INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Aplicação de Redes Neuronais Artificiais para classificação das operações de perfuração: O caso de poços *deepwater* da Galp em diferentes geografias

Valter José Chaile

Mestrado em Informática e Gestão

Orientadores:
Doutor Sérgio Moro, Professor Auxiliar com Agregação,
Iscte - Instituto Universitário de Lisboa

Doutor Aristides Carneiro, Engenheiro de Petróleo
Galp Exploração e Produção

Agosto, 2020

“A Man's mind stretched to a new idea never goes back to its original dimensions.”

Oliver Wendell Holmes, Jr.

Dedico este trabalho aos meus Pais que sempre lutaram
para que minha mente fosse aberta a novas ideias,
a minha esposa e ao meu filho Ayan, que me deram forças para que
eu concluísse este projeto antes dele nascer

Agradecimentos

Sem a ajuda e colaboração de muitas pessoas não teria sido possível a realização deste trabalho.

A Deus, por me ter guiado e iluminado em toda esta caminhada.

Aos professores do Instituto Universitário de Lisboa: Rúben Pereira, Sérgio Moro, Fernando Brito e Abreu, José Barateiro, Carlos Coutinho, Álvaro Rosa, Luís Rodrigues, Pedro Ramos, Manuela Aparício, João Pina e João Guerreiro, pelos valiosos ensinamentos.

Ao Professor Sergio Moro, meu orientador, por ter aceite e acreditado no meu trabalho e pela dedicação demonstrada em todas as suas fases, pelo apoio, paciência, orientação, supervisão, incentivo e por ter corrigido pacientemente as minhas incorreções.

Ao Aristides Carneiro, meu coorientador e colega de trabalho, que sabiamente e com sua larga experiência na área de E&P ajudou a tornar este trabalho real.

Aos amigos do Mestrado Elen, Edson, Yotelma, Pedro, Carlota, Celício, Jônatas, Sérgio, Sabrina e Yolanda, pelo apoio, amizade, carinho, convívio e por terem dividido todos os momentos alegres e tristes comigo.

Aos meus Pais José Salomão F. Chaile (em memória) e Marta Isaías Chaile, que sempre investiram, incentivaram e lutaram pela minha educação.

À minha esposa Helena Chaile que, de forma incansável, esteve sempre ao meu lado em todos os momentos difíceis, acompanhando e motivando para a concretização deste projeto.

Aos meus irmãos Salomão T. Chaile e Isac Wilson S. Chaile, pelo encorajamento e forças quando decidi embarcar nesta aventura, e a minha Família em geral, pelo apoio, atenção e forças dadas.

A Galp E&P, e aos meus colegas, Jorge Lourenço, Daniel Patrocinio, Diana Braceiro, Luis Klem, Frederico Henriques, Igor Sobral, Paulo Gomes, Tiago Monte, Laura Carvoeiro, Ana Rute, Inês Couto, pela ajuda, paciência e compressão durante a minha formação.

Ao Professor Rodrigues Fazenda, meu docente e orientador na licenciatura pelo incentivo para que eu fizesse o Mestrado.

Aos meus amigos do IT Project, pelo apoio.

A todos, o meu sincero “Obrigado”.

Resumo

As grandes descobertas de petróleo em águas ultra-profundas são um grande incentivo à produção petrolífera e têm representado grandes investimentos na Galp. No entanto, explorar estes valiosos recursos em altas profundidades requer a utilização de ferramentas e metodologias que permitam prever como vai ser o comportamento dos poços durante a perfuração.

A equipa de perfuração é responsável por estimar os custos e duração das operações em novos poços e simular a perfuração, estimando um resultado que culmina com um plano de perfuração. A perfuração é uma operação que acarreta custos elevados que são proporcionais à duração das atividades. Em poços exploratórios, caso não sejam encontrados hidrocarbonetos, não haverá retorno do investimento. Por esta razão, a classificação das operações durante a perfuração é muito importante para gerar premissas de duração para o projeto de novos poços.

Para este estudo, três procedimentos independentes foram propostos. O primeiro consiste na classificação das operações de perfuração nos atributos: (i)Tipo de operação, (ii) causa de tempo não produtivo e (iii) fase de operação, usando as redes *Multi-Layer Perceptron* (MLP). O segundo procedimento diz respeito à classificação usando redes *Long Short-Term Memory* (LSTM) para os mesmos atributos, O terceiro e último, diz respeito à comparação dos resultados dos dois modelos propostos com o objetivo de identificar aquele que apresentou o melhor resultado.

Através desse trabalho é possível concluir que os dados de perfuração atualmente disponíveis representam uma fonte rica de informação e podem ser utilizados para otimizar o processo de construção de poços de petróleo.

Palavras-Chave: redes neuronais artificiais; inteligência artificial; classificação; aprendizagem de máquina; perfuração; completação.

