs, ao invés da descrição completa.

Tabela 2 – Resumo da Classificação Visual (CLV).

CLV	DESCRIÇÃO
AT	Aterro
CG	Canga
MD	Máfica Decomposta
MSD	Máfica Semi Decomposta
MS	Máfica Sã
JPC	Jaspelito Compacto
JPF	Jaspelito Friável
HF	Hematitito Friável
RIF	Riolito Friável
HGOF	Hematitito Goethítico Friável
SR	Sem Recuperação
HCTF	Hematitito Contaminado Friável
RIS	Riolito Semi Compacto
JPS	Jaspelito Semi Compacto
RIC	Riolito Compacto
HC	Hematitito Compacto
RCF	Rocha Intrusiva Ácida Friável
AT	Aterro

A Tabela 3 mostra a relação de ferramentas (e suas descrições) que compõe a base de dados de perfilagem utilizada nesse projeto. Nessa mesma tabela é apresentada a corrida de cada ferramenta por furo.

Tabela 3 – Relação dos furos disponíveis na base de dados de perfilagem, das ferramentas perfiladas em cada furo e a descrição de cada curva perfilada por ferramenta.

Ferramenta	Curva	Unidade	Descrição da Curva	SSD-FD01038	SSD-FD01001	SSD-FD01006	SSD-FD00998	SSD-FD00995
DD6	CADE	.mm	Caliper from DD6	✓	✓	✓	✓	✓
	DENL	.g/c3	Density Long Spaced	✓	✓	✓	✓	✓
	DD3L	.cps	Density Long Spaced Raw	✓	✓	✓	✓	✓
	DNLO	.g/c3	Density Long Spaced OPEN	✓	✓	✓	✓	✓
	DENB	.g/c3	Density Short Spaced	✓	✓	✓	✓	✓
	DNBO	.g/c3	Density Short Spaced OPE	✓	✓	✓	✓	✓
	DD3B	.cps	DensityShort Spaced Raw	✓	✓	✓	✓	✓
	GRDE	.api	Gamma Ray from DD6	✓	✓	✓	✓	✓
	GRDO	.api	Gamma Ray from DD6 Open	✓	✓	✓	✓	✓
GC2	CCO1	.mm	Caliper 3-Arm CO1-GC2	✓	✓	✓	✓	✓
	CO1C	.cps	Caliper Raw	✓	✓	✓	✓	✓
	DD3C	.cps	Caliper Raw	✓	✓	✓	✓	✓
	GRC1	.gapi	Gamma Ray from GC1-GC2	✓	✓	✓	✓	✓
	DD3G	.cps	Gamma Ray Raw	✓	✓	✓	✓	✓
	GC1G	.cps	Gamma Ray Raw	✓	✓	✓	✓	✓
GTMP	DEPT	.m	Logged depth	✓	✓	✓	✓	✓
	MSUS	-	Magnetic Susc. 1	✗	✗	✓	✓	✓
	MSS4	-	Magnetic Susc. 4	✓	✗	✗	✗	✗
	FE2	.ohm.m	Resistivity Deep	✓	✗	✗	✗	✗
	FE1	.ohm.m	Resistivity Shallow	✓	✗	✗	✗	✗
	GTMP	.degc	Temperature	✓	✓	✓	✓	✓
	CCLF	.cps	Casing Collar Locator	✓	✓	✓	✓	✓
-	BIT	mm	Bit size	✓	✓	✓	✓	✓

Para dinamizar o tratamento e análise os dados de perfilagem e da descrição geológica foram consolidados em uma única base, incluindo-se todos os furos. Um trecho dessa base é exibido na Tabela 4, nela as linhas representam as amostras, ou seja pontos de medidas ao longo dos furos, e nas colunas os atributos, ou propriedades físicas medidas pela perfilagem com a CLV correspondente para cada amostra.

Tabela 4 - Trecho da base de dados consolidada dos dados de descrição geológica (CLV) e perfilagem geofísica

CLV	DEPTH	FURO	CADE	GRDE	DD3L	DD3B	DENB	DENL	GRC1	···
AT	-3.84	SSD-FD01038	77.622	46.381	19	653	3.658	3.188	44.818	···
AT	-3.85	SSD-FD01038	77.636	42.615	20	588	3.646	3.198	41.714	···
AT	-3.86	SSD-FD01038	77.669	44.598	10	593	3.645	3.208	41.52	···
AT	-3.87	SSD-FD01038	77.712	44.796	20	581	3.65	3.213	38.803	···
AT	-3.88	SSD-FD01038	77.717	52.922	10	650	3.648	3.218	50.056	···
AT	-3.89	SSD-FD01038	77.675	50.94	29	640	3.637	3.213	52.773	···
AT	-3.9	SSD-FD01038	77.664	52.922	20	660	3.632	3.213	55.489	···
AT	-3.91	SSD-FD01038	77.675	51.139	19	632	3.618	3.214	52.579	···
AT	-3.92	SSD-FD01038	77.633	51.337	19	614	3.608	3.211	49.862	···
AT	-3.93	SSD-FD01038	77.605	57.283	20	660	3.592	3.21	55.489	···
AT	-3.94	SSD-FD01038	77.613	59.265	10	663	3.588	3.205	58.399	···
AT	-3.95	SSD-FD01038	77.627	59.463	19	687	3.584	3.193	61.309	···
AT	-3.96	SSD-FD01038	77.6	59.463	20	677	3.581	3.178	55.877	···
...	...	...	...	...	...	...	...	...	...	...

Os Striplogs (Figura 12 a Figura 16) são a representação gráfica da base de dados consolidada. São compostos pelas curvas da perfilagem geofísica convencional e pelo perfil da descrição para cada um dos furos.

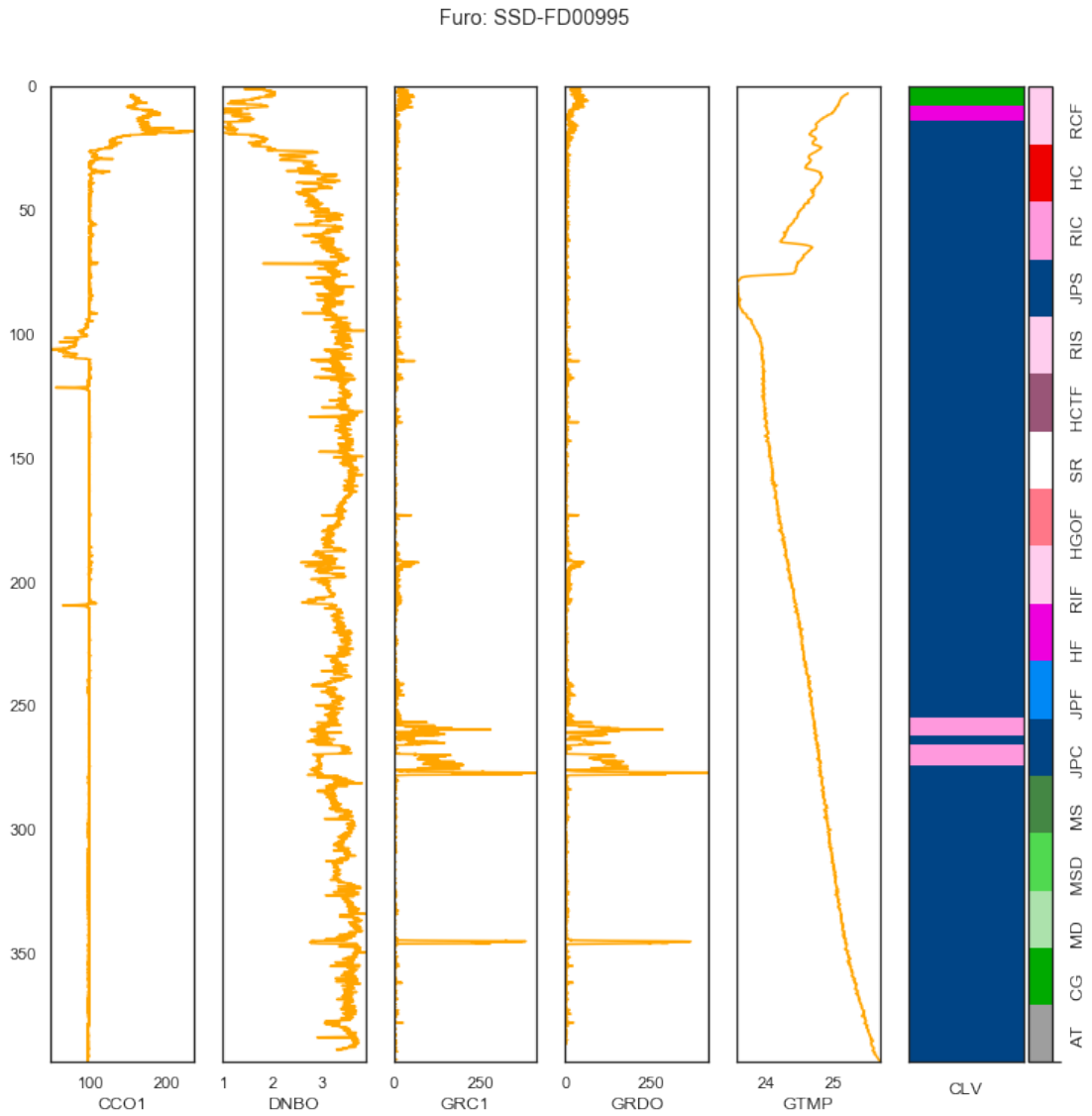


Figura 12 - Striplog da descrição geológica (CLV) juntamente com as curvas da perfilagem geofísica convencional para o Furo SSD-FD00995. Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6), Temperatura (GTMP).

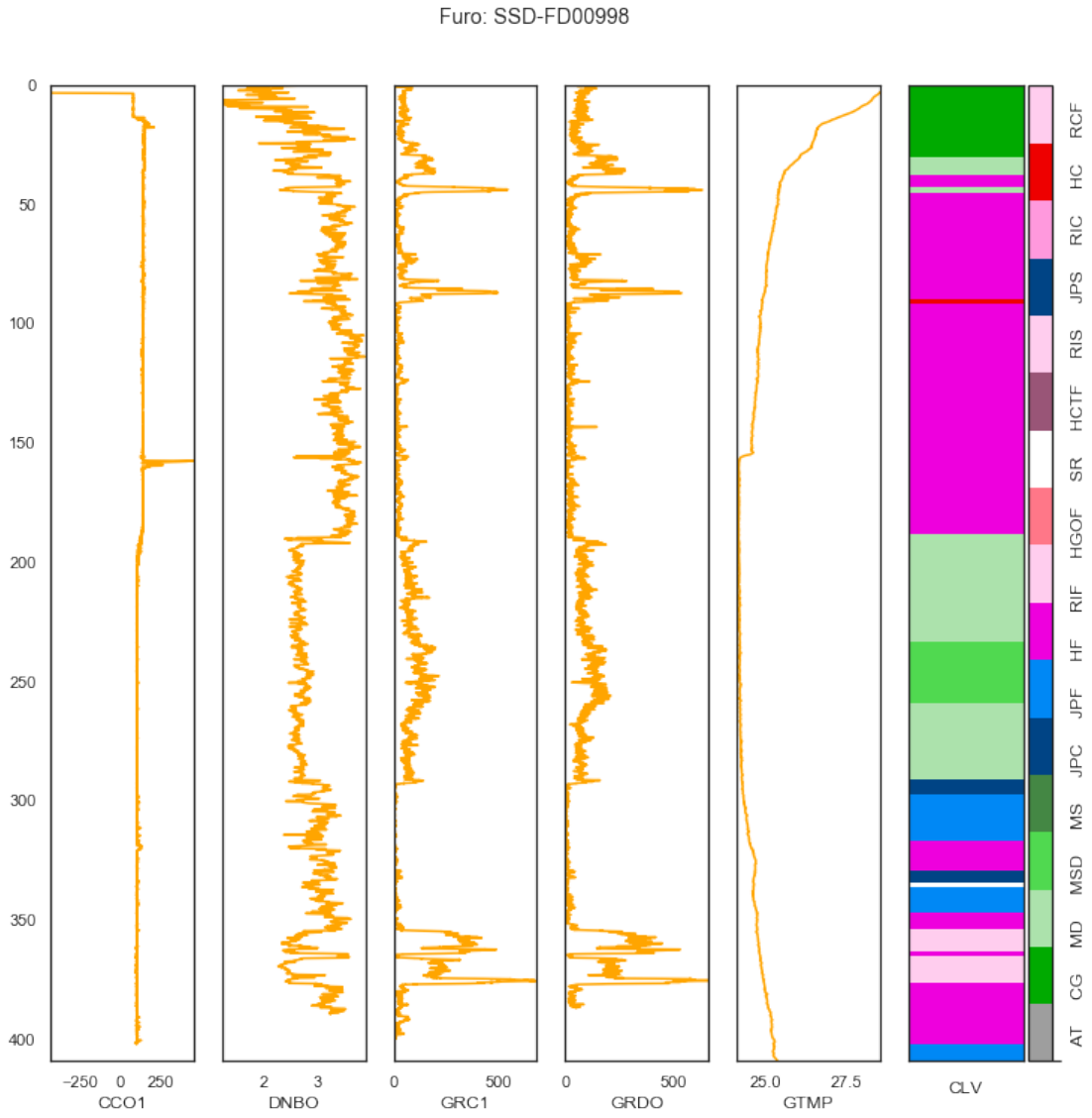


Figura 13 - Striplog da descrição geológica (CLV) juntamente com as curvas da perfilagem geofísica convencional para o Furo SSD-FD00998. Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6), Temperatura (GTMP).

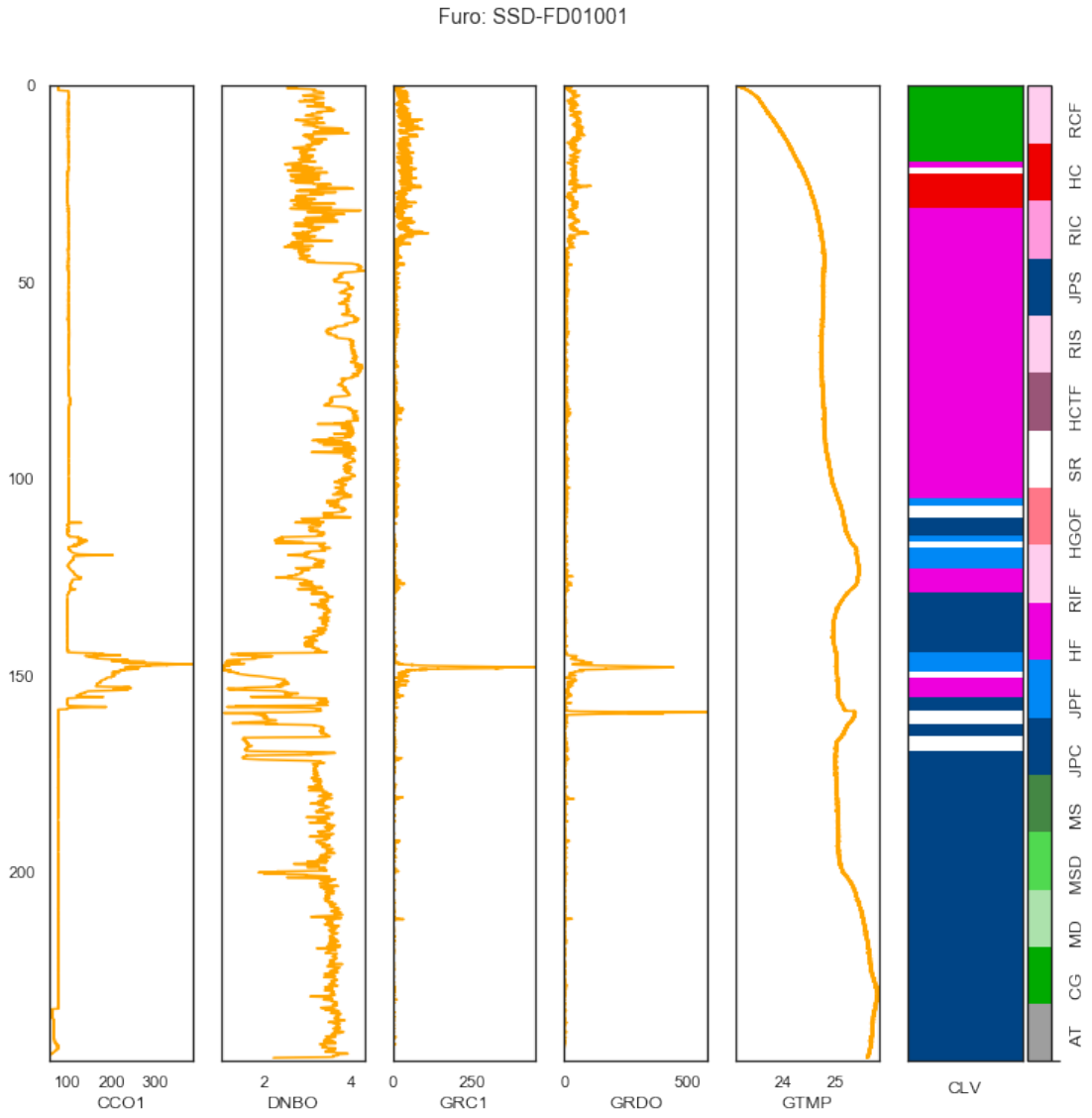


Figura 14 - Striplog da descrição geológica (CLV) juntamente com as curvas da perfilagem geofísica convencional para o Furo SSD-FD01001. Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6), Temperatura (GTMP).