Abstract

The large Ultra-Deepwater oil discoveries are a major incentive for oil production and have been a major investment in Galp, but the exploration of these valuable natural resources at high depths requires the use of tools and methodologies that are capable to predict how It will be the behaviour while drilling.

The drilling team is responsible for estimating the costs and duration of operations in new simulated wells and estimating a result that culminates with a drilling plan. Drilling is an operation that carries high costs that are proportional to the duration of activities. For exploration wells, if no hydrocarbons are found, there will be no return on investment. For this reason, the classification of the various operations during drilling is key to generate duration assumptions for the design of new wells.

For this study, three independent procedures have been proposed. The first consists of the classification of drilling operations in the attributes: (i)Type of operation, (ii) cause of non-productive time and (iii) phase of operation, using Multi-Layer Perceptron (MLP) networks. The second procedure concerns the classification of the same attributes using Long Short-Term Memory (LSTM) networks. The third and last one concerns the comparison of the results of the two models proposed with the objective of identifying the model that presented the best result.

Through this work it is possible to conclude that the available drilling data represent a potential source of information and can be used to optimize the petroleum well construction.

Keywords: Artificial Neural Network; Artificial Intelligence; Classification; Machine Learning; Drilling; Completion.

Índice

Índice de quadros

Índice de figuras

AAPE *Average Absolute Percentage Error*

ANN *Artificial Neural Network*

BPD Barril de petróleo por dia

BDP Boletim Diário de Perfuração

BHA *Bottom Hole Assembly*

BOP *Blow Out Preventor*

CRISP-DM *Cross Industry Standard Process for Data Mining*

CPU *Central Processing Unit*

DT *Decision Trees*

D&C *Drilling & Completion*

ECD *Equivalente Circulation Density*

EUA Estados Unidos da América

E&P *Exploration and Production*

FP *False Positive*

FN *False Negative*

FTE *Full-Time Equivalent*

GA *Genetic Algorithm*

GIS *Geographic information system*

HDD *Hard Disk Drive*

HIIP *Hidrocarbons Initially In Place*

IDE *Integrated Development Environment*

IA Inteligência Artificial

K-NN *K-Neurest Neighbor*

KDD *Knowledge Discovery in Databases*

LSTM *Long Short Term Memory*

ML *Machine Learning*

MAE *Maximum Absolute Error*

MAPE *Mean Absolute Percentage Error*

MSE *Mean Square Error*

m/ft *Meters / feets*

MLP *Multilayer Perceptron*

NFLTO *No Free Lunch Theorems for Optimization*

NPT *Non-Productive Time*

O&G *Oil & Gas*

TP *True Positive*

PWDa *Pressure While Drilling Analyzer*

PoS *Probability of Success*

PT *Productive Time*

R^2 *Correlation Coefficient*

ROC *Receiver Operating Characteristics*

RELU *Rectified Linear Unit*

RNN *Recurrent Neural Network*

RNA Redes Neuronais Artificial

ROI *Return of Investiment*

RSL Revisão Sistemática de Literatura

RMSE *Root Mean Square Error*

SA Sociedade Anónima

SIG *Sistemas de Informação Geografica*

SPE *Society of Petroleum Engineers*

SVM *Support Vetor Machines*

Capítulo 1

1. Introdução

Este capítulo trata sobre o enquadramento do trabalho a desenvolver na dissertação, evidenciando as razões que motivam o desenvolvimento de um estudo nesta área, assim como os objetivos a alcançar e resultados que se esperam obter. Além disso, é descrita a abordagem metodológica adotada e qual a estratégia de pesquisa bibliográfica utilizada.

1.1. Enquadramento

Segundo Fayyad & Shapiro (1996), a noção de encontrar padrões úteis nos dados tem recebido uma variedade de nomes, incluindo *Data Mining* (DM), extração de conhecimento, descoberta de informações, coleta de informação, assim como processamento de padrão de dados. O termo DM tem sido principalmente usado por estatísticos, analistas de dados e por operadores de Sistemas de Informação Geográfica (GIS). Além disso, vem ganhando muita popularidade nas áreas de bases de dados. O volume de dados armazenados por estes sistemas chegou a um nível onde a análise dos dados, por meio de processos manuais ou de técnicas tradicionais, não explora toda a informação implícita disponível.

Em 1997, Adriaans e Zantinge já afirmavam que as pessoas não seriam capazes de analisar estes dados de forma manual, e que as organizações teriam maior capacidade de sobreviver no mercado se utilizassem ferramentas automáticas para extração de conhecimento a partir de seus dados. A utilização de recursos para este tipo de análise já vem sendo feita desde o final dos anos 80, num processo chamado Descoberta de Conhecimento em Base de Dados (KDD, do inglês *Knowledge Discovery in Databases*), que foi popularizado num workshop, em 1989, por Piatetsky-Shapiro, para enfatizar que o conhecimento é o produto final de uma descoberta orientada a dados. Esta abordagem foi popularizada nos campos de Inteligência Artificial (IA) e *Machine Learning* (ML) (Fayyad, U. M., Piatetsky-Shapiro, 1996).