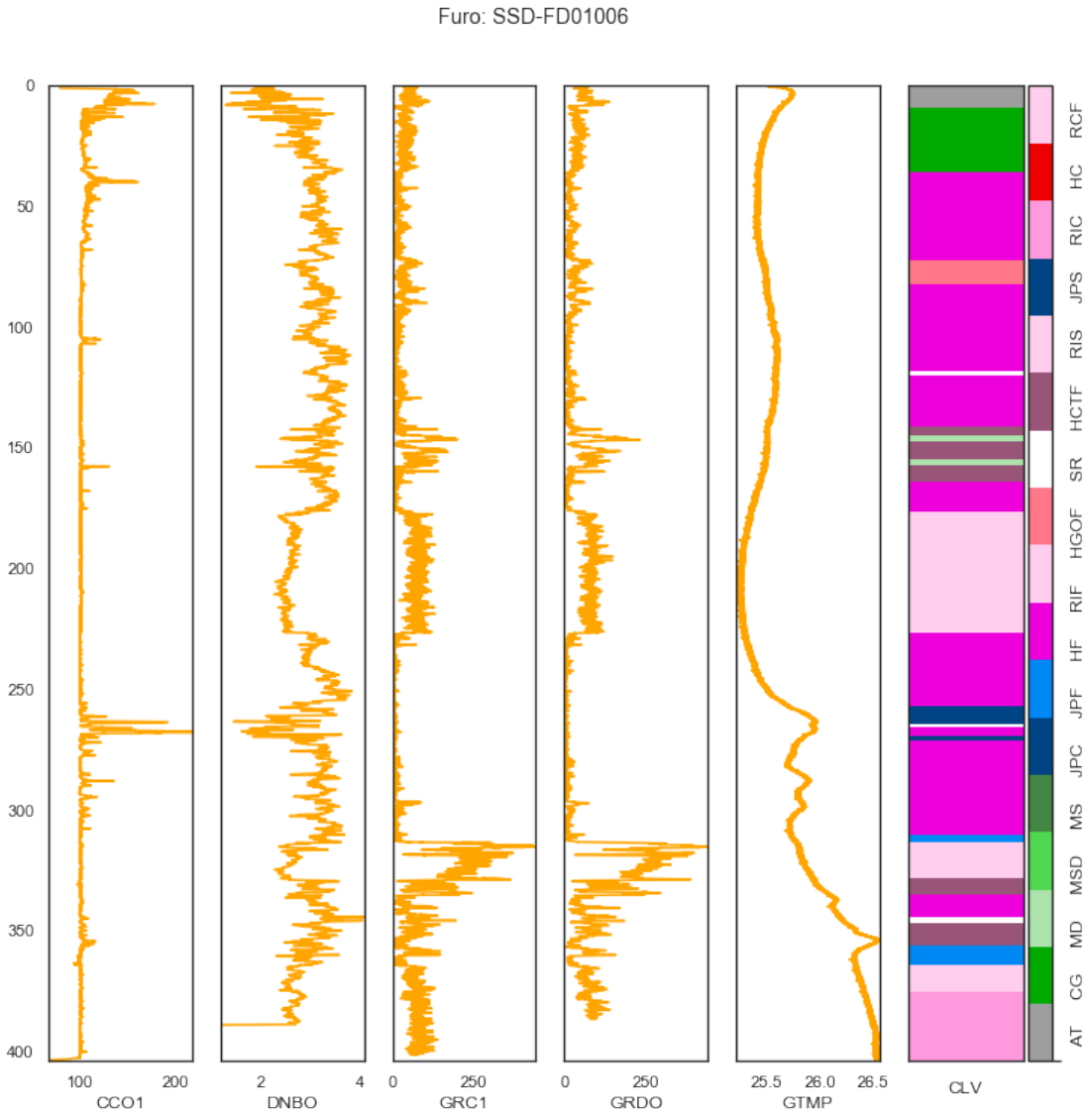


Figura 15 - Striplog da descrição geológica (CLV) juntamente com as curvas da perfilagem geofísica convencional para o Furo SSD-FD01006. Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6), Temperatura (GTMP).

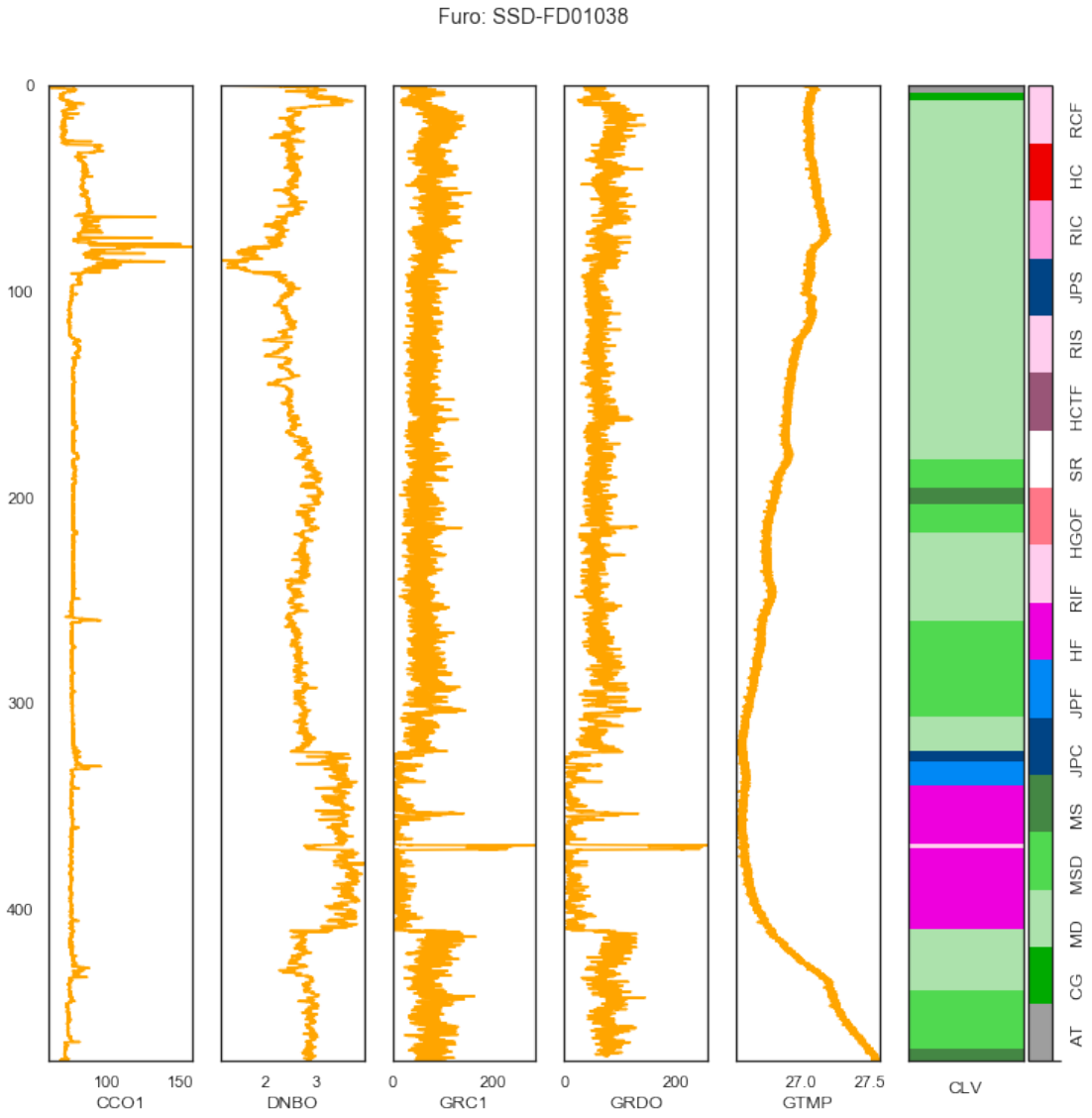


Figura 16 - Striplog da descrição geológica (CLV) juntamente com as curvas da perfilagem geofísica convencional para o Furo SSD-FD01038. Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6), Temperatura (GTMP).

3.2 Análise Exploratória dos Dados

A análise exploratória busca, através dos dados, acessar as informações sobre o domínio do problema estudado, apoiando-se na sua descrição quantitativa e qualitativa. Possibilita assim o aumento do conhecimento e de *insights* necessários para a escolha dos diferentes algoritmos (e seus parâmetros) a serem utilizados nas demais etapas do processo de determinação do modelo de classificação de litotipos.

Como a perfilagem geofísica teve intervalo de amostragem de 1cm, a descrição geológica, que possui um suporte amostral maior, foi reamostrada para este mesmo intervalo. A Figura 17 mostra os furos com os intervalos geológicos descritos pela CLV.

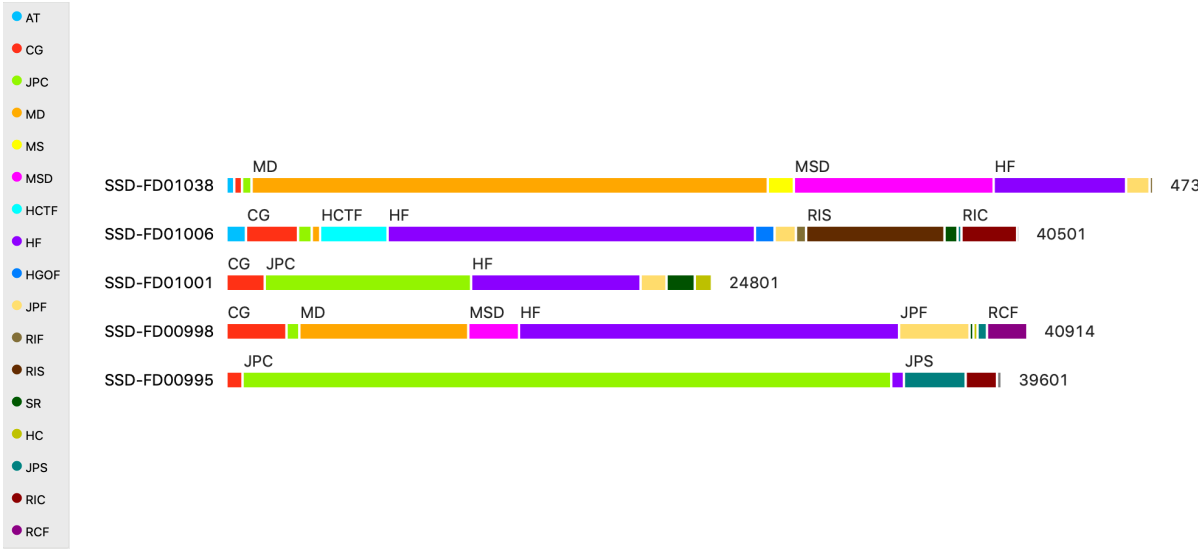


Figura 17 – Furos e descrição geológica (CLV)

3.2.1 Estatística Descritiva da Base de Dados

Na descrição estatística dos dados foram utilizadas as bibliotecas *python: Pandas* (McKinney, 2010) e de programação visual *Orange* (Demsar et al., 2013).

A Tabela 5 apresenta a estatística descritiva da base de dados. Para os atributos categóricos foi contabilizado o número de classes (*únicos*), a classe mais frequente

(*moda e*) e número de instâncias ausentes (*# ausentes*) e seu respectivo percentual (*% ausentes*).

Para os atributos numéricos (Tabela 5, inferior), foram determinadas as métricas estatísticas (Borradaile, 2003; Sauter, 2002) de tendência central (*mean* e *50%*), de dispersão desvio padrão (*std*) e os quartis *25%* e *75%*, valores máximos e mínimos (respectivamente *max* e *min*), bem como foram contabilizadas as instâncias ausentes (*# ausentes*) e calculado o seu percentual (*% ausentes*).

A leitura da Tabela 5 mostra que dos 17 litotipos presentes na base de dados há uma maior frequência de ocorrência do Hematitito Friável (HF), bem como o furo SSD-FD01038 sendo o de maior representatividade numérica.

Tabela 5 - Estatística descritiva dos atributos categóricos (superior) e numéricos (inferior)

Atributo	contagem	únicos	moda	# ausentes	% ausentes
CLV	192805	17	HF	352	0.18
FURO	193157	5	SSD-FD01038	0	0

Atributo	contagem	mean	std	min	25%	50%	75%	max	# ausentes	% ausente
BIT	193157	80.03	4.40	77.80	77.80	77.80	77.80	88.70	0.00	0.00
CADE	188480	78.30	0.81	53.38	77.81	78.05	78.25	80.17	4677.00	2.42
CCLF	191467	-0.02	7.84	-399.99	-0.71	0.01	0.75	460.39	1690.00	0.87
CCO1	191684	97.76	31.05	-430.15	77.92	99.08	101.65	446.52	1473.00	0.76
CO1C	191671	1596.47	135.94	0.00	1527.00	1559.00	1677.00	3702.00	1486.00	0.77
DD3B	188807	616.06	279.44	0.00	409.00	540.00	770.00	2298.00	4350.00	2.25
DD3C	188465	1678.16	15.19	1178.00	1668.00	1673.00	1692.00	1726.00	4692.00	2.43
DD3G	187540	14.14	22.49	0.00	0.00	10.00	20.00	812.00	5617.00	2.91
DD3L	188750	34.71	119.28	0.00	10.00	10.00	30.00	3596.00	4407.00	2.28
DENB	188498	3.71	0.46	1.96	3.36	3.76	4.06	5.59	4659.00	2.41
DENL	188505	4.95	2.37	0.25	2.92	3.58	8.00	8.00	4652.00	2.41
DNBO	188224	3.03	0.50	0.98	2.67	3.10	3.42	4.31	4933.00	2.55
DNLO	172446	3.15	0.67	0.48	2.67	3.11	3.66	5.64	20711.00	10.72
FE1	42219	2232.12	9177.44	11.15	147.25	281.21	430.17	46681.90	150938.00	78.14
FE2	42219	2099.51	7463.18	12.33	207.31	404.87	927.93	38094.60	150938.00	78.14
GC1G	191026	18.97	31.41	0.00	0.00	0.00	28.00	505.00	2131.00	1.10
GRC1	191059	40.90	54.48	0.00	4.81	22.31	62.86	678.40	2098.00	1.09
GRDE	187565	31.39	42.86	0.00	3.88	16.34	47.55	503.30	5592.00	2.90
GRDO	187495	42.73	57.91	0.00	4.55	22.73	64.84	663.76	5662.00	2.93
GTMP	192311	25.48	1.01	23.12	24.71	25.33	26.56	28.53	846.00	0.44
MSS4	27507	32.26	154.29	0.00	14.38	19.58	24.98	2209.38	165650.00	85.76
MSUS	100419	817.70	1253.04	0.00	44.83	189.05	884.01	4986.64	92738.00	48.01
DEPTH	193157	-200.32	121.71	-473.39	-300.10	-193.15	-96.57	0.00	0.00	0.00

3.2.2 Distribuição de Litotipos (Classes) por Furo

Outra forma de se visualizar a base de dados é através do gráfico de frequência de ocorrência dos litotipos (*CLV*), e de frequência de ocorrência da *CLV* por furo. Como pode ser observado na Figura 18, há um claro desbalanceamento na frequência de ocorrência dos litotipos. Por exemplo, os litotipos HF, JPC e MD apresentam frequência aproximadamente quatro vezes maior que o MSD, que é o litotipo subsequente de maior frequência.

Quando avaliamos a frequência de ocorrência dos litotipos por furo (Figura 19), observamos outra característica que, apesar de soar óbvia, merece ser destacada: a frequência de ocorrência dos diferentes litotipos ao longo dos furos não se mantém constante. Ou seja, nem sempre os furos de sondagem exploratória amostram os mesmos litotipos com a mesma frequência.

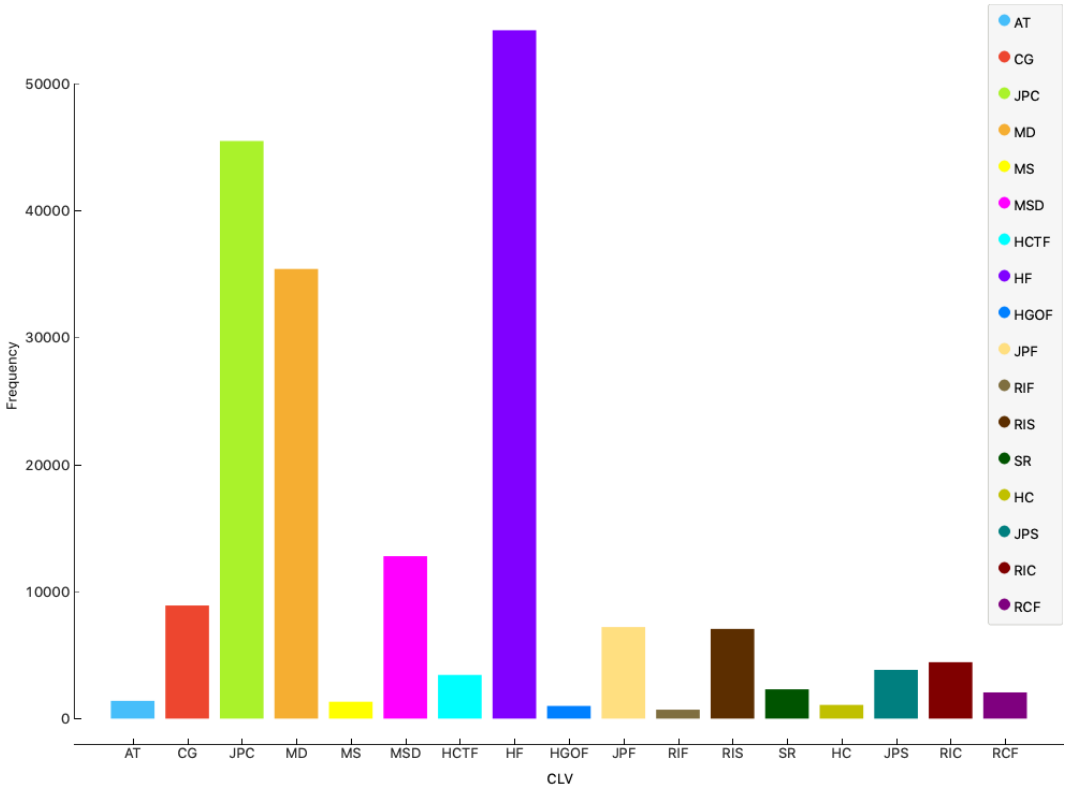


Figura 18 - Frequência dos litotipos (CLV)

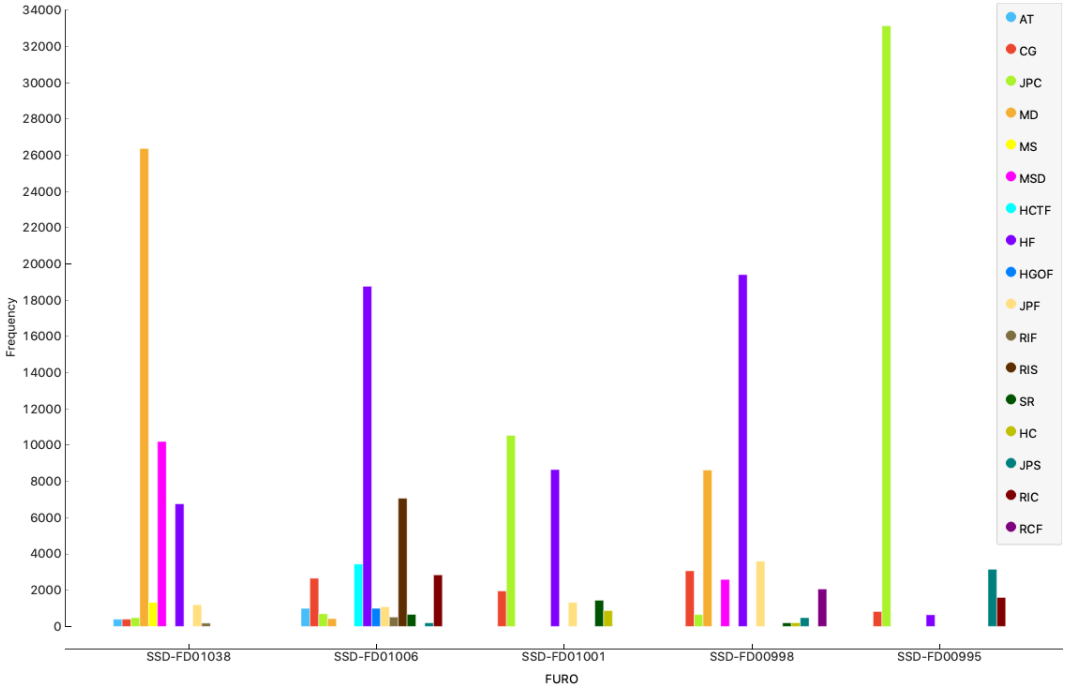


Figura 19 - Frequência de ocorrência dos litotipos (CLV) por Furo.

Com relação aos atributos numéricos, a Tabela 5 mostra que esses apresentam escalas diferentes nas medidas. Isso é evidenciado quando comparamos a ordem de grandeza dos valores das medidas de localização tendência central (*mean* e mediana *50%*) ou os valores máximos medidos (*max*). Por exemplo os atributos relacionados à densidade (*DENB* e *DENL*, etc.) apresentam valores médios da ordem de 10^1 enquanto os atributos relacionados à radiação gama apresentam valores médios da ordem de 10^3.

No que diz respeito ao percentual de instâncias ausentes, os atributos apresentam percentual de ausente inferior a 5% em sua grande maioria, salvas as exceções DNLO (10.72%), FE1 e FE2 (78.14%), MSS4 (85.76%) e MSUS (48.01%).

3.3 Pré-processamento dos dados

O tratamento dos dados visa aumentar a qualidade dos mesmos, mantendo as características originais intrínsecas descritas durante a exploração dos dados, e preparando-os para as etapas subsequentes de treinamento e validação. O pré-processamento dos dados, bem como as demais etapas de treinamento e validação, foi realizado com a aplicação de programação visual *Orange* (Demsar et

al., 2013). A Figura 20 é a visão geral do fluxo de processamento adotado neste trabalho.

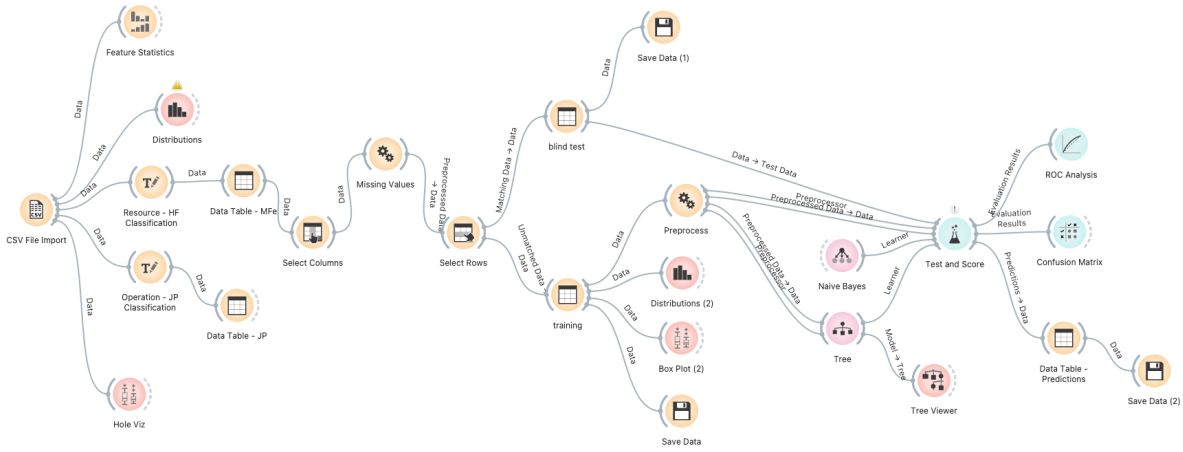


Figura 20 - Exemplo de fluxo de processamento no contexto de classificação supervisionada utilizando a aplicação de programação visual *Orange*.

3.3.1 *Separação do conjunto de dados – Treinamento e Validação (Teste-Cego)*

Com o objetivo de validar a aplicação do modelo de classificação de litotipos, simulando um regime de produção, um dos furos foi separado dos demais furos da base de dados. Esse furo será apresentado ao modelo de classificação em um teste-cego, o treinamento será realizado nos demais furos, ou seja, no conjunto de treinamento. Essa estratégia é uma forma de avaliar a eficácia de generalização do modelo.

A base de dados de treinamento consistiu nos furos: SSD-FD00995, SSD-FD00998, SSD-FD01006 e SSD-FD01038. Esse conjunto foi utilizado para ajustar/otimizar os parâmetros dos algoritmos de classificação em cada um dos problemas, bem como avaliar o desempenho, ou sucesso, na tarefa de predição das classes.

O furo SSD-FD01001 foi selecionado aleatoriamente para realização da validação via teste cego. O conjunto de dados teste (teste -cego), é utilizado para avaliar o sucesso final dos modelos de classificação.

3.3.2 Seleção de atributos e edição de domínios

A tarefa de seleção dos atributos tem o objetivo de preparar o conjunto de dados para a etapa subsequente de estimativa do modelo de classificação. Essa tarefa pode ser realizada utilizando-se informações *a priori*, tipicamente obtidas com os especialistas do domínio da base de dados, ou baseando-se em métodos matemáticos e estatísticos que elaboram um *ranking* dos atributos (Yu & Liu, 2003) pontuando-os, de acordo com diferentes critérios, e baseando-se na relação dos demais atributos com o atributo-alvo. Consequentemente, isso permite a remoção daqueles atributos que não contribuem de forma significativa para a classificação, por serem redundantes, por exemplo, promovendo uma diminuição do custo computacional, mas sem comprometer a qualidade final da classificação. Esse processo reduz a dimensionalidade do problema, uma vez que cada atributo da base de dados corresponde a uma dimensão da mesma.

A Figura 21 mostra a tela do *widget* de seleção de atributos (*features*) da aplicação *Orange* (Demsar et al., 2013). Nela podemos ver que os atributos DNLO, FE1, FE2, MSS4 e MSUS (devido ao percentual de ausentes), juntamente com os atributos BIT (diâmetro nominal do furo) e CCLF (contagem de hastes de perfuração) não foram utilizados na etapa de treinamento do modelo, ou seja o critério de seleção dos atributos foi baseado no conhecimento dos especialistas sobre cada um dos atributos bem como no percentual de dados faltantes.

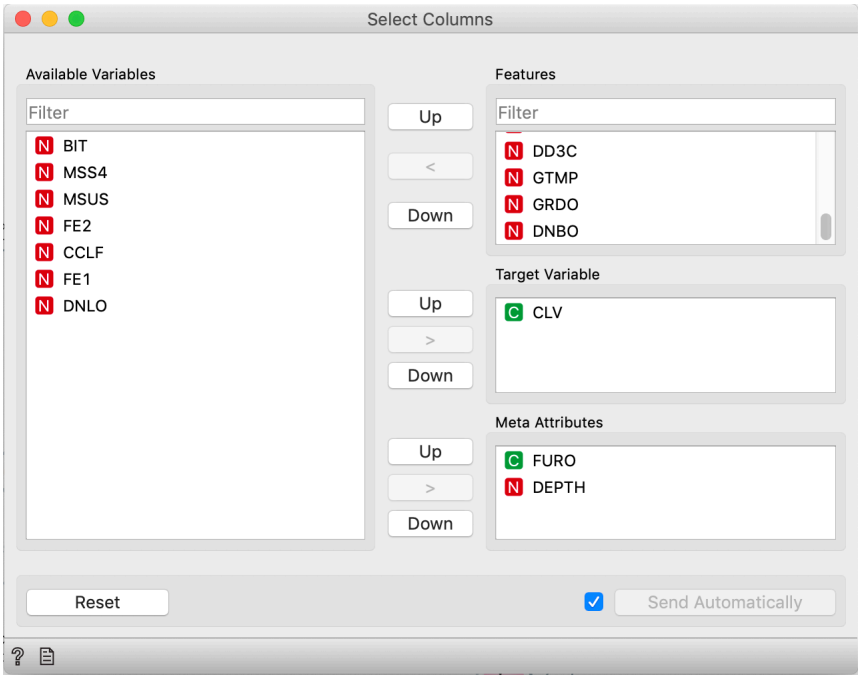


Figura 21 - Tela de seleção dos atributos da aplicação *Orange*

Nesse *widget* também é feito o apontamento do atributo alvo (*target variable*), neste trabalho o atributo alvo é CLV. Os atributos FURO e DEPTH são mantidos no conjunto de treinamento como *meta attributes*, ou seja, serão utilizados apenas para finalidade de visualização. Os demais atributos compõem a base de dados de treinamento.

Os atributos foram selecionados como segue:

- Features: GRDE, DENB, GRC1, CCO1, CO1C, GC1G, GTMP, GRDO, DNBO
- Meta attributes: FURO, DEPTH, BIT
- Target: CLV
- Removed: 12 (DD3G, DENL, DD3C, CADE, MSS4, DNLO, FE2, FE1, CCLF, DD3B, DD3L, MSUS)

A base de dados de treinamento, composta pelos demais furos, teve o atributo CLV (target) editado para novos domínios, em função dos problemas que serão modelados, conforme Tabela 6.

Tabela 6 - Edição dos domínios (target) para cada problema.

Classificação MFe	Classificação CLV
AT → other	AT
CG → other	CG
JPC → other	JPC
MD → other	MD
MS → other	MS
MSD → other	MSD
HCTF → MFe	HCTF
HF → MFe	HF
HGOF → MFe	HGOF
JPF → other	JPF
RIF → other	RIF
RIS → other	RIS
SR → other	SR
HC → MFe	HC
JPS → other	JPS
RIC → other	RIC
RCF → other	RCF

Dessa forma, para cada um dos problemas, o target CLV é composto pelas classes da Tabela 6.

3.3.3 Tratamento de valores ausentes

O tratamento de valores ausente acontece em duas dimensões – em nível dos atributos e em nível das instâncias – e consiste na imputação ou remoção do atributo e/ou instância. Nesse trabalho a escolha foi pela remoção baseada em um critério de limite percentual de valores ausentes. Essa escolha foi feita em função da complexidade elevada associada à realização da imputação dos valores ausentes de forma criteriosa, esse estudo por si só seria tema de um mestrado.

Os critérios de tolerância foram:

i. Atributos com percentual de ausentes superior a 10% foram removidos integralmente da base de dados na etapa de seleção de atributos (seção 3.3.2) e desconsiderados das demais etapas. São eles: DNLO, FE1, FE2, MSS4 e MSUS.
ii. Instâncias (linha da base de dados) com valores ausentes em qualquer um dos atributos restantes também foram removidas.

3.3.4 Adequação da escala de valores dos atributos - Normalização

Como foi observado durante a exploração da base de dados (seção 3.2.1), a perfilagem geofísica apresenta alguns atributos com ordens de grandeza distintas (*e.g.* GRDO e DNBO). Dessa forma, para minimizar a influência dessa característica, foi adotada a normalização *MinMax* (Basheer & Hajmeer, 2000), no intervalo [-1, 1], para toda a base de dados. Ou seja, os atributos têm os valores de suas instâncias normalizados para o intervalo acima.

Esses autores argumentam em sua revisão, sobre Redes Neurais Artificiais, que a técnica de normalização *MinMax*:

i. Previne que números grandes se sobreponham aos menores no processo de aprendizagem (treinamento);
ii. Impede a saturação prematura dos neurônios nas camadas ocultas, o que impediria o processo de aprendizagem.

3.4 Aprendizado Supervisionado e Validação

O objetivo dessa seção é apresentar, dentro do domínio das técnicas de aprendizado de máquina, alguns dos principais algoritmos de classificação (Géron, 2017), disponíveis nas bibliotecas *python Orange e Scikit-learn* (Pedregosa et al., 2011).

3.4.1 Treinamento por Validação Cruzada

O processo de validação cruzada (James et al., 2013; Stone, 1974) consiste em subdividir o conjunto de dados de treinamento novamente, com o objetivo auxiliar a capacidade de generalização do modelo de classificação durante o processo de treinamento do algoritmo. A capacidade de generalização do modelo é avaliada, qualitativamente, pela sua capacidade de predizer uma nova instância, que não tenha sido utilizada no seu processo de treinamento.

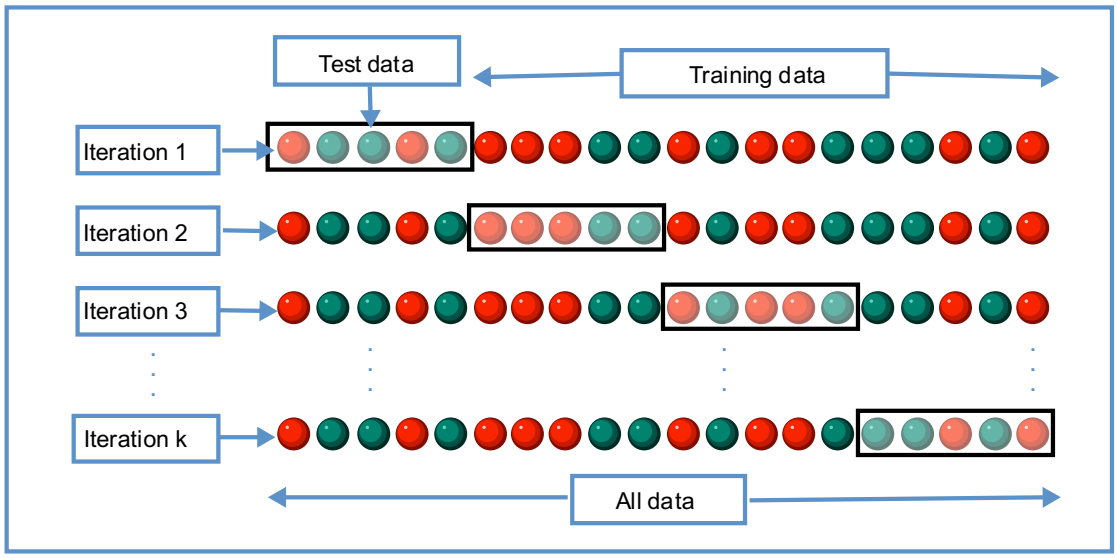


Figura 22 - Exemplificação do processo de validação cruzada. (fonte: By Gufosowa - Own work, CC BY-SA 4.0)

A Figura 22 ilustra o processo pelo qual passa essa nova divisão da base de dados de treino. Subdivide-se a base em um número fixo de trechos *chunks*, correspondente a k-*folds*, ou número de vezes (iterações) que o processo será repetido. A cada iteração do processo de validação cruzada um *chunk* é separado para teste e o treinamento é executado nos demais *chunks*. Ao término do processo, todos os *chunks* são percorridos como teste e treino, e a performance média de predição do modelo pode ser estimada.

A característica de desbalanceamento da frequência de ocorrência da CLV, identificada na exploração dos atributos categóricos, também deve ser levada em consideração sempre que seja necessário reamostrar a base de dados (Figura 23).

Figura 23 - Composição das cinco *folds* utilizadas do processo de validação cruzada. Nota-se que a frequência relativa de ocorrência da CLV é mantida em cada *fold*.

Para isso todo processo de reamostragem é feito de forma estratificada, ou seja, o conjunto reamostrado preserva as frequências relativas dos atributos, como pode ser visto na Figura 23. O processo de validação cruzada considerou 5 *folds* (*k=5*), mantendo-se a frequência relativa de ocorrência da CLV em cada *fold* (estratificação).

No processo de treinamento por validação cruzada também foi realizado a exploração dos hiperparâmetros do algoritmo *Decision Tree*, utilizando o método *gridSearch* do *Scikit-Learn*. Nesse método é criado um pipeline contendo os vetores com a lista de valores para cada parâmetro do modelo, e de forma recursiva o treinamento ocorre fazendo a combinação entre todos os parâmetros passados para o método.

3.4.2 Naive Bayes

O algoritmo estima um modelo classificador baseado do teorema de Bayes (Géron, 2017; James et al., 2013; Sauter, 2002), que calcula a probabilidade de um evento ocorrer dado o conhecimento a priori de condições que podem estar relacionadas a esse evento, com a premissa de independência dos atributos. Ou seja, o valor de um determinado atributo é independente do valor dessa mesma instância nos demais atributos.

Em sua implementação na aplicação de programação visual *Orange* (Demsar et al., 2013), esse método não possui hiperparâmetros que possam ser ajustados, entretanto ele necessita que as variáveis (atributos) contínuos sejam discretizados para o cálculo da probabilidade conjunta. Dessa forma é realizada a discretização dos atributos contínuos em *bins*, com critério de mesma frequência de ocorrência por *bin*, assumindo-se uma distribuição normal.

O fluxo de processamento com o método recebe como entrada a base de dados e os preprocessadores (parâmetros de saída do pré-processamento). As

saídas do método são o próprio algoritmo de aprendizado *Naïve Bayes* e o modelo classificador treinado na base de dados.

3.4.3 Decision Tree

O algoritmo estima um modelo classificador subdividindo progressivamente o conjunto de dados em grupos menores, baseado em atributos descritivos, ou seja, que explicam a variabilidade dos dados de forma significativa. Por esse motivo os modelos *Decision Tree* possuem forte característica de explicabilidade, pois permitem a compreensão de como o modelo faz a classificação de uma dada amostra. O algoritmo agrupa os dados em conjuntos por similaridade, e busca uma regra baseada naquele atributo que melhor explica essa divisão por similaridade. Para cada nova regra dois novos grupos agrupamentos são feitos.