KDD é um processo interativo e iterativo de várias etapas, desde limpeza, integração, seleção, transformação, DM, avaliação dos padrões e, por último, apresentação do conhecimento. É repetitivo e muitas vezes precisa de auxílio do usuário (Han, Kamber, & Pei, 2011). Para George & Bennett (2005), os estudos de casos não devem ser escolhidos de forma aleatória, mas por sua relevância de contributo para o objetivo da pesquisa.

Por esta razão, neste estudo, é apresentada uma organização com grande representatividade em Portugal na área de exploração e produção de petróleo e gás: A Galp S.A[1]. Trata-se um grupo de empresas portuguesas no setor de energia. É detentora da Petrogal e da Gás de Portugal, sendo hoje um grupo integrado de produtos petrolíferos e de gás natural, com atividades que se estendem, desde a exploração e produção de petróleo e gás natural, à refinação e distribuição de produtos petrolíferos, distribuição e venda de gás natural e a geração de energia elétrica. Atualmente está entre as maiores

[1] https://www.galp.com/corp/pt/sobre-nos/a-galp/organizacao

empresas de Portugal, controlando cerca de 50% do comércio de combustíveis no país e a totalidade da capacidade refinadora de Portugal.

A Galp Exploração e Produção (E&P), é uma unidade de negócio do grupo Galp S.A., que é responsável pelas atividades da Galp S.A. no sector *upstream*[2] de petróleo e gás natural. A esta área compete a coordenação, supervisão, estudo e execução de todos os trabalhos relacionados com a prospeção, pesquisa, avaliação, desenvolvimento e produção de hidrocarbonetos nos projetos em que ela está envolvida, bem como a responsabilidade de identificar, analisar e promover oportunidades de desenvolvimento em novos projetos de exploração e produção petrolífera.

1.2. Motivação e relevância do tema

Com as grandes descobertas de petróleo em águas ultra profundas no Brasil (pré-sal), a Galp E&P tem estado a investir na exploração destes recursos. No entanto, de acordo com Empressa de Pesquisa Energética (2018) explorar estas reservas em grandes profundidades, requer um desafio enorme para os engenheiros de perfuração.

A equipa de perfuração e completação, representada pela sigla D&C, é responsável por definir e implementar os controlos necessários relacionados a todas as atividades operacionais e de engenharia de poço, designadamente medidas de projeto e integridade, desempenho e otimização de custos, com o objetivo de maximizar o desempenho dos ativos de E&P da Galp e a maior aderência possível (Galp, 2017).

A perfuração de um poço de petróleo é uma operação bastante complexa que está sujeita à ocorrência de uma série de anormalidades inerentes à incerteza geológica que existe nestas operações. A evolução das tecnologias utilizadas na perfuração, como, por exemplo, Programa de Diagnóstico de Problemas de Perfuração em Tempo Real (do em inglês *Pressure While Drilling Analyzer* – PWDa) têm proporcionado uma redução na frequência de ocorrência de problemas ao longo dos anos, mas ainda assim, alguns desafios persistem. As complicações que podem ocorrer durante a perfuração são altamente indesejáveis e causam grandes prejuízos, como a elevação do custo da perfuração e, muitas vezes, a perda do poço (Guilherme, Queiroz, Urgal, & Chavarette, 2010).

Um dos grandes desafios que as equipas de D&C, nas empresas não operadoras têm, é o de extrair informação de forma automática e estruturada, reduzindo a quantidade de extração de dados realizada manualmente, o que trará benefícios para a equipa, permitindo que os engenheiros despendam mais tempo a analisar dados de operações em vez da extração dos mesmos.

Segundo Rich & Knight (1991), IA é o estudo de como os computadores podem fazer tarefas que hoje são desempenhadas por pessoas. Os autores Rich & Knight (1991), afirmam ainda que existe uma

[2] A frase *Upstream* caracteriza-se pelas atividades de busca, identificação e localização das fontes de óleo, e ainda o transporte deste óleo extraído até as refinarias, onde será processado. Resumindo, são as atividades de exploração, perfuração e produção.

perceção generalizada de que a análise de dados de perfuração ocorre durante o planeamento de um novo poço. Dessa forma, as lições apreendidas através da análise de dados poderiam contribuir para a melhoria do processo de perfuração de novos poços. No entanto, Heriot Watt (2013) afirma que devido às pressões diárias a que um engenheiro de perfuração de poços é submetido, a análise de dados nem sempre é conduzida de maneira apropriada e muitas vezes é feita de forma manual.

Tendo em vista o exposto, a presente dissertação propõe um modelo de IA para classificação de fases de operações de perfuração e completação, baseado em técnicas de redes neuronais artificiais.

1.3. Contribuições esperadas

Pretende-se que o modelo seja implementado dentro da unidade de negócios Galp E&P, com a equipa de perfuração, numa fase inicial e, depois, seja estendida para diversas áreas tais como: Engenharia de Reservatórios, Garantia de Escoamento & Processo, e Projetos. Para além disso, pretende-se contribuir para a ciência através de classificação de texto usando modelos de redes neuronais com um caso de estudo real.

1.4. Questões e objetivos de investigação

1.4.1. Objetivo Geral

O objetivo do trabalho é construir um modelo de IA para indústria de E&P usando redes neuronais artificias.

1.4.2. Objetivos Específicos

Para alcançar o objetivo geral, foi desenvolvido o trabalho em três etapas:

- realizar uma Revisão Sistemática de Literatura (RSL) acerca do tema;
- construir o modelo de classificação usando redes neuronais;
- classificar as operações de perfuração de um poço de petróleo através do modelo construído;

1.5. Abordagem metodológica

1.5.1. Estratégia de pesquisa bibliográfica

De modo a compreender os conceitos fundamentais para o trabalho de dissertação e identificar os trabalhos relevantes na área, bem como efetuar um levantamento do estado da arte, foi necessário fazer uma pesquisa e seleção criteriosa dos artigos. Nesse sentido, foi definido que, para encontrar as publicações com interesse no tema, seriam utilizadas as plataformas *B-On*, *Google Scholar*, *Web of Science*, *ScienceDirect*, *SpringerLink*, *OnePetro* e *Scopus*.

Para a pesquisa dos artigos de cada uma das áreas, foram utilizados os seguintes termos de pesquisa: '*Artificial Intelligence, 'neural network', 'natural language processing', 'data classification', 'drilling and completion'.*

Para encontrar artigos que abordassem a aplicação de técnicas de *Machine Learning* em *Drilling and Completion* para efetuar revisões de literatura, recorreu-se à seguinte *query*: ('*data classification'* OR '*machine learning*' OR *natural language processing*' OR '*Neural Networks*') AND ('*drilling'* OR '*Oil&Gas*') YEAR (2000 TO 2019)).

Dos cerca de 50 artigos obtidos através das pesquisas, nem todos eram relevantes. Após a leitura dos títulos e *abstracts* dos mesmos, considerando também o número de citações, foi possível fazer uma triagem, priorizando os catorze artigos mais recentes e considerados relevantes para as áreas abordadas na dissertação.

1.5.2. Tipo de pesquisa

O presente trabalho baseia-se numa pesquisa bibliográfica e de estudo de caso, com abordagem qualitativa com natureza aplicada. De acordo com a definição de Gil (2002), esta pesquisa é considerada de estudo de caso, por buscar analisar um tema observado na realidade, explicando como e porque ele ocorre. Além de identificar os fatores que contribuem para que o tema em questão se materialize.

Quanto à natureza, de acordo com Silva & Menezes (2005) é considerada aplicada, pois tem como objetivo gerar conhecimento prático, buscando a solução de problemas específicos. A abordagem, de acordo com Neves (2002), é uma investigação qualitativa, uma vez que se trata de recolha de dados, através de relatórios diários recebidos durante a perfuração de um poço.

1.6. Estrutura e organização da dissertação

O presente estudo está organizado em cinco capítulos que pretendem refletir as diferentes fases até à sua conclusão.

O primeiro capítulo introduz o tema da investigação e objetivos da mesma, bem como uma breve descrição da estrutura do trabalho, assim como a estratégia de pesquisa bibliográfica adotada.

O segundo capítulo reflete o enquadramento teórico, designado por Revisão da literatura, em que são explorados e descritos os vários conceitos inerentes à dissertação, mais concretamente os conceitos relativos a áreas de redes neuronais e perfuração e completação de poços de petróleo e gás.

O terceiro capítulo é dedicado à Metodologia utilizada no processo de recolha e tratamento de dados, bem como os métodos de análise utilizados.

O quarto capítulo apresenta a análise dos resultados obtidos, de acordo com a metodologia que se entendeu apropriada.

No quinto e último capítulo apresentam-se as conclusões deste estudo, assim como as recomendações, limitações e possibilidade de trabalhos futuros.

Capítulo 2

2. Revisão da Literatura

Neste capítulo é apresentada uma revisão da literatura sobre os conceitos de IA, os seus modelos e comparação entre eles. É descrita, em detalhe, o modelo escolhido para o estudo e é feita uma revisão de literatura sobre a perfuração e completação de um poço de petróleo.