Sua implementação na aplicação de programação visual *Orange* permite o ajuste dos seguintes hiperparâmetros:

- Induce binary tree: constrói uma arvore binária (divide os dados em dois nodes-filhos);
- Min. number of instances in leaves: O algoritmo impede a divisão em *branchs* que disponibilizariam instâncias de treinamento nas folhas em número menos que a referência;
- Do not split subsets smaller than: Impede a divisão dos conjuntos que resultariam em um subconjunto menor que o valor de referência;
- Limit the maximal tree depth: Limita a profundidade da árvore em número de níveis de *nodes;*
- Stop when majority reaches: Critério de parada do modelo em função de um percentual dos critérios de crescimento atingidos.

O fluxo de processamento com o método recebe como entrada a base de dados e os preprocessadores (parâmetros de saída do pré-processamento). As saídas do método são o algoritmo de aprendizado *Decision Tree* e o modelo

classificador treinado na base de dados. Os hiperparâmetros que apresentaram melhor desempenho no treinamento (seleção pelo método *gridSearch*) foram: *Induce binary trees* (sim); *min. number of instances in leaves* (2); *do not split subsets smaller than* (5); *limit maximal tree depth* (10) e *stop when majority reaches* (95%)

3.4.4 Métrica de performance dos modelos de classificação

Existem diversas formas de se avaliar a performance de um modelo no contexto de aprendizado de máquina supervisionado disponíveis, ficando a critério do usuário a escolha da métrica mais adequada para refletir a capacidade do modelo na resolução do problema estudado. Na sequência são apresentadas as métricas que foram utilizadas nesse trabalho (Caté et al., 2017; Géron, 2017; Pedregosa et al., 2011; Shi, 2014; Xie et al., 2018).

As definições abaixo são utilizadas de forma ampla na definição das métricas:

VP = Verdadeiro Positivo; VN = Verdadeiro Negativo;

FP = Falso Positivo; FN = Falso Negativo.

3.4.4.1 Matriz de Confusão

A matriz de confusão é uma forma de visualização da performance de um modelo de aprendizado em uma tarefa de classificação. A tabela que apresenta nas linhas o número (ou proporção relativa) de instâncias na classe correta e as colunas o número de instâncias na classe predita. A diagonal principal mostra o número de amostras preditas (classificadas) corretamente pelo modelo.

		Predição		
		Classe A	Classe B	
Verdade	Classe A	VN	FP	VN + FP
	Classe B	FN	VP	FN + VP
		VN + FN	FP + VP	

Figura 24 - Matriz de Confusão

3.4.4.2 Acurácia (CA)

A acurácia *CA – Classification Accuracy*), é a proporção de predições corretas feitas pelo modelo e pode ser calculada conforme a expressão:

$$CA = \frac{VP + VN}{VP + VN + FP + FN}$$

A acurácia é a métrica de performance mais intuitiva, por refletir o percentual de acerto de forma direta. Ela é tipicamente utilizada no processo de treinamento para demonstrar, em linhas gerais, como o modelo está evoluindo. O ponto de atenção aqui diz respeito ao número de FP e FN, que a depender do problema serão alvo de otimização.

3.4.4.3 Revocação (Recall)

A métrica Revocação é a proporção de verdadeiro positivos, calculada segundo a expressão:

$$Recall = \frac{VP}{VP + FN}$$

Essa métrica traduz a proporção de casos Verdadeiros Positivos realmente capturados pelo modelo. Problemas nos quais é necessário que o modelo seja sensível o suficiente para minimizar o número de Falso Negativos, ou seja, não recuperada as ocorrências de uma dada classe, essa métrica poderia ser utilizada para avaliar a performance.

3.4.4.4 Precisão (Precision)

A métrica precisão é a proporção de predições positivas feitas corretamente pelo modelo, calculada segundo a expressão:

$$Precision = \frac{VP}{VP + FP}$$

Logo um modelo que tenha boa precisão em dada tarefa significa que ele tem capacidade de predizer corretamente uma saída positiva, minimizando predições positivas incorretas (FP).

3.4.4.5 F1-Score

A métrica F1-Score é a média harmônica entre Precisão e Revocação, calculada pela expressão:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{VP}{VP + \frac{FN + FP}{2}}$$

A métrica F1 é muito útil quando é necessário comparar dois classificadores distintos, levando-se em consideração o número de FN e FP, ponderando pela precisão e revocação.

4 Resultados e Discussão

4.1 Caracterização Petrofísica dos Litotipos

A caracterização petrofísica consiste na determinação dos valores e características estatísticas, que serão descritivos para cada litotipo (CLV) presente na base de dados. Dos dados obtidos pela perfilagem convencional apenas as curvas contagem de radiação gama (GRDO) e Densidade (DNBO) são propriedades físicas dos materiais geológicos. As demais curvas refletem uma mistura de informações da geologia com o furo e com o processo de perfilagem. A base de dados conta com 189260 amostras (pontos de leitura), distribuídas nos 5 furos.

A determinação de valores de referência para os litotipos contribui diretamente no aumento do conhecimento geológico no contexto da exploração de minério de ferro em S11D, pois tais valores serão utilizados em outras fases de prospecção geológica e geofísica (*e.g.* Inversão geofísica de métodos potenciais e modelagem geológica) ou mesmo na modelagem geoestatística e estimativa de recursos.

Tais valores de referência são definidos em termos das suas medidas de tendência central (Média e Mediana) e de sua dispersão (desvio padrão, ou std). A Tabela 7 consolida essas medidas dos atributos Densidade (DNBO – g/cm^3) e Contagem Total (GRDO - cps) para cada litotipo (CLV).

Observa-se que tanto o hematitito quanto o jaspelito apresentam densidades (DNBO) muito próximas, ao redor de 3.33 g/cm^3 (diferença de 0.03 g/cm^3 na média). Com relação à contagem total, o hematitito (HF) apresenta média quatro vezes maior e mediana seis vezes maior com relação ao jaspelito (JPC).

A densidade média mínima corresponde a 2.36 g/cm^3 (Aterro - AT), e máxima de 3.35 g/cm^3 (Hematitito Friável). Com relação à contagem total, a menor média corresponde a 5.50 cps (Jaspelito Compacto – JPC) e a maior média a 279.66 cps (Rocha Intrusiva Ácida Friável - RCF).

Tabela 7 - Consolidação das medidas de tendência central (Média e Mediana) e dispersão (desvio Padrão - std) dos atributos petrofísicos para cada litotipo (CLV).

	DNBO			GRDO		
	mean	median	std	mean	median	std
CLV						
AT	2.36	2.34	0.44	60.68	58.39	22.01
CG	2.75	2.88	0.57	52.32	44.34	28.67
HC	3.09	3.08	0.29	41.27	37.03	28.44
HCTF	3.13	3.16	0.27	74.08	56.11	64.88
HF	3.35	3.38	0.40	22.05	12.93	39.96
HGOF	2.95	2.97	0.15	37.39	33.56	16.91
JPC	3.32	3.38	0.33	5.50	2.02	22.99
JPF	3.05	3.12	0.49	21.78	9.44	41.85
JPS	2.91	3.10	0.55	10.66	7.58	9.87
MD	2.54	2.56	0.27	74.80	68.08	40.25
MS	2.96	2.99	0.10	62.00	59.53	14.84
MSD	2.77	2.77	0.13	84.17	73.55	35.43
RCF	2.60	2.54	0.27	279.66	247.02	124.99
RIC	2.91	2.90	0.32	72.74	81.30	49.29
RIF	2.74	2.71	0.24	256.05	255.93	94.95
RIS	2.57	2.57	0.12	102.38	82.55	63.22

Além das medidas de tendência central e dispersão tabeladas, podemos fazer uso de análises gráficas para visualizar o comportamento das propriedades físicas dos litotipos. A Figura 25 mostra os Boxplots da petrofísica para cada um dos litotipos. Nela é possível observar que os litotipos de maior variabilidade de densidade (DNBO, painel superior), são em sua maioria friáveis ou provenientes de processos alteração/intemperismo (AT, CG, JPF, HF, HCTF), exceto o RIC (Riolito Compacto). Outra característica visível é a simetria da distribuição de densidade para a maioria dos litotipos. Exceções feita para MS e RIF (respectivamente Máfica Sã e Riolito Friável). Com relação a contagem total (GRDO, painel inferior) observa-se que as maiores variabilidade correspondem ao Riolito Friável (RIF) e à Rocha Intrusiva Ácida Friável (RCF). Os jaspelitos (JPC, JPF e JPS) e hemattitos (HF, HC, HGOF) apresentam sistematicamente baixos valores de contagem total (GRDO) e variabilidade.

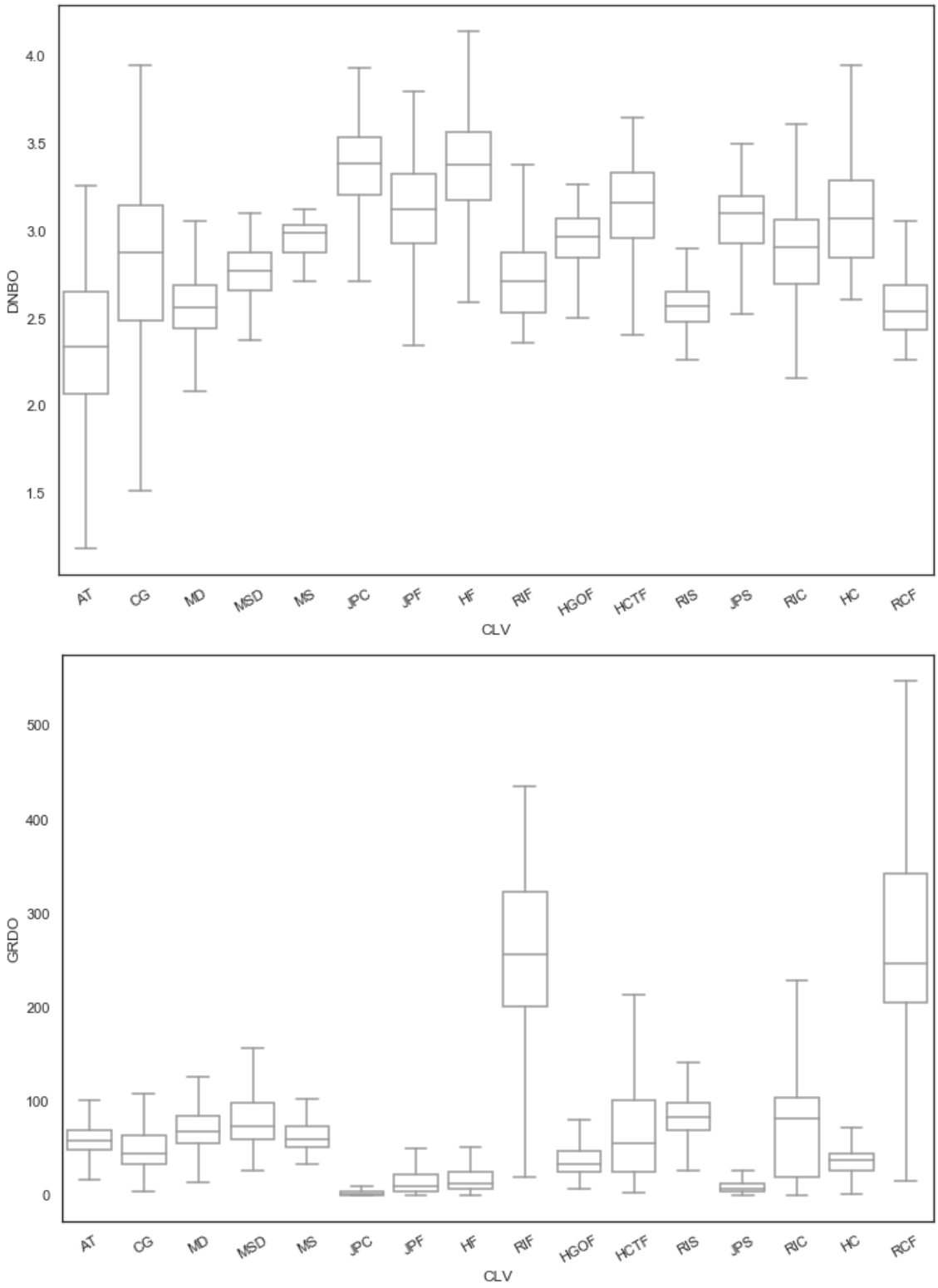


Figura 25 - Boxplots das propriedade físicas densidade (DNBO, em g/cm3, painel superior) e Contagem Total (GRDO, em cps, painel inferior) agrupados por Litotipo (CLV)

A análise gráfica de distribuição conjunta (Figura 26) foi utilizada para avaliar o comportamento das distribuições de densidade e contagem total para cada litotipo. Os litotipos JPC, JPF e HF apresentam comportamentos semelhantes de baixo

valor de GRDO e alto valor de DNBO e suas distribuições estão individualizadas. O litotipos AT, MSD, MS, HC, RIF, RIS e RCF apresentam duas populações em pelo menos um dos atributos. Os valores de referência para esses litotipos devem ser utilizados com atenção. Por exemplo, em um estudo de inversão geofísica esses valores podem refletir domínios distintos, ou mesmo incorporar incertezas e variância nos modelos se forem incorporados como um único domínio.

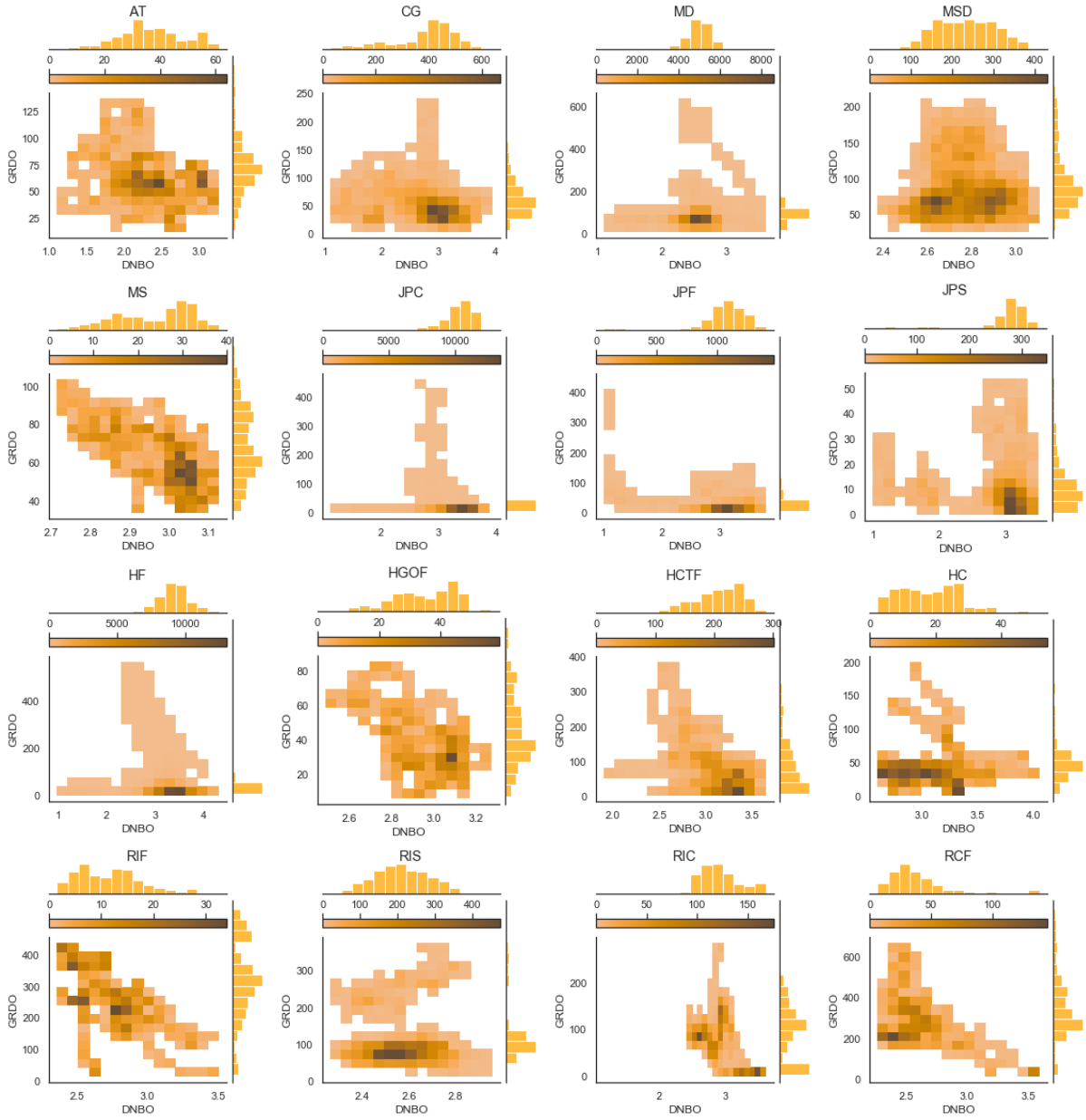


Figura 26 -Diagrama de distribuição conjunta da densidade (DNBO) e contagem total (GRDO). Nos eixos complementares estão as distribuições de cada um dos atributos. A escala de cor é baseada no número de amostras.

4.2 Aprendizado Supervisionado - Classificação de Litotipos (CLV)

A dinamização do processo de descrição geológica, ou mesmo a disponibilização de um modelo de descrição preliminar, baseado na perfilagem geofísica convencional (gama natural e densidade), por si só formalizou um problema de classificação multiclasse (ou multi-rótulos). Nesse tipo de problema, como não há uma classe alvo específica, a performance foi avaliada pela métrica F1.

4.2.1 Treinamento

Como explicado na seção 3.3.1, a base de dados de treinamento consistiu nos furos: SSD-FD00995, SSD-FD00998, SSD-FD01006 e SSD-FD01038. Esse conjunto foi utilizado para ajustar/otimizar os parâmetros dos algoritmos de classificação, no processo de treinamento, e também avaliar o desempenho na tarefa de predição das 16 classes a partir dos dados da perfilagem convencional.

A Tabela 8 mostra as métricas de performance média, sobre cada classe, obtidas durante o treinamento, considerando-se a validação cruzada (k=5), com estratificação dos dados (Seção 3.4.1). Podemos observar que o ambos os algoritmos apesentam capacidade de gerar modelos com bons resultados para todas as métricas (> 0.5500).