2.1. Inteligência Artificial

Rich & Knight (1991), afirmam que um dos avanços mais importantes na comunidade científica que atraiu praticamente todos os campos do esforço humano é o conceito de IA. Ao definir o termo, Leia (2015) e Agwu, Akpabio, Alabi, & Dosunmu (2018) consideram que "*artificial*" pode significar simplesmente não ter ocorrência na vida real ou não ter a mesma configuração na natureza. No entanto, a definição de inteligência é ampla e variada (Legg & Hutter, 2007). Enquanto que Agwu et al. (2018), defendem que o campo da IA sofre situações extremas, tais como definições insuficientes e demais. Para reforçar ainda mais esse ponto, Nilsson (2009) afirmou que a IA carece de uma definição universalmente aceite. Enquanto as tentativas de dar uma definição abrangente ao conceito de IA são prejudicadas pela complexidade explosiva, essas complexidades não diminuem o fato de que a IA é uma grande promessa como ferramenta para entender o relacionamento entre entidades complexas (Agwu et al., 2018). No entanto, este trabalho se alinhará à definição de IA proposta por Rich & Knight (1991), que afirma IA como o estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas.

2.1.1. Evolução de IA

Antes de investigar sobre a IA, é pertinente examinar como surgiu o conceito. Mcculloch & Pitts (1943) referem que o ser humano sempre quis uma máquina que fizesse o trabalho de agir e pensar como ele. Assim, estudos de várias áreas começaram a estudar essa possibilidade.

Em 1943, Warren McCulloch e Walter Pitts apresentam um artigo que fala pela primeira vez de redes neuronais, estruturas de raciocínio artificiais em forma de modelo matemático que imitam o nosso sistema nervoso (Barreto, 2002).

Durante o período de 1981-1990, o Governo Japonês começou a produzir computadores de quinta geração, que poderiam ser capazes de processar inteligência. Esses computadores podem processar linguagem natural, jogar *game*, reconhecer imagens de objetos e provar teoremas matemáticos, todos eles inseridos no domínio da IA (Konar, 1999). Em geral, os cronogramas no desenvolvimento das técnicas de IA e os seus desenvolvedores estão organizados cronologicamente (Tabela 2.1). Observa-se que as redes neuronais artificiais foram as primeiras técnicas de IA a serem desenvolvidas, enquanto os sistemas inteligentes híbridos são os mais recentes.

Tabela 2.1:Tendência evolutiva de várias técnicas de inteligência artificial, adaptado de Agwu et al. (2018).

Técnicas de IA	Ano de desenvolvimento	Desenvolvedor
Redes Neuronais Artificiais (RNA)	1943	McCulloch and Pitts
Fuzzy logic	1965	Lofti A. Zadeh
Genetic algorithm (GA)	1970	John Holland
Case based reasoning	1977	Schank and Abelson
Support vetor machines (SVM)	1995	Vapnik, V
Particle swarm algorithm	1995	Eberhart and Kennedy

2.2. Machine Learning

Machine Learning (ML) envolve um conjunto de técnicas que visam ensinar os computadores a fazerem o que é natural para os humanos e animais: aprender com a experiência. Os algoritmos de ML usam métodos computacionais para "aprender" informações diretamente dos dados, sem depender de uma equação predeterminada como modelo. Os algoritmos melhoram adaptando o seu desempenho à medida que os números de amostras disponíveis para aprendizagem aumentam (Mathworks, 2016).

De acordo com Monard & Baranauskas (2003), ML usa dois tipos de técnicas de aprendizagem: aprendizagem supervisionada, que treina um modelo em dados conhecidos de entrada e saída para que ele possa prever resultados futuros, e aprendizagem não supervisionado, que encontra padrões ocultos ou estruturas intrínsecas nos dados de entrada.

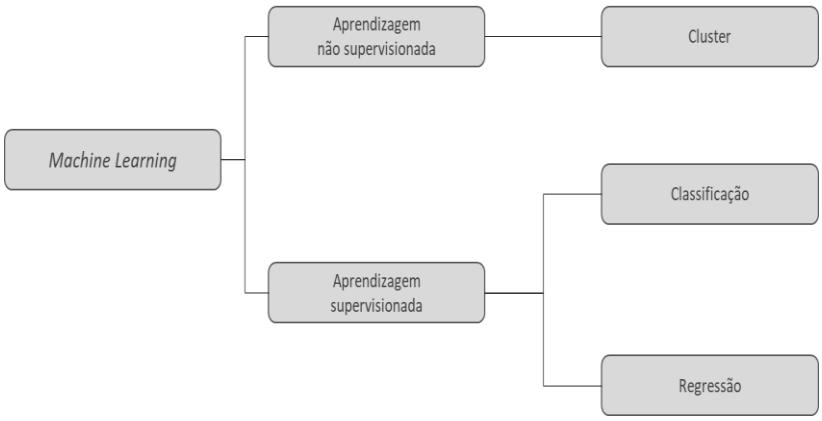


Figura 2.1: Tipos de aprendizagem (Mathworks, 2016)