Tabela 8 - Performance do treinamento considerando validação cruzada (k=5). Classificação de Litotipos (CLV).

Model	CA	F1	Precision	Recall
Tree-Default	0.9172	0.9153	0.9164	0.9172
Naive Bayes	0.7061	0.7068	0.7187	0.7061

O modelo *Decision Tree* teve a melhor performance durante o treinamento (F1 = 0.9153) contra F1 = 0.7068 do modelo *Naïve Bayes*.

A matriz de confusão (Figura 27), mostra o detalhamento da performance através da proporção de predições corretas com relação à classe real. É observado na matriz que o modelo gerado pelo algoritmo *Decision Tree*, em linhas gerais, apresenta maior proporção de predições corretas para ambas as classes. Exceção feita

para a classe HC (hematitito compacto), que apesar do desempenho baixo (< 55%), o modelo *Naïve Bayes* teve melhor performance (17.6% de verdadeiros positivos contra 5.9% para o modelo *Decision Tree*).

Confusion matrix for Tree-Default (showing proportion of actual)

		Predicted																	
		AT	CG	JPC	MD	MS	MSD	HCTF	HF	HGOF	JPF	RIF	RIS	SR	HC	JPS	RIC	RCF	∑
Actual	AT	81.5 %	2.1 %	0.4 %	11.5 %	0.0 %	0.1 %	0.1 %	4.2 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	1365
	CG	0.9 %	83.9 %	0.0 %	1.9 %	0.0 %	0.2 %	0.4 %	9.3 %	3.4 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	6335
	JPC	0.0 %	0.0 %	96.1 %	0.4 %	0.0 %	0.0 %	0.0 %	1.4 %	0.0 %	0.6 %	0.0 %	0.0 %	0.1 %	0.0 %	0.7 %	0.7 %	0.0 %	34248
	MD	0.0 %	0.2 %	0.0 %	96.6 %	0.1 %	1.8 %	0.4 %	0.7 %	0.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	35427
	MS	0.0 %	0.0 %	0.0 %	0.2 %	85.0 %	14.8 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	1075
	MSD	0.0 %	0.0 %	0.0 %	15.1 %	1.8 %	83.0 %	0.0 %	0.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	12800
	HCTF	0.3 %	2.0 %	0.1 %	2.5 %	0.0 %	0.0 %	75.4 %	15.7 %	2.0 %	0.4 %	0.1 %	1.5 %	0.1 %	0.0 %	0.0 %	0.0 %	0.0 %	3443
	HF	0.0 %	0.5 %	1.3 %	0.1 %	0.0 %	0.0 %	0.8 %	95.3 %	0.3 %	1.1 %	0.0 %	0.1 %	0.1 %	0.0 %	0.1 %	0.1 %	0.2 %	43950
	HGOF	0.0 %	4.7 %	0.0 %	0.0 %	0.0 %	0.0 %	2.1 %	7.0 %	86.2 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	1000
	JPF	0.0 %	0.0 %	9.3 %	0.1 %	0.0 %	0.0 %	2.7 %	15.3 %	0.0 %	66.8 %	0.2 %	2.8 %	0.3 %	0.0 %	2.2 %	0.3 %	0.0 %	5230
	RIF	0.0 %	0.0 %	0.1 %	1.1 %	0.0 %	0.0 %	0.1 %	4.8 %	0.0 %	1.0 %	78.0 %	14.8 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	705
	RIS	0.0 %	0.2 %	0.0 %	0.1 %	0.0 %	0.0 %	0.7 %	0.2 %	0.0 %	0.5 %	0.6 %	97.5 %	0.0 %	0.0 %	0.0 %	0.3 %	0.0 %	7070
	SR	0.0 %	1.5 %	14.6 %	0.0 %	0.0 %	0.0 %	0.6 %	32.3 %	0.0 %	3.2 %	0.0 %	0.0 %	47.4 %	0.0 %	0.5 %	0.0 %	0.0 %	865
	HC	0.0 %	0.0 %	0.0 %	0.5 %	0.0 %	0.0 %	0.0 %	93.7 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	5.9 %	0.0 %	0.0 %	0.0 %	205
	JPS	0.2 %	0.0 %	7.2 %	0.7 %	0.0 %	0.0 %	0.0 %	6.8 %	0.0 %	8.8 %	0.0 %	0.0 %	0.3 %	0.0 %	75.7 %	0.2 %	0.0 %	3538
	RIC	0.0 %	0.1 %	16.9 %	0.3 %	0.0 %	0.1 %	0.5 %	0.3 %	0.0 %	0.1 %	0.0 %	2.0 %	0.0 %	0.0 %	0.0 %	78.8 %	0.8 %	2699
	RCF	0.0 %	0.0 %	1.6 %	0.0 %	0.0 %	0.0 %	0.0 %	2.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	1.1 %	95.3 %	2065
	∑	1209	5774	34872	36766	1175	11448	3343	45504	1305	4578	624	7294	498	14	3065	2465	2086	162020

Confusion matrix for Naive Bayes (showing proportion of actual)

		Predicted																	
		AT	CG	JPC	MD	MS	MSD	HCTF	HF	HGOF	JPF	RIF	RIS	SR	HC	JPS	RIC	RCF	∑
Actual	AT	31.5 %	11.3 %	0.0 %	15.2 %	7.3 %	11.2 %	0.1 %	4.5 %	1.0 %	0.0 %	0.0 %	17.9 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	1365
	CG	3.4 %	42.6 %	0.0 %	18.1 %	2.1 %	3.3 %	2.2 %	16.7 %	6.0 %	0.0 %	3.5 %	2.3 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	6335
	JPC	0.0 %	0.0 %	85.0 %	0.4 %	0.0 %	0.2 %	0.0 %	4.1 %	0.0 %	2.5 %	0.0 %	0.0 %	0.1 %	0.0 %	5.7 %	0.9 %	1.0 %	34248
	MD	0.2 %	0.3 %	0.0 %	71.4 %	0.4 %	22.7 %	0.6 %	0.9 %	0.0 %	0.1 %	0.5 %	2.5 %	0.1 %	0.1 %	0.0 %	0.1 %	0.0 %	35427
	MS	0.0 %	0.0 %	0.0 %	1.6 %	35.2 %	62.3 %	0.0 %	0.9 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	1075
	MSD	0.0 %	0.1 %	0.0 %	25.8 %	1.6 %	72.3 %	0.0 %	0.1 %	0.0 %	0.0 %	0.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	12800
	HCTF	0.1 %	5.0 %	0.0 %	0.3 %	0.0 %	0.0 %	34.3 %	33.6 %	4.0 %	0.4 %	11.1 %	10.3 %	0.0 %	0.0 %	0.0 %	1.0 %	0.0 %	3443
	HF	0.0 %	2.9 %	5.7 %	0.3 %	0.1 %	0.1 %	2.4 %	78.7 %	1.2 %	3.2 %	0.3 %	1.0 %	0.5 %	1.8 %	0.8 %	1.0 %	0.3 %	43950
	HGOF	1.4 %	17.7 %	0.0 %	1.6 %	0.0 %	0.0 %	4.3 %	29.4 %	38.4 %	1.4 %	0.6 %	4.9 %	0.3 %	0.0 %	0.0 %	0.0 %	0.0 %	1000
	JPF	0.0 %	0.5 %	23.0 %	0.2 %	0.0 %	0.0 %	3.7 %	36.2 %	2.8 %	20.4 %	0.8 %	2.2 %	0.0 %	0.0 %	10.2 %	0.0 %	0.0 %	5230
	RIF	0.0 %	0.7 %	0.0 %	0.4 %	0.0 %	20.1 %	7.7 %	2.6 %	0.0 %	0.0 %	29.4 %	39.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	705
	RIS	0.3 %	0.8 %	0.0 %	0.8 %	0.0 %	0.0 %	0.3 %	0.2 %	3.0 %	0.1 %	4.4 %	90.2 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	7070
	SR	0.0 %	2.9 %	10.3 %	0.0 %	0.0 %	0.0 %	24.2 %	40.5 %	2.8 %	12.6 %	0.2 %	0.8 %	2.0 %	0.0 %	3.8 %	0.0 %	0.0 %	865
	HC	0.0 %	8.8 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	58.5 %	0.0 %	0.0 %	0.0 %	15.1 %	0.0 %	17.6 %	0.0 %	0.0 %	0.0 %	205
	JPS	0.0 %	0.4 %	38.0 %	1.4 %	0.0 %	1.0 %	0.0 %	18.1 %	0.2 %	10.2 %	0.0 %	0.0 %	0.3 %	0.0 %	30.1 %	0.3 %	0.0 %	3538
	RIC	0.2 %	0.5 %	20.3 %	0.6 %	0.0 %	0.0 %	1.1 %	0.4 %	5.8 %	0.0 %	3.5 %	29.1 %	0.0 %	0.0 %	5.8 %	18.1 %	14.7 %	2699
	RCF	0.0 %	0.1 %	1.5 %	0.0 %	0.0 %	0.0 %	0.0 %	3.2 %	0.0 %	0.0 %	0.0 %	1.5 %	0.0 %	0.0 %	0.0 %	3.7 %	89.9 %	2065
	∑	759	4745	34821	30376	1003	18625	3116	42022	1981	3903	1567	9706	326	892	4093	1385	2700	162020

Figura 27 - Matrizes de Confusão (Proporção do Atual) das predições realizadas no treinamento pelos modelos Decision Tree (superior) e Naive Bayes (inferior) para o problema de classificação multiclasse – Classificação de Litotipos.

Avaliando-se as predições erradas Falsos Negativos, é observado que as classes que apresentam as maiores proporções de predições erradas não são aquelas de menor frequência de ocorrência na base de treino (Figura 18). Isso significa que o treinamento realizado segundo a validação cruzada, com estratificação, foi suficiente para mitigar os efeitos do desbalanceamento das classes. Observa-se um viés para ambos os modelos na proporção de falsos positivos para a as classes preditas HF e JPC, que são a de maior frequência na base de dados de treinamento (Figura

18). Apenas a inserção de novas amostras das demais classes seria capaz de mitigar tal efeito, o que depende da coleta de mais dados.

Os resultados do treinamento também podem ser visualizados nos Striplogs (Figura 28 e Figura 29). Em linhas gerais ambos os modelos demonstraram boa capacidade de predição dos intervalos mineralizados, preservando as características geométricas e recuperando a posição dos contatos. O modelo *Naïve Bayes* aparenta ter mais sensibilidade ao sinal de alta frequência dos atributos da perfilagem, evidenciado pela alta variabilidade nas predições.

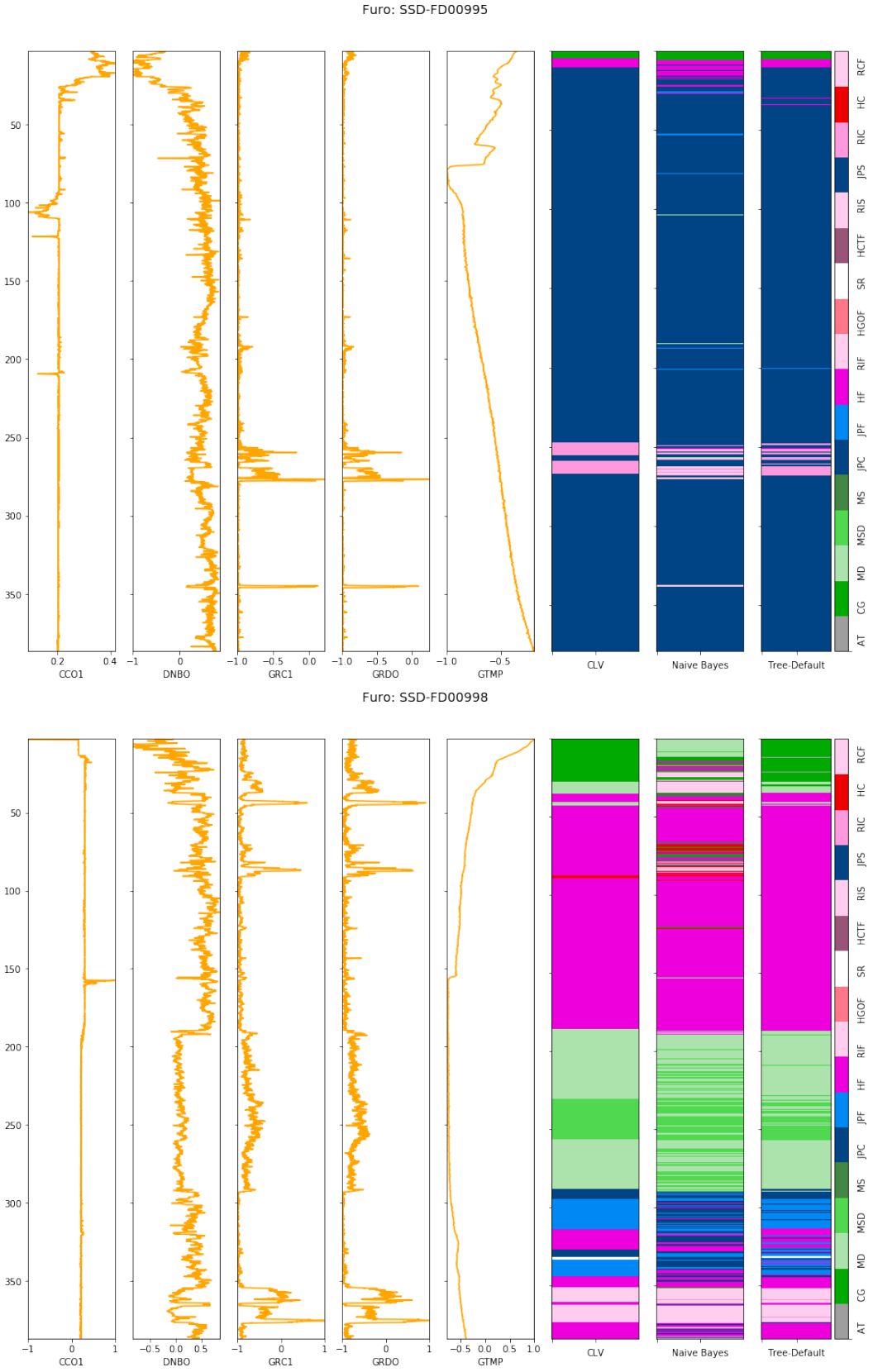


Figura 28 - Striplog dos litotipos (CLV) juntamente com as predições (*Naïve Bayes e Tree-Default*), *com* as curvas da perfilagem geofísica convencional (Atributos de entrada), para os Furos SSD-FD00995 (Superior) e SSD-FD00998 (Inferior). Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6).

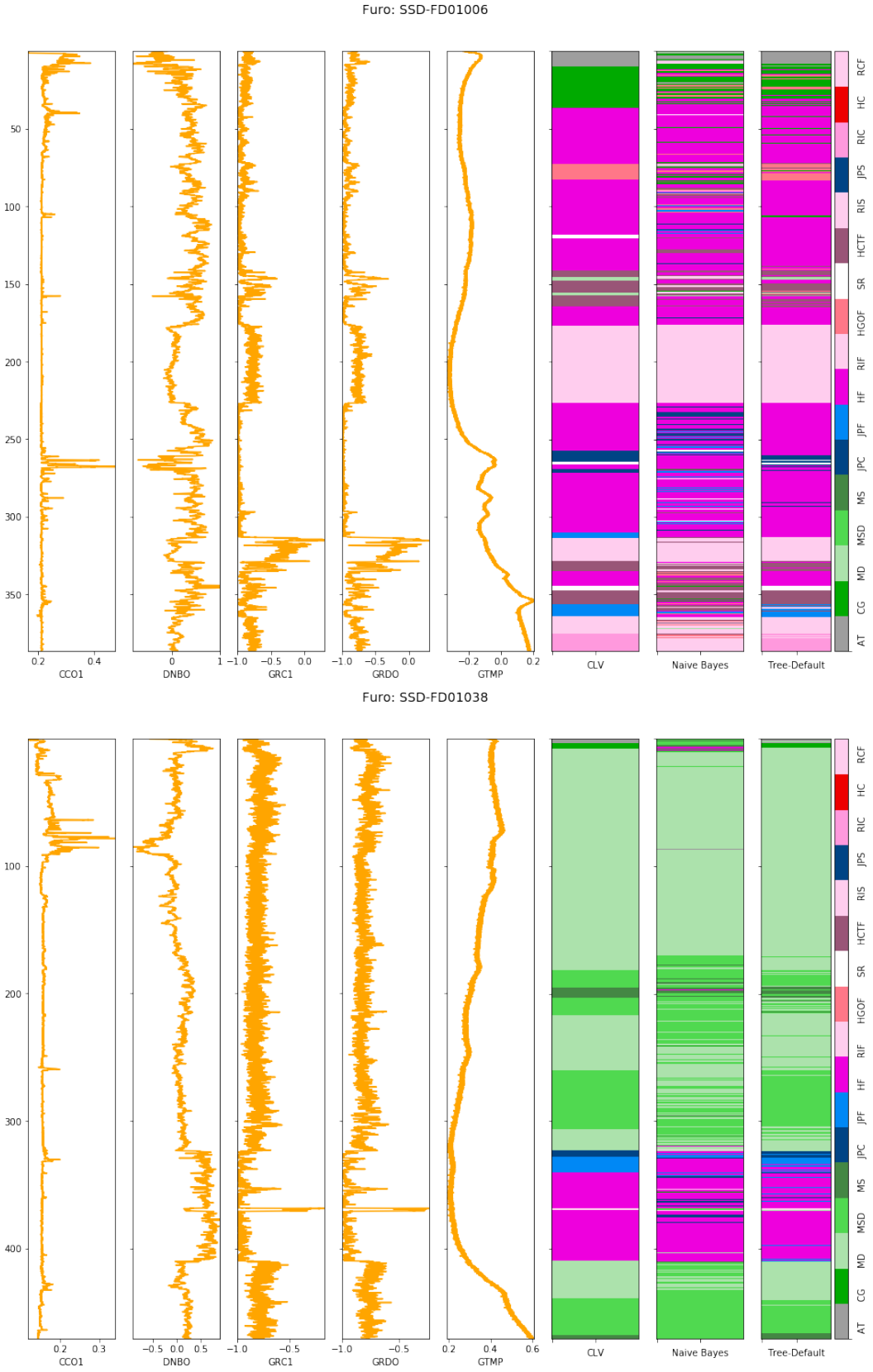


Figura 29 - Striplog dos litotipos (CLV) juntamente com as predições (*Naïve Bayes* e *Tree-Default*), *com* as curvas da perfilagem geofísica convencional (Atributos de entrada), para os Furos SSD-FD00995 (Superior) e SSD-FD00998 (Inferior). Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6).