2.2.1. Aprendizagem supervisionada

Aprendizagem supervisionada, de acordo com Shobha & Rangaswamy (2018), é um modelo de aprendizagem criado para fazer previsões, dado um conjunto de instâncias do problema, caracterizadas por determinadas variáveis de entrada e com uma variável de saída que reflete o objetivo do problema. Um algoritmo de aprendizagem supervisionada utiliza um conjunto de dados de entrada e suas respostas conhecidas aos dados (saída) para aprender o modelo de regressão ou classificação. Ou seja, assume a presença de um "professor" (variável de saída), onde são fornecidas as respostas corretas para cada situação. A aprendizagem é realizada a partir de exemplos (instâncias ou casos de treino) compostos por um vetor de entrada e por um vetor de saída.

2.2.2. Aprendizagem não-supervisionada

Neste tipo de aprendizagem, segundo Cheeseman & Stutz (1990), o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados formando *clusters*. A aprendizagem é realizada pela descoberta de regularidades (semelhanças) nos dados de entrada, procurando agrupamentos (*clustering*) dos exemplos de treino. Um exemplo na área de E&P é a amostra de perfuração em poços de O&G (Figura 2.2). Neste caso, o algoritmo vai procurar agrupamento a partir das características dos fluidos que podem ser encontrados nos relatórios diários de perfuração.

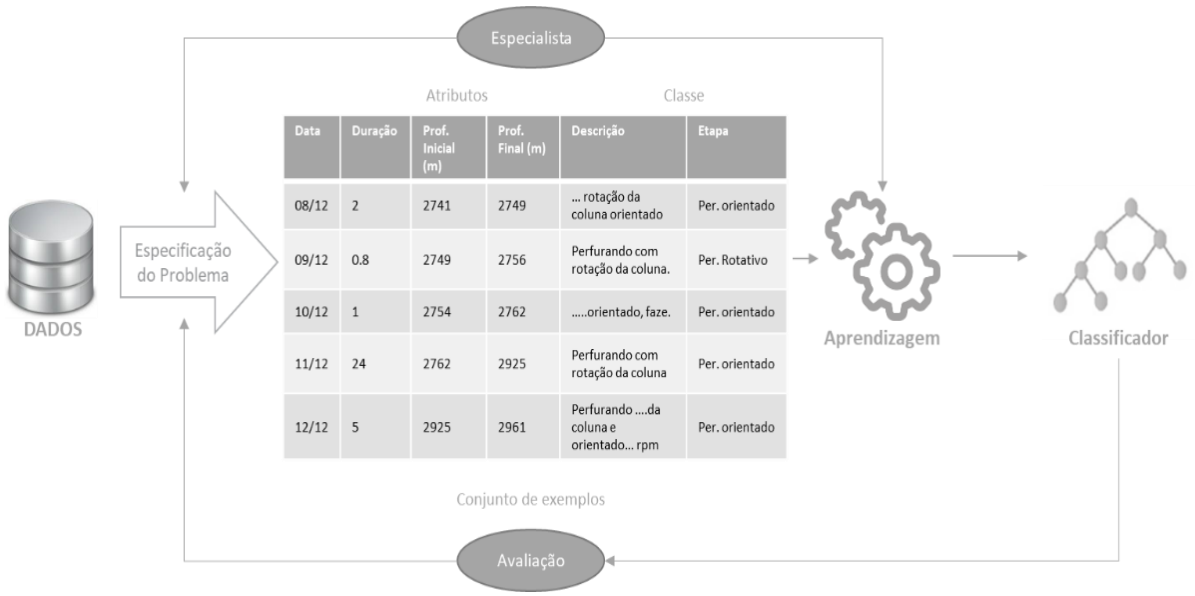


Figura 2.2: Exemplo de paradigma de aprendizagem, adaptado de (Monard & Baranauskas, 2003)

Han et al. (2011), afirmam que as técnicas mais populares que podem ser usadas em problemas de ML são: Redes Neuronais Artificiais (ANN), Máquinas de Vetores de Suporte (SVM), logica fuzzy, Algoritmos Genéticos e K-Nearest Neighbor (K-NN).