4.2.2 Validação (Teste-Cego)

O conjunto de teste, composto pelo furo SSD-FD01001, foi utilizado para avaliar a performance dos modelos *Naïve Bayes* e *Decision* Tree, na tarefa de classificação dos litotipos. A Tabela 9 mostra a performance média para cada classe, na tarefa de classificação binária. No teste-cego, assim como ocorreu no treinamento, os modelos tiveram performance satisfatória para todas as métricas (>0.5500).

Tabela 9 - Performance do Teste-Cego. Classificação de Litotipos (CLV) no furo SSD-FD01001.

Model	CA	F1	Precision	Recall
Tree-Default	0.6537	0.6079	0.5787	0.6537
Naive Bayes	0.6177	0.5725	0.5363	0.6177

Quanto comparada a performance dos modelos no teste-cego e no treinamento é observado que o modelo *Decision Tree* mantém o melhor desempenho (F1 = 0.6079 contra F1= 0.5725) na tarefa de classificação multiclasse. Contudo este mesmo modelo apresenta a maior queda no desempenho, saindo de F1=0.9153 para F1=0.6073 (variação de 33%), ao passo que o modelo *Naïve Bayes* apresenta variação de 18% a menos na performance medida pela métrica F1. Este efeito é interpretado como uma menor predisposição, ou capacidade, de generalização para o modelo *Decision Tree*.

Confusion matrix for Tree-Default (showing proportion of actual)

		Predicted																	
		AT	**CG**	**JPC**	**MD**	**MS**	**MSD**	**HCTF**	**HF**	**HGOF**	**JPF**	**RIF**	**RIS**	**SR**	**HC**	**JPS**	**RIC**	**RCF**	**∑**
Actual	**AT**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	CG	0.0 %	0.0 %	69.3 %	28.4 %	0.0 %	1.9 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.4 %	0.0 %	0.0 %	**1955**
	JPC	0.0 %	0.0 %	95.2 %	0.0 %	0.0 %	0.0 %	0.0 %	4.6 %	0.0 %	0.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.1 %	**10159**
	MD	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	MS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	MSD	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	HCTF	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	HF	0.0 %	0.7 %	17.5 %	1.2 %	0.0 %	0.0 %	0.0 %	71.6 %	0.0 %	4.4 %	0.0 %	0.0 %	0.0 %	0.0 %	2.8 %	0.0 %	1.8 %	**8655**
	HGOF	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	JPF	0.0 %	22.6 %	25.3 %	0.0 %	0.0 %	0.0 %	0.0 %	50.5 %	0.0 %	1.5 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.2 %	**1214**
	RIF	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	RIS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	SR	0.6 %	2.1 %	49.6 %	2.4 %	0.0 %	0.0 %	0.0 %	36.9 %	0.0 %	1.5 %	0.0 %	0.0 %	0.0 %	0.0 %	7.0 %	0.0 %	0.0 %	**1445**
	HC	0.0 %	0.0 %	12.5 %	14.2 %	0.0 %	0.0 %	0.0 %	3.0 %	0.0 %	19.8 %	0.0 %	0.0 %	0.1 %	0.0 %	50.5 %	0.0 %	0.0 %	**875**
	JPS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	RIC	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	RCF	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	∑	**9**	**364**	**13667**	**817**		**39**		**7845**		**607**			**1**		**794**		**160**	**24303**

Confusion matrix for Naive Bayes (showing proportion of actual)

		Predicted																	
		AT	**CG**	**JPC**	**MD**	**MS**	**MSD**	**HCTF**	**HF**	**HGOF**	**JPF**	**RIF**	**RIS**	**SR**	**HC**	**JPS**	**RIC**	**RCF**	**∑**
Actual	**AT**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	CG	0.0 %	0.0 %	0.4 %	3.2 %	0.0 %	23.0 %	0.0 %	50.2 %	0.0 %	22.0 %	0.0 %	0.0 %	0.0 %	0.0 %	1.2 %	0.0 %	0.0 %	**1955**
	JPC	0.0 %	0.2 %	92.0 %	0.0 %	0.0 %	0.0 %	0.0 %	7.4 %	0.0 %	0.2 %	0.0 %	0.0 %	0.0 %	0.0 %	0.1 %	0.1 %	0.0 %	**10159**
	MD	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	MS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	MSD	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	HCTF	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	HF	0.0 %	1.5 %	28.1 %	1.5 %	0.0 %	0.8 %	0.0 %	65.0 %	0.1 %	2.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.2 %	0.2 %	0.4 %	**8655**
	HGOF	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	JPF	0.0 %	14.0 %	20.3 %	0.0 %	0.0 %	0.0 %	0.0 %	56.1 %	0.0 %	3.6 %	0.0 %	4.6 %	0.2 %	0.0 %	0.0 %	1.2 %	0.0 %	**1214**
	RIF	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	RIS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	SR	0.0 %	6.2 %	50.0 %	1.9 %	0.0 %	4.4 %	0.0 %	32.0 %	0.0 %	1.3 %	0.0 %	0.0 %	0.0 %	0.0 %	0.1 %	4.0 %	0.0 %	**1445**
	HC	0.0 %	0.0 %	0.0 %	10.3 %	0.0 %	24.1 %	0.0 %	34.4 %	0.0 %	26.3 %	0.0 %	0.0 %	0.6 %	0.0 %	4.3 %	0.0 %	0.0 %	**875**
	JPS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	RIC	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	RCF	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	∑	**1**	**408**	**12751**	**314**		**793**		**8798**	**9**	**932**		**56**	**9**		**100**	**100**	**32**	**24303**

Figura 30 - Matrizes de Confusão (Proporção do Atual) das predições realizadas no teste-cego pelos modelos Decision Tree (superior) e Naive Bayes (Inferior) para o problema de classificação multiclasse – Classificação de Litotipos

A Figura 31 mostra o Striplog das predições no teste-cego para ambos os modelos, juntamente com os perfis para os atributos de entrada da perfilagem geofísica convencional. Os resultados são satisfatórios, a posição dos principais contatos e a relação dos domínios litológicos é recuperada com grande concordância por ambos os modelos. Os litotipos Canga (CG), Hematitito Compacto (HC) e Jaspelito Friável (JPF) foram aqueles que apresentaram pior recuperação por ambos os modelos. O modelo *Naïve Bayes* apresentou maior sensibilidade as variações de alta frequência do dado de entrada, como pode ser observado na maior variabilidade de predições.

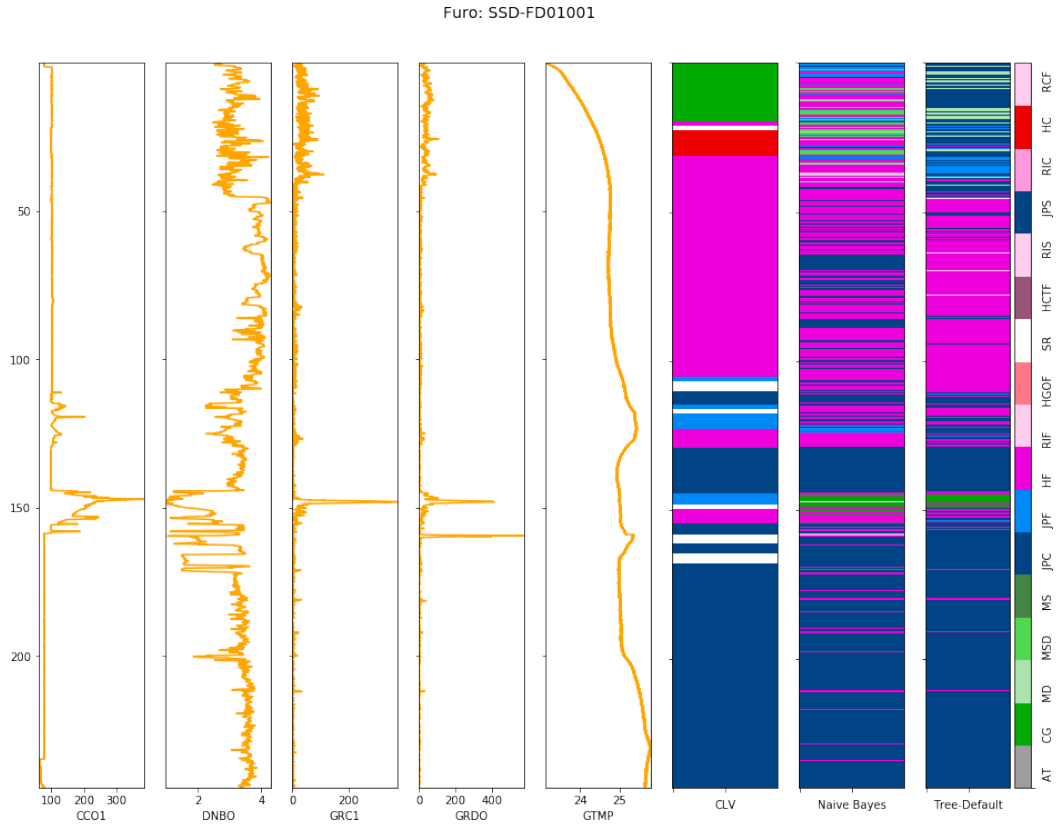


Figura 31 - Striplog dos litotipos (CLV) juntamente com as predições (*Naïve Bayes* e *Tree-Default*), *com* as curvas da perfilagem geofísica convencional (Atributos de entrada), para os Furos SSD-FD00995 (Superior) e SSD-FD00998 (Inferior). Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6).

4.3 Detecção de Minério de Ferro (MFe)

A individualização de diferentes litotipos de acordo com seu valor econômico foi formalizada por um problema de classificação binária, com duas classes possíveis: Minério de Ferro (classe "MFe") e não-minério (classe "others"). Esse problema é de particular interesse para às áreas de exploração mineral (na identificação de novos alvos minerais), de recursos (delineação de alvos já conhecidos).

Nesse tipo de problema, para que não seja gerada expectativa de ocorrência de um bem de valor sustentada por uma mentira (superestimativa da ocorrência de minério), busca-se um modelo de classificação conservador. Ou seja, um modelo que tenha a capacidade de maximizar o número de predições corretas (verdadeiros positivos), ao passo que seja minimizado o número de classificações falsos positivos, com tolerância aos falsos negativos.

4.3.1 Treinamento

Como explicado na seção 3.3.1, a base de dados de treinamento consistiu nos furos: SSD-FD00995, SSD-FD00998, SSD-FD01006 e SSD-FD01038. Esse conjunto foi utilizado para ajustar/otimizar os parâmetros dos algoritmos de classificação em cada um dos problemas, bem como avaliar o desempenho, ou sucesso, na tarefa de predição das classes.

A Tabela 10 mostra a compilação das métricas de performance média, sobre cada classe, obtidas durante o treinamento, considerando-se a validação cruzada em 5 *folds*, estratificados (Seção 3.4.1). Podemos observar que o ambos os algoritmos apesentam capacidade de gerar modelos com bons resultados para todas as métricas (> 0.5500).

Tabela 10 – Performance do treinamento considerando validação cruzada (k=5). Classificação de intervalos de minério de Ferro (MFe).

Model	CA	F1	Precision	Recall
Tree-Default	0.9570	0.9282	0.9299	0.9265
Naive Bayes	0.8868	0.8166	0.7939	0.8408

Para esse problema é esperado um modelo que minimize a classificação errada de MFe, portanto, simultaneamente capaz de minimizar o número de Falsos Positivos (métrica Precisão). Isso ocorre ao custo de não ser classificado como MFe intervalos que seriam MFe, ou seja, incorrer em classificações do tipo Falsos Negativos, subestimando-se as ocorrências de minério. Por tal motivo, para problemas dessa natureza é interessante observar a métrica F1 para a classe MFe, em conjunto com a Precisão.

A matriz de confusão (Figura 32), mostra a proporção de acertos das predições com relação ao valor real. Nela observa-se que o modelo gerado pelo algoritmo *Decision Tree*, em geral, apresenta maior proporção de predições corretas para ambas as classes. Em números são corretas 97.0% das predições para não-minério (contra 90.6% para o modelo *Naïve Bayes*) e de 92.7% de para MFe (contra 84.1% para o modelo *Naïve Bayes*).

Confusion matrix for Tree-Default (showing proportion of actual)

		Predicted		
		other	**MFe**	**∑**
Actual	**other**	97.0 %	3.0 %	**113422**
	MFe	7.3 %	92.7 %	**48598**
	∑	**113601**	**48419**	**162020**

Confusion matrix for Naive Bayes (showing proportion of actual)

		Predicted		
		other	**MFe**	**∑**
Actual	**other**	90.6 %	9.4 %	**113422**
	MFe	15.9 %	84.1 %	**48598**
	∑	**110550**	**51470**	**162020**

Figura 32 - Matrizes de Confusão (Proporção do Atual) das predições realizadas no treinamento pelos modelos Decision Tree (esquerda) e Naive Bayes (direita), para o problema de classificação binária - Detecção de Minério de Ferro.

Para as predições erradas (Falsos Positivos e Negativos) o *Decision Tree* também apresenta melhor desempenho, ou seja, apresenta menor proporção de predições de cada um desses tipos: 3% de Falsos Positivos (contra 9.4% para o modelo *Naïve Bayes), e 7.3% para os* Falsos Negativos (contra 15.9% para o modelo *Naïve Bayes)*.

Ambos os modelos apresentaram menor número, ou proporção, de Falsos Positivos em relação aos Falsos Negativos. Isso demonstra que para o problema específico os modelos apresentam baixa predisposição em classificar como minério intervalos estéreis, ou seja, não superestimam os intervalos mineralizados.

Os resultados do treinamento também podem ser visualizados nos Striplogs (Figura 33 e Figura 34). Em linhas gerais ambos os modelos demonstraram boa capacidade de predição dos intervalos mineralizados, preservando as características geométricas e recuperando a posição dos contatos. Assim como no problema de classificação de litotipos (seção 4.2), o modelo *Naïve Bayes* aparenta ter maior sensibilidade ao sinal de alta frequência dos atributos de entrada da perfilagem geofísica, evidenciado pela maior variabilidade das predições quando comparado ao modelo *Decision Tree*.

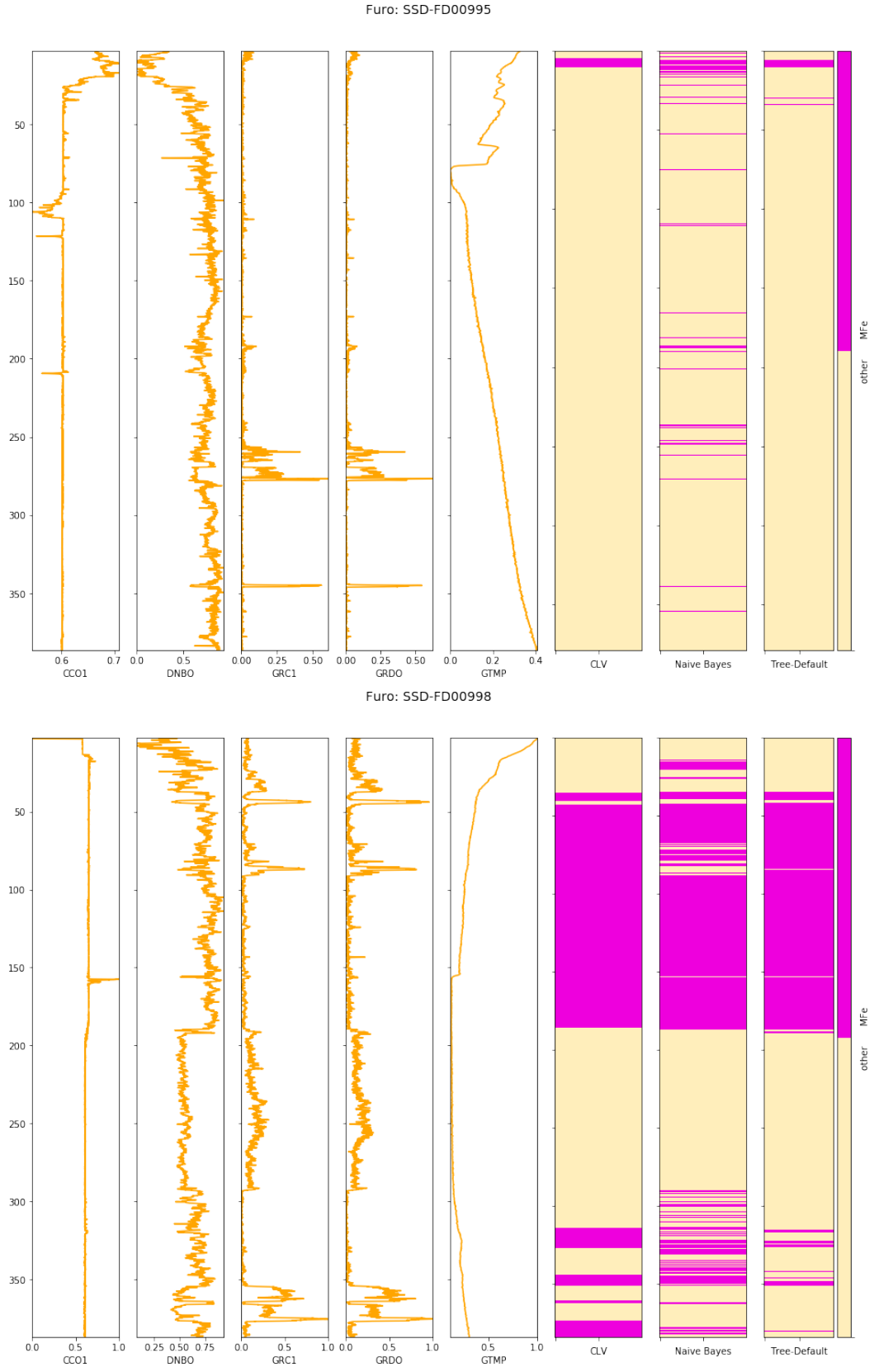


Figura 33 - Striplog dos intervalos mineralizados (CLV) juntamente com as predições (*Naïve Bayes e Tree-Default*), *com* as curvas da perfilagem geofísica convencional (Atributos de entrada), para os Furos SSD-FD00995 (Superior) e SSD-FD00998 (Inferior). Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6).

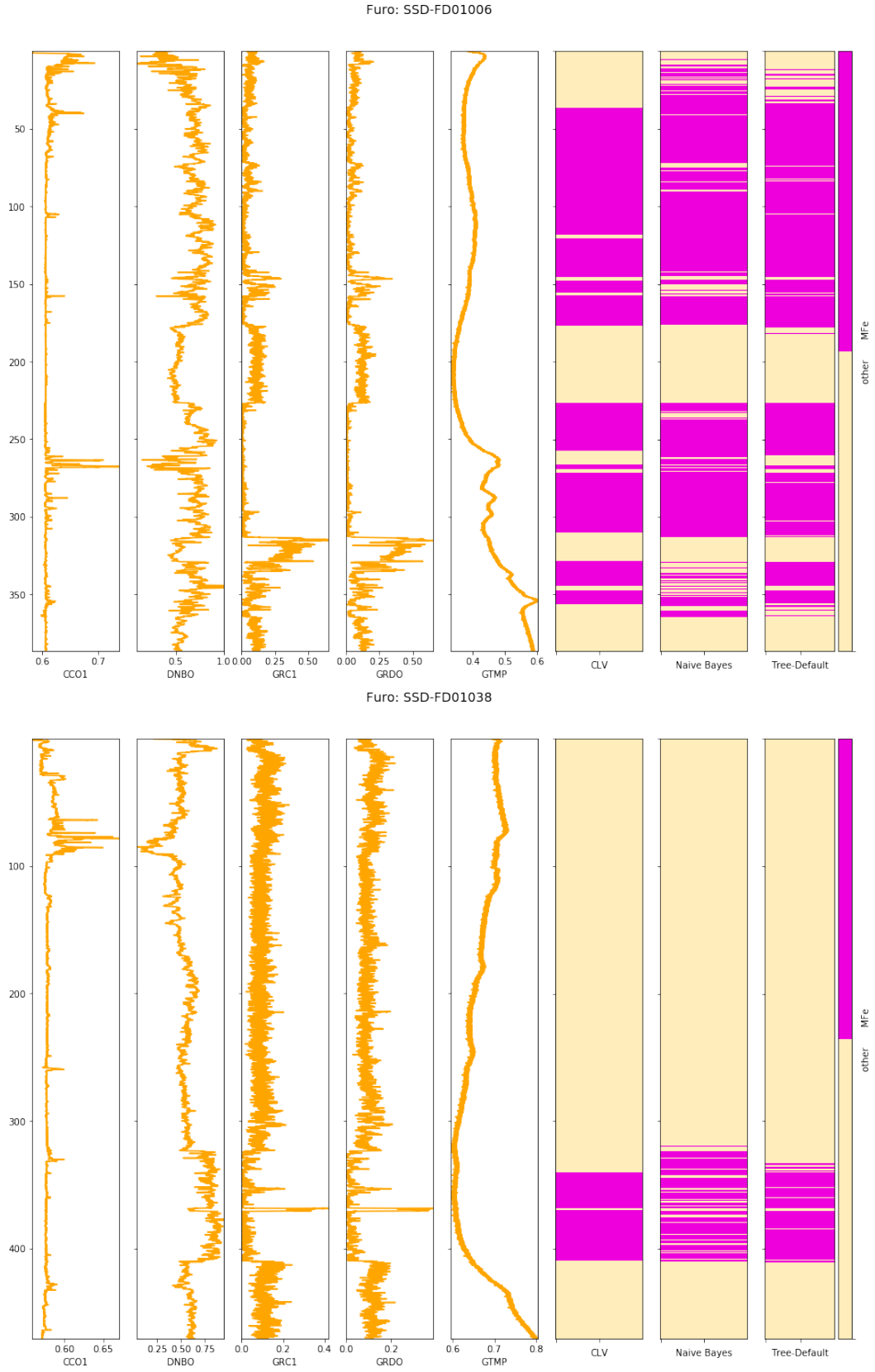


Figura 34 - Striplog dos intervalos mineralizados (CLV) juntamente com as predições (Naïve Bayes e Tree-Default), com as curvas da perfilagem geofísica convencional (Atributos de entrada), para os Furos SSD-FD01006 (Superior) e SSD-FD01038 (Inferior). Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6).

4.3.2 Validação (Teste-Cego)

O conjunto de teste, composto pelo furo SSD-FD01001, foi utilizado para avaliar a performance dos modelos *Naïve Bayes* e *Decision* Tree, na tarefa de classificação de intervalos mineralizados. A Tabela 11 mostra a performance média para cada classe, na tarefa de classificação binária. No teste-cego, assim como ocorreu no treinamento, os modelos tiveram performance satisfatória para todas as métricas (>0.5500).

Tabela 11 - Performance durante teste -cego. Classificação de intervalos de minério de Ferro (MFe) no furo SSD-FD01001.

Model	CA	F1	Precision	Recall
Naive Bayes	0.8229	0.8246	0.8322	0.8229
Tree-Default	0.8027	0.7989	0.8020	0.8027

A principal diferença no comportamento das performances, quando comparadas às do treinamento, é o fato do modelo *Naïve Bayes* ter apresentado as melhores performances. Tal diferença é mais bem compreendida quando avaliamos a Matriz de confusão (Figura 35), nela observamos que essa melhora na performance do modelo *Naïve Bayes* é consequência da maior proporção de predições corretas para intervalos mineralizados (85.1%), combinada à menor proporção de Falsos Negativos (14.9%), contra respectivamente as proporções de 66.8% e 33.2% para o modelo *Decision Tree*.

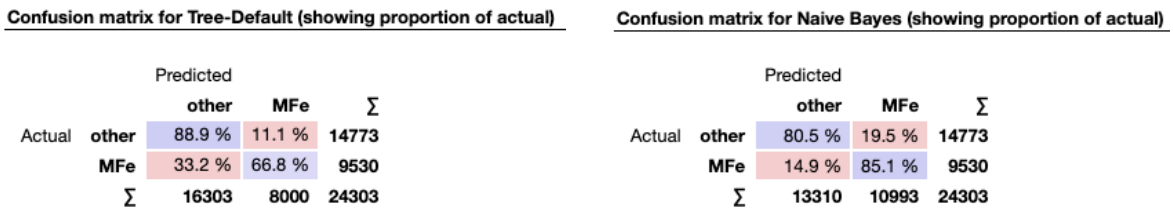


Figura 35 - Matriz de Confusão (Proporção do Atual) das predições realizadas em teste-cego pelos modelos Decision Tree e Naive Bayes para o problema de classificação binária - Detecção de Minério de Ferro.

Em contrapartida o modelo *Naïve Bayes* apresenta maior proporção de Falsos Positivos (19.5%), comparada às predições do modelo *Decision Tree* (11.1%), e consequentemente uma menor precisão para a predição correta da classe de MFe.

Quando comparadas as métricas de performance durante o teste-cego para ambos os modelos com os valores obtidos durante o treinamento, é observado que o modelo *Naïve Bayes* praticamente não apresenta variação. Ao contrário do modelo *Decision Tree*, que apresenta uma queda de aproximadamente 10% para todas as métricas de performance. Isso significa que, quando comparadas as matrizes de confusão, é observada a piora de todas as proporções nessa mesma magnitude.

Esses resultados são muito satisfatórios, e podem ser visualizados nos Striplogs (Figura 36), que mostram que as predições realizadas por ambos os modelos preservam sistematicamente a geometria e sucessão dos intervalos mineralizados a ferro. Todas as zonas mineralizadas, para todos os furos, tiveram amostras corretamente classificadas como minério de ferro (MFe).

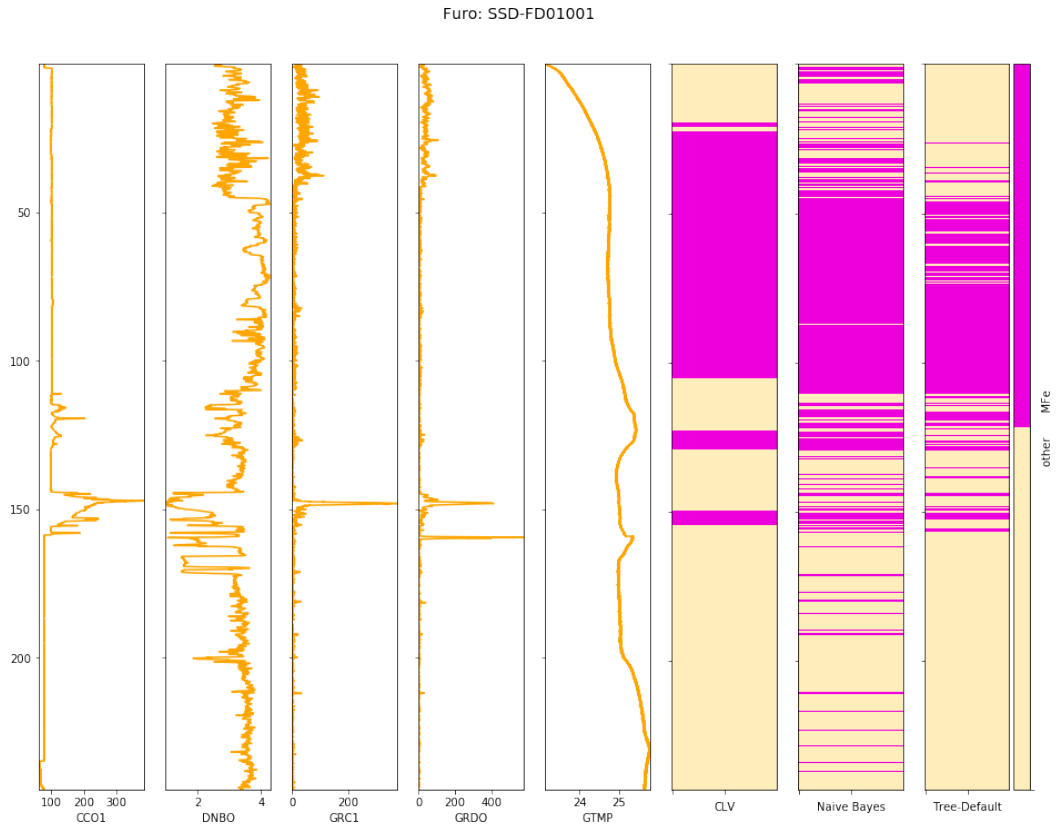


Figura 36 - Striplog dos intervalos mineralizados (CLV) juntamente com as predições (Naïve Bayes e Tree-Default), com as curvas da perfilagem geofísica convencional (Atributos de entrada), para o Furo SSD-FD01001. Na ordem: Caliper (CCO1), Densidade (DNBO), Contagem Total (GRC1 - ferramenta GTC), Contagem Total (GRDO - ferramenta DD6).

5 Conclusões

Os resultados obtidos mostram que o problema de classificação da descrição geológica tem potencial para ser dinamizado, ou parcialmente automatizado, pois os atributos da perfilagem geofísica convencional refletem diferenças que permitiram, sob a abordagem de aprendizado supervisionado, a classificação e predição bem sucedidas tanto dos litotipos quanto para identificação dos intervalos contendo minério de ferro.

A etapa de exploração da base de dados trouxe consigo bons insights sobre domínio dos dados, e permitiu a imaginação de tratamentos específicos dos dados. Podemos pensar no uso de ferramentas de perfilagem complementares, que meçam outras grandezas física e/ou químicas, com o intuíto de possibilitar a separação das classes mal resolvidas nesse trabalho.

A caracterização petrofísica dos litotipos mostra que, para o conjunto de mais de 189 mil amostras, não há distinção significativa das médias das densidades dos jaspelitos e hematititos (3.33 g/cm^3). Entretanto os Hematititos apresentam contagens totais com média pelo menos 4 vezes maiores que as do jaspelitos, podendo os valores de referência obtidos nesse trabalho ser utilizados nos processos de modelagem geológica, estimativa de recursos e inversão geofísica.

Os modelos de classificação obtidos tiveram boa performance para ambos os testes-cegos de classificação do furo SSD-FD01001 (> 0.5500). considerando-se a baixa dimensionalidade da base de dados (5 atributos). A diferença entre as performances durante o treinamento e teste-cego mostram que o modelo *Decision Tree* apresenta menor capacidade de generalização, ou seja, predizer classes baseando-se em amostras não utilizadas no treinamento, do que o modelo *Naïve Bayes*.

As piores performances foram obtidas para o problema multiclasse de classificação de litotipos (*Naïve Bayes* F1= 0.5725 e *Decision Tree* F1 = 0.6079). Esses resultados são considerados satisfatórios. Os litotipos Canga (CG), Hematitito

Compacto (HC) e Jaspelito Friável (JPF) foram aqueles que apresentaram pior recuperação por ambos os modelos (maior proporção de Falsos Negativos).

A performance de ambos os modelos na resolução do problema de classificação binária de identificação de intervalos mineralizados foi bastante satisfatória (*Naïve Bayes* F1= 0.8246 e *Decision Tree* F1 = 0.7989). O modelo *Naïve Bayes* apresentou maior proporção de Verdadeiros Positivos e menor proporção de Falsos Negativos, entretanto o melhor resultado para a minimização dos Falsos Positivos foi do modelo *Decision Tree*, que em linhas gerais apresentou comportamento mais conservador.

Apesar apesar da sua relativa simplicidade, o modelo estimado pelo algoritmo Naïve Bayes teve a melhor capacidade de generalização, classificando os litotipos de menor frequência de ocorrência em maior proporção que os demais modelos.

Outro aspecto relevante no contexto do problema de exploração mineral, que merece ser destacado, é o fato que durante a validação visual da classificação observamos que ambos os modelos apresentaram boa correlação visual e geométrica com o CLV original, preservando em grande parte a posição dos contatos entre os diferentes litotipos ao longo do furo. O Modelo *Naïve Bayes* aparenta ser mais sensível as variações de alta frequência dos sinais (atributos) de entrada.

Como saída positiva, temos que a avaliação e validação dos modelos de classificação de litotipos a partir dos dados de perfilagem geofísica, dentre as opções estudadas nesse trabalho, deverá ser realizada em função dos seguintes aspectos:

i. Performance global– avaliada pelo F1-Score média para todas as classes (problema multiclasse) e deverá ser avaliada em conjunto com a Precisão para problemas binários relacionados a intervalos mineralizados;

ii. Generalização – avaliada pela eficiência em classificar corretamente, durante a etapa de teste, as classes com menor amostragem no conjunto de treino, ou seja, menos conhecidas pelo modelo;
iii. Performance específica – avaliada pela capacidade do modelo classificar corretamente uma ou mais classes de interesse, sob a demanda da área de negócio. Por exemplo, um modelo com performance global menor poderia ser mais eficiente para classificar dada classe de interesse econômico, em detrimento das demais.

O uso de tecnologia como perfilagem geofísica, potencializada pelo aprendizado de máquina, certamente contribuirá significativamente nos processos de exploração mineral e modelagem geológica, tanto em tempo quanto em qualidade. A metodologia desenvolvida nesse trabalho fez uso de um conjunto restrito de furos. A taxa de sucesso das predições poderá ser positivamente impactada pela introdução de novos dados. Esse método é iterativo e os modelos poderão ser refinados à medida em que as campanhas de sondagem e perfilagem nos depósitos de Serra Sul evoluírem.

Referências

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*, 3–31. https://doi.org/10.1016/s0167-7012(00)00201-3

Bateman, R. M. (2015). Cased-hole log analysis and reservoir performance monitoring, second edition. In *Cased-Hole Log Analysis and Reservoir Performance Monitoring, Second Edition*. https://doi.org/10.1007/978-1-4939-2068-6

Blouin, M., Caté, A., Perozzi, L., & Gloaguen, E. (2017). Automated facies prediction in drillholes using machine learning. *79th EAGE Conference and Exhibition 2017 - Workshops*, *June*, 12–15. https://doi.org/10.3997/2214-4609.201701657

Borradaile, G. (2003). *Statistics of Earth Science Data - Their Distribuition in Time, Space and Orientation* (1st editio). Springer-Verlag Berlin Heidelberg GmbH.

Caté, A., Perozzi, L., Gloaguen, E., & Blouin, M. (2017). Machine learning as a tool for geologists. *Leading Edge*, *36*(3), 215–219. https://doi.org/10.1190/tle36030215.1

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., & Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, *14*, 2349–2553. http://jmlr.org/papers/v14/demsar13a.html

Ellis, D. v., & Singer, J. M. (2008). *Well Logging for Earth Scientists* (2nd Editio, Vol. 2). Springer.

Figueiredo e Silva, R. C., Lobato, L. M., Zucchetti, M., Hagemann, S., & Vennemann, T. (2020). Geotectonic signature and hydrothermal alteration of metabasalts under- and overlying the giant Serra Norte iron deposits, Carajás mineral Province. *Ore Geology Reviews*, *120*(June 2019). https://doi.org/10.1016/j.oregeorev.2020.103407

Fullagar, P. K., Livelybrooks, D. W., Zhang, P., Calvert, A. J., & Wu, Y. (2000). Radio tomography and borehole radar delineation of the McConnell nickel sulfide deposit, Sudbury, Ontario, Canada. *Geophysics*, *65*(6), 1920–1930. https://doi.org/10.1190/1.1444876

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and Tensor Flow: Concepts, Tools, and Techniques to Build Intelligent Systems* (N. Tache, Ed.; March 2017). O'Reilly Media, INC. http://oreilly.com/catalog/errata.csp?isbn=9781491962299

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to Statistical Learning* (1st Editio). © Springer Science+Business Media New York 2013. https://doi.org/10.1007/978-1-4614-7138-7_8

Maiti, S., Krishna Tiwari, R., & Kümpel, H. J. (2007). Neural network modelling and classification of lithofacies using well log data: A case study from KTB borehole site. *Geophysical Journal International*, *169*(2), 733–746. https://doi.org/10.1111/j.1365-246X.2007.03342.x

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, *1697900*(Scipy), 51–56. http://conference.scipy.org/proceedings/scipy2010/mckinney.html

Pechnig, R., Haverkamp, S., Wohlenberg, J., Zimmermann, G., & Burkhardt, H. (1997). Integrated log interpretation in the German Continental Deep Drilling Program: Lithology, porosity, and fracture zones. *Journal of Geophysical Research: Solid Earth*, *102*(B8), 18363–18390. https://doi.org/10.1029/96JB03802

Pedregosa, F., Varoquaux, 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, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Pereira, W. R. (2017). *Dissertação de Mestrado. Perfilagem Geofísica Aplicada à Determinação de Parâmetros Geomecânicos em Maciços Rochosos*.