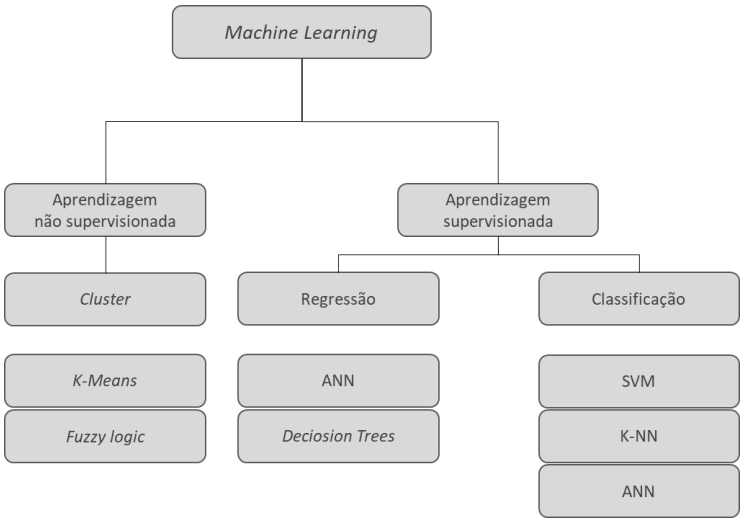


Figura 2.3: Técnicas de Machine Learning (Mathworks, 2016)

2.3. Algoritmos de inteligência artificial

A inteligência artificial tem sido usada em muitas áreas de aplicações na solução de problemas de classificação, diagnóstico, seleção e previsão. As técnicas dos modelos capturam a incerteza entre os cenários reais de causa e efeito, incorporando Episteme disponível com probabilidades e inferência de probabilidades (Han et al., 2011).

2.3.1. K-Means

O algoritmo *k-means* define o centroide de um *cluster* como o valor médio dos pontos no *cluster*. Primeiro, seleciona aleatoriamente *k* dos objetos em *D*, cada um dos quais, inicialmente, representa uma média ou centro de *cluster*. Para cada um dos objetos restantes, é atribuído ao *cluster* o mais semelhante, com base na distância euclidiana entre o objeto e a média do *cluster* (Han et al., 2011).

Onde:

k: número de clusters,

D: um conjunto de dados contendo *n* objetos.

2.3.2. Logica Fuzzy

A lógica *fuzzy* foi introduzida nos meios científicos, em 1965, por Lofti Asker Zadeh, por meio da publicação do artigo *Fuzzy Sets no Jornal of Information and Control*. Diferente da Lógica Booleana, que admite apenas valores booleanos, ou seja, verdadeiro ou falso, Han et al. (2011) afirmam que a lógica difusa ou *fuzzy*, trata de valores que variam entre 0 e 1. Assim, uma pertinência de 0.5 pode representar meio verdade, logo 0.9 e 0.1, representam quase verdade e quase falso, respetivamente. Com a necessidade de lidar com a complexidade dos problemas, a teoria da probabilidade era usada com sucesso em muitas áreas da ciência. Porém, com essa teoria era mais difícil tratar a incerteza. Um exemplo disso era considerar o período meia-idade que começa aos 35 anos e termina aos 55 anos (Gabril et al., 2011).

2.3.3. Árvores de Decisão (*Decision Trees*, DT)

Uma árvore de decisão (DT), utiliza uma estratégia de dividir para conquistar. DT são algoritmos de decisão que pretendem modelar uma variável dependente, a partir de um conjunto de variáveis explicativas (independentes) que vão sendo divididas sequencialmente, dando origem a ramos que melhor discriminam as classes de um problema. Este método classifica uma população em segmentos semelhantes a ramificações que constroem uma árvore invertida com um nó raiz, nós internos e nós folhas (Song & Lu, 2015).

2.3.4. Máquinas de Vetores de Suporte (SVM)

Máquinas de Vetores de Suporte (SVM), é um método de IA muito poderoso usado para resolver problemas de reconhecimento de padrões de duas classes. Analisa os dados e identifica padrões para fazer uma classificação (Mushtaq & Mellouk, 2017). Os métodos variam na estrutura e nos atributos do classificador. Gove & Faytong (2012), afirmam que o SVM mais conhecido é um classificador linear, prevendo a classe de membro de cada entrada entre duas classificações possíveis. Uma definição mais precisa indicaria que uma SVM constrói um híper plano ou conjunto de híper planos para classificar todas as entradas num espaço de alta dimensão ou mesmo infinito (Gove & Faytong, 2012). Os valores mais próximos da margem de classificação são conhecidos como vetores de suporte. O objetivo do SVM é maximizar a margem entre o híper plano e os vetores de suporte.

2.3.5. Algoritmos Genéticos (GA)

Um Algoritmo Genético (GA) é um algoritmo de otimização de busca, baseado na mecânica do processo de seleção natural (Holland, 1992). O conceito básico desse algoritmo é imitar o conceito de "sobrevivência do mais apto"; simula os processos observados num sistema natural em que os fortes tendem a adaptar-se e sobreviver, enquanto os fracos tendem a parecer (Ab Wahab, Nefti-Meziani, & Atyabi, 2015).