Roldão, D., Ribeiro, D., Cunha, E., Noronha, R., Madsen, A., & Masetti, L. (2012). *Combined Use of Lithological and Grade Simulations for Risk Analysis in Iron Ore, Brazil* (pp. 423–434). https://doi.org/10.1007/978-94-007-4153-9_34

Sauter, R. M. (2002). Introduction to Statistics and Data Analysis. In *Technometrics* (Vol. 44, Issue 1). https://doi.org/10.1198/tech.2002.s664

Shi, G. (2014). *Data Mining and Knowledge Discovery For Geoscientists*. Elsevier Inc.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

VALE. (2016). *Relatório Final - Revisão dos Recursos - Modelo de Serra Sul, Corpos C e D - Minério de Ferro - Diretoria de Planejamento e Desenvolvimento de Ferrosos - Gerência de Recursos Minerais Ferrosos*.

Wanstedt, S. (1992). Geophysical logging applied to ore characterization in the Zinkgruvan mine, Sweden. *Exploration Geophysics*, *23*(2), 401–406. https://doi.org/10.1071/EG992401

Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., & Tu, M. (2018). Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *Journal of Petroleum Science*

and Engineering, *160*(October 2017), 182–193. https://doi.org/10.1016/j.petrol.2017.10.028

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings, Twentieth International Conference on Machine Learning*, *2*, 856–863.

6 Anexos

6.1 Códigos Gerais para Litotipos

MATERIAIS SUPERFICIAIS					
Código	Descrição	Código	Descrição	Código	Descrição
AT	Aterro	CO	Colúvio	PI	Pilha de Itabirito
CG	Canga	LT	Laterita	RA	Rolado Argiloso
CLC	Cloritito Compacto	LTC	Laterita Compacta	RJ	Rejeito
CLF	Cloritito Friável	LTF	Laterita Friável	RO	Rolado
CLS	Cloritito Semi Compacto	LTS	Laterita Semi Compacta	SO	Solo
CM	Canga de Minério	PE	Pilha de Estéril		

MINÉRIO LIMPO					
Código	Descrição	Código	Descrição	Código	Descrição
HC	Hematitito Compacto	HS	Hematitito Semi Compacto	IFR	Itabirito Friável Rico
HF	Hematitito Friável	IC	Itabirito Compacto	IP	Itabirito Pulverulento
HP	Hematitito Pulverulento	IF	Itabirito Friável	IS	Itabirito Semi Compacto

OUTROS	
Código	Descrição
DT	Rocha Destruida
SR	Sem recuperação

MINÉRIO CONTAMINADO					
Código	Descrição	Código	Descrição	Código	Descrição
HACC	Hematitito Arcoseano Compacto	HGOS	Hematitito Goethítico Semi Compacto	ICTP	Itabirito Contaminado Pulverulento
HACF	Hematitito Arcoseano Friável	HMNC	Hematitito Manganesífero Compacto	ICTS	Itabirito Contaminado Semi Compacto
HACP	Hematitito Arcoseano Pulverulento	HMNF	Hematitito Manganesífero Friável	IDOC	Itabirito Dolomítico Compacto
HACS	Hematitito Arcoseano Semi Compacto	HMNP	Hematitito Manganesífero Pulverulento	IDOF	Itabirito Dolomítico Friável
HALC	Hematitito Aluminoso Compacto	HMNS	Hematitito Manganesífero Semi Compacto	IDOP	Itabirito Dolomítico Pulverulento
HALF	Hematitito Aluminoso Friável	HPOC	Hematitito Fosforoso Compacto	IDOS	Itabirito Dolomítico Semi Compacto
HALP	Hematitito Aluminoso Pulverulento	HPOF	Hematitito Fosforoso Friável	IGOC	Itabirito Goethítico Compacto
HALS	Hematitito Aluminoso Semi Compacto	HPOP	Hematitito Fosforoso Pulverulento	IGOF	Itabirito Goethítico Friável
HANC	Hematitito Anfibolítico Compacto	HPOS	Hematitito Fosforoso Semi Compacto	IGOP	Itabirito Goethítico Pulverulento
HANF	Hematitito Anfibolítico Friável	HTAC	Hematitito Talcoso Compacto	IGOS	Itabirito Goethítico Semi Compacto
HANP	Hematitito Anfibolítico Pulverulento	HTAF	Hematitito Talcoso Friável	IHCF	Itabirito com lentes de HC Friável
HANS	Hematitito Anfibolítico Semi Compacto	HTAP	Hematitito Talcoso Pulverulento	IMNC	Itabirito Manganesífero Compacto
HARC	Hematitito Argiloso Compacto	HTAS	Hematitito Talcoso Semi Compacto	IMNF	Itabirito Manganesífero Friável
HARF	Hematitito Argiloso Friável	IALC	Itabirito Aluminoso Compacto	IMNP	Itabirito Manganesífero Pulverulento
HARP	Hematitito Argiloso Pulverulento	IALF	Itabirito Aluminoso Friável	IMNS	Itabirito Manganesífero Semi Compacto
HARS	Hematitito Argiloso Semi Compacto	IALP	Itabirito Aluminoso Pulverulento	IOCC	Itabirito Ocre Compacto
HCTC	Hematitito Contaminado Compacto	IALS	Itabirito Aluminoso Semi Compacto	IOCF	Itabirito Ocre Friável
HCTF	Hematitito Contaminado Friável	IANC	Itabirito Anfibolítico Compacto	IOCP	Itabirito Ocre Pulverulento
HCTP	Hematitito Contaminado Pulverulento	IANF	Itabirito Anfibolítico Friável	IOCS	Itabirito Ocre Semi Compacta
HCTS	Hematitito Contaminado Semi Compacto	IANP	Itabirito Anfibolítico Pulverulento	IPOC	Itabirito Fosforoso Compacto
HDOC	Hematitito Dolomítico Compacto	IANS	Itabirito Anfibolítico Semi Compacto	IPOF	Itabirito Fosforoso Friável
HDOF	Hematitito Dolomítico Friável	IARC	Itabirito Argiloso Compacto	IPOP	Itabirito Fosforoso Pulverulento
HDOP	Hematitito Dolomítico Pulverulento	IARF	Itabirito Argiloso Friável	IPOS	Itabirito Fosforoso Semi Compacto
HDOS	Hematitito Dolomítico Semi Compacto	IARP	Itabirito Argiloso Pulverulento	ITAC	Itabirito Talcoso Compacto
HGOC	Hematitito Goethítico Compacto	IARS	Itabirito Argiloso Semi Compacto	ITAF	Itabirito Talcoso Friável
HGOF	Hematitito Goethítico Friável	ICTC	Itabirito Contaminado Compacto	ITAP	Itabirito Talcoso Pulverulento
HGOP	Hematitito Goethítico Pulverulento	ICTF	Itabirito Contaminado Friável	ITAS	Itabirito Talcoso Semi Compacto

ROCHA ESTERIL					
Código	Descrição	Código	Descrição	Código	Descrição
ACC	Arcósio Compacto	ASF	Ardósia Friável	BVF	Brecha Vulcânica Friável
ACF	Arcósio Friável	ASS	Ardósia Semi Compacto	BVS	Brecha Vulcânica Semi Compacta
ACS	Arcósio Semi-compacto	ATC	Anortosito Compacto	BXC	Bauxita Compacta
ADC	Andesito Compacto	ATF	Anortosito Friável	BXF	Bauxita Friável
ADF	Andesito Friável	ATS	Anortosito Semi Compacto	BXS	Bauxita Semi Compacta
ADS	Andesito Semi Compacto	AXC	Anfibólio Clorita Xisto Compacto	CAC	Calcário Compacto
AGC	Argilito Compacto	AXF	Anfibólio Clorita Xisto Friável	CAF	Calcário Friável
AGF	Argilito Friável	AXS	Anfibólio Clorita Xisto Semi Compacto	CAS	Calcário Semi Compacto
AGS	Argilito Semi-compacto	BCC	Brecha Cataclástica Compacta	CDC	Calcário Dolomitico Compacto
AHEC	Arcóseo Hematítico Compacto	BCF	Brecha Cataclástica Friável	CDF	Calcário Dolomitico Friável
AHEF	Arcóseo Hematítico Friável	BCS	Brecha Cataclástica Semi Compacta	CDS	Calcário Dolomitico Semi Compacto
AHEP	Arcóseo Hematítico Pulverulento	BDC	Brecha Sedimentar Compacta	CHC	Metachert Compacto
AHES	Arcóseo Hematítico Semi Compacto	BDF	Brecha Sedimentar Friável	CHF	Metachert Friável
AMNC	Arcóseo Manganesífero Compacto	BDS	Brecha Sedimentar Semi Compacta	CHS	Metachert Semi-compacto
AMNF	Arcóseo Manganesífero Friável	BHC	Brecha Hidrotermal Compacta	COC	Conglomerado Compacto
AMNP	Arcóseo Manganesífero Pulverulento	BHF	Brecha Hidrotermal Friável	COF	Conglomerado Friável
AMNS	Arcóseo Manganesífero Semi Compacto	BHS	Brecha Hidrotermal Semi Compacta	COS	Conglomerado Semi-compacto
ANC	Anfibolito Compacto	BIC	Biotitito Compacto	CQ	Canga Química
ANF	Anfibolito Friável	BIF	Biotitito Friável	CRC	Cromitito Compacto
ANS	Anfibolito Semi-compacto	BIS	Biotitito Semi Compacto	CRF	Cromitito Friável
APC	Aplito Compacto	BRC	Brecha Compacta	CRS	Cromitito Semi Compacto
APF	Aplito Friável	BRF	Brecha Friável	CSC	Calciossilicática
APS	Aplito Semi Compacto	BRS	Brecha Semi Compacta	CSF	Calciossilicática Friável
ARC	Arenito Compacto	BSC	Basalto Compacto	CSS	Calciossilicática Semi Compacta
ARF	Arenito Friável	BSF	Basalto Friável	CTC	Cataclasito Compacto
ARS	Arenito Semi-compacto	BSS	Basalto Semi Compacto	CTF	Cataclasito Friável
ASC	Ardósia Compacto	BVC	Brecha Vulcânica Compacta	CTS	Cataclasito Semi Compacto

ROCHA ESTERIL					
Código	Descrição	Código	Descrição	Código	Descrição
CXC	Clorita Xisto Compacto	GBF	Gabro Friável	MCS	Metaconglomerado Semi Compacto
CXF	Clorita Xisto Friável	GBS	Gabro Semi Compacto	MD	Máfica Decomposta
CXS	Clorita Xisto Semi Compacto	GDC	Granodiorito Compacto	MGC	Magnetitito Compacto
DBC	Diabásio Compacto	GDF	Granodiorito Friável	MGF	Magnetitito Friável
DBF	Diabásio Friável	GDS	Granodiorito Semi Compacto	MGS	Magnetitito Semi Compacto
DBS	Diabásio Semi Compacto	GNC	Gnaisse Compacto	MHC	Metachert Compacto
DHEC	Diamictito Hematítico Compacto	GNF	Gnaisse Friável	MHF	Metachert Friável
DHEF	Diamictito Hematítico Friável	GNS	Gnaisse Semi Compacto	MHS	MetachertSemi Compacto
DHEP	Diamictito Hematítico Pulverulento	GOC	Gondito Compacto	MIC	Milonito Compacto
DHES	Diamictito Hematítico Semi Compacto	GOF	Gondito Friável	MIF	Milonito Friável
DIC	Diamictito Compacto	GOS	Gondito Semi Compacto	MIS	Milonito Semi-compacto
DIF	Diamictito Friável	GRC	Granito Compacto	MLC	Mármore Dolomítico Compacto
DIS	Diamictito Semi Compacto	GRF	Granito Friável	MLF	Mármore Dolomítico Friável
DOC	Dolomito Compacto	GRS	Granito Semi Compacto	MLS	Mármore Dolomítico Semi Compacto
DOF	Dolomito Friável	GSC	Gossan Compacto	MNC	Manganês Compacto
DOS	Dolomito Semi Compacto	GSF	Gossan Friável	MNF	Manganês Friável
DRC	Diorito Compacto	GSS	Gossan Semi Compacto	MNP	Manganês Pulverulento
DRF	Diorito Friável	GTC	Granitóide Compacto	MNS	Manganês Semi Compacto
DRS	Diorito Semi Compacto	GTF	Granitóide Friável	MRC	Metarenito Compacto
DSC	Dacito Compacto	GTS	Granitóide Semi Compacto	MRF	Metarenito Friável
DSF	Dacito Friável	GVC	Grauvaca Compacta	MRS	Metarenito Semi Compacto
DSS	Dacito Semi Compacto	GVF	Grauvaca Friável	MS	Máfica Sã
DUC	Dunito Compacto	GVS	Grauvaca Semi Compacta	MSD	Máfica Semi Decomposta
DUF	Dunito Friável	GXC	Granada Xisto Compacto	MTC	Migmatito Compacto
DUS	Dunito Semi Compacto	GXF	Granada Xisto Friável	MTF	Migmatito Friável
FCC	Filito Carbonoso Compacto	GXS	Granada Xisto Semi Compacto	MTS	Migmatito Semi Compacto
FCF	Filito Carbonoso Friável	HIC	Hidrotermalito Compacto	MXC	Mica Xisto Compacto
FCS	Filito Carbonoso Semi Compacto	HIF	Hidrotermalito Friável	MXF	Mica Xisto Friável
FDC	Filito Dolomítico Compacto	HIS	Hidrotermalito Semi Compacto	MXS	Mica Xisto Semi Compacto
FDF	Filito Dolomítico Friável	HOC	Hornfels Compacto	NOC	Norito Compacto
FDS	Filito Dolomítico Semi Compacto	HOF	Hornfels Friável	NOF	Norito Friável
FGC	Filito Grafitoso Compacto	HOS	Hornfels Semi Compacto	NOS	Norito Semi Compacto
FGF	Filito Grafitoso Friável	IN	Rocha Intrusiva	OCC	Rocha Ocre Compacta
FGS	Filito Grafitoso Semi Compacto	INC	Rocha Intrusiva Compacto	OCF	Rocha Ocre Friável
FHC	Folhelho Compacto	INF	Rocha Intrusiva Friável	OCS	Rocha Ocre Semi Compacta
FHF	Folhelho Friável	INS	Rocha Intrusiva Semi Compacta	PGC	Pegmatito Compacto
FHS	Folhelho Semi Compacto	JPC	Jaspilito Compacto	PGF	Pegmatito Friável
FLC	Filito Compacto	JPF	Jaspilito Friável	PGS	Pegmatito Semi Compacto
FLF	Filito Friável	JPS	Jaspilito Semi Compacto	PXC	Piroxenito Compacto
FLS	Filito Semi Compacto	LNC	Linhito Compacto	PXF	Piroxenito Friável
FNC	Folhelho Negro Compacto	LNF	Linhito Friável	PXS	Piroxenito Semi Compacto
FNF	Folhelho Negro Friável	LNS	Linhito Semi Compacto	QCC	Quartzo Clorita Xisto Compacto
FNS	Folhelho Negro Semi Compacto	MAC	Mármore Compacto	QCF	Quartzo Clorita Xisto Friável
FSC	Filito Sericítico Compacto	MAF	Mármore Friável	QCS	Quartzo Clorita Xisto Semi Compacto
FSF	Filito Sericítico Friável	MAS	Mármore Semi Compacto	QFC	Quartzito Ferruginoso Compacto
FSS	Filito Sericítico Semi Compacto	MBC	Metabasalto Compacto	QFF	Quartzito Ferruginoso Friável
GAC	Granada Anfibólio Xisto Compacto	MBF	Metabasalto Friável	QFS	Quartzito Ferruginoso Semi Compacto
GAF	Granada Anfibólio Xisto Friável	MBS	Metabasalto Semi Compacto	QXC	Quartzo Xisto Compacto
GAS	Granada Anfibólio Xisto Semi Compacto	MCC	Metaconglomerado Compacto	QXF	Quartzo Xisto Friável
GBC	Gabro Compacto	MCF	Metaconglomerado Friável	QXS	Quartzo Xisto Semi Compacto

ROCHA ESTERIL			
Código	Descrição	Código	Descrição
QZC	Quartzito Compacto	SPF	Serpentinito Friável
QZF	Quartzito Friável	SPS	Serpentinito Semi Compacto
QZS	Quartzito Semi Compacto	SXC	Sericita Xisto Compacto
RBC	Rocha Intrusiva Básica Compacta	SXF	Sericita Xisto Friável
RBF	Rocha Intrusiva Básica Friável	SXS	Sericita Xisto Semi Compacto
RBS	Rocha Intrusiva Básica Semi Compact	TLC	Lapilli Tufo Compacto
RCC	Rocha Intrusiva Ácida Compacta	TLF	Lapilli Tufo Friável
RCF	Rocha Intrusiva Ácida Friável	TLS	Lapilli Tufo Semi Compacto
RCS	Rocha Intrusiva Ácida Semi Compacta	TXC	Talco Xisto Compacto
RDC	Riodacito Compacto	TXF	Talco Xisto Friável
RDF	Riodacito Friável	TXS	Talco Xisto Semi Compacto
RDS	Riodacito Semi Compacto	VQ	Veio de Quartzo
RIC	Riolito Compacto	VQC	Veio de Quartzo Compacto
RIF	Riolito Friável	VQF	Veio de Quartzo Friável
RIS	Riolito Semi Compacto	VQS	Veio de Quartzo Semi Compacto
RUC	Rocha intrusiva Ultramáfica Compacta	VUC	Rocha Extrusiva Compacto
RUF	Rocha intrusiva Ultramáfica Friável	VUF	Rocha Extrusiva Friável
RUS	Rocha intrusiva Ultramáfica Semi Com	VUS	Rocha Extrusiva Semi Compacto
SCAC	Siltito Carbonoso Compacto	XGC	Xisto Grafitoso Compacto
SCAF	Siltito Carbonoso Friável	XGF	Xisto Grafitoso Friável
SCAP	Siltito Carbonoso Pulverulento	XGS	Xisto Grafitoso Semi Compacto
SCAS	Siltito Carbonoso Semi Compacto	XTC	Xisto Compacto
SIC	Siltito Compacto	XTF	Xisto Friável
SIF	Siltito Friável	XTS	Xisto Semi Compacto
SIS	Siltito Semi Compacto	ZC	Zona de Cisalhamento
SPC	Serpentinito Compacto	ZT	Zona de Transição