2.3.6. K-Nearest Neighbor (K-NN)

O método K-NN tem sido amplamente utilizado na área de reconhecimento de padrões. Os classificadores K-NN são baseados na aprendizagem por analogia, comparando uma determinada tupla de teste com as tuplas de treinamento semelhantes a ela. As tuplas de treinamento são descritas por *n* atributos. Cada tupla representa um ponto num espaço *n*-dimensional. Dessa maneira, todas as tuplas de treinamento são armazenadas num espaço padrão *n*-dimensional. Quando uma tupla desconhecida é fornecida, um classificador *k*-vizinho mais próximo, pesquisa no espaço padrão as *k* tuplas de treinamento mais próximas da tupla desconhecida. Essas *k* tuplas de treinamento são os *k* "vizinhos mais próximos" da tupla desconhecida (Han et al., 2011).

2.3.7. Redes Neuronais Artificias (ANNs)

O modelo de ANN tenta imitar processos simplificados de aprendizagem biológica e simular algumas funções do sistema nervoso humano. Essa técnica inteligente e adaptativa possui um sistema de processamento de informações paralelo que pode desenvolver associações, transformações ou mapeamentos entre objetos ou dados. Uma ANN consiste em unidades de processamento simples chamadas neurónios. Deve-se notificar que a abordagem de ANN não usa um algoritmo pré-descrito para solucionar um problema. Aprende o modelo da solução automaticamente, treinando algumas entradas e resultados esperados (Bishop, 1996). As especificações das ANN’s são descritas nos próximos subcapítulos.

2.3.8. Modelo de ANNs

Antes de mais, serão introduzidas o modelo simplificado de um neurónio e as capacidades de processamento associadas. O modelo de um neurónio artificial de Mcculloch & Pitts (1943), tenta simular as realidades biológicas que ocorrem dentro de uma célula do sistema nervoso (figura 2.4). A informação fornecida por outros neurónios entra em $\mathbb{D}$ entradas x_y (=sinapses) no neurónio processador. O processamento consiste numa combinação linear das entradas.

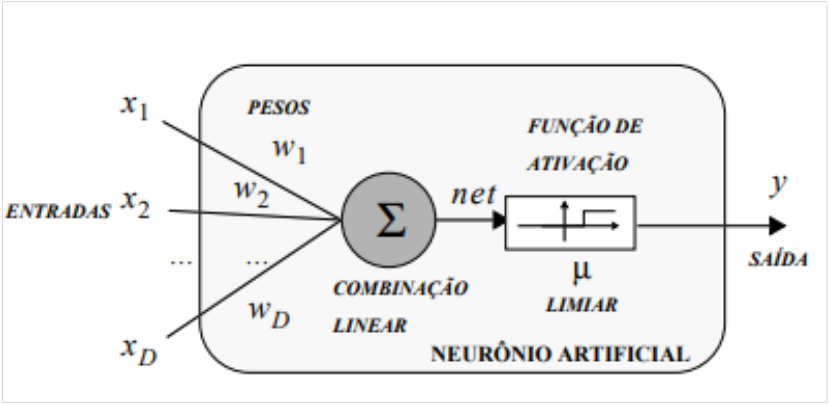


Figura 2.4: Modelo de ANN (Mcculloch & Pitts, 1943)

2.3.9. Topologia de Redes Neuronais Artificias

A topologia de ANN, de acordo com Rauber (2014), pode distinguir as redes recorrentes (*recurrent*) e as redes de propagação para frente (*feedforward*). No caso de rede redes neuronais recorrentes (RNN), segundo Elman (1990), são um tipo de ANN projetada para reconhecer padrões em sequências de dados, como texto, caligrafia, palavra falada ou dados de séries numéricas que emanam de sensores, bolsas de valores e agências governamentais. Esses algoritmos consideram tempo e sequência, têm uma dimensão temporal. Este modelo analisa um texto palavra por palavra e armazena a semântica de todo o texto anterior numa camada oculta de tamanho fixo. Um exemplo de uma rede RNN, é a *Long Short-Term Memory* – LSTM.

2.3.10. LSTM

De acordo com Zaremba, Sutskever, Vinyals, & Brain (2015), LSTM é uma arquitetura de RNN que "lembra" valores em intervalos arbitrários. A LSTM é bem adequada para classificar, processar e prever séries temporais com intervalos de tempo de duração desconhecida. Foi introduzida inicialmente por Hochreiter & Schmidhuber (1997) e funcionam muito bem numa grande variedade de problemas, sendo atualmente muito utilizadas.

A dinâmica RNN pode ser descrita usando transições determinísticas do anterior para o estado atual oculto. O LSTM possui dinâmicas complicadas que permitem "memorizar" informações com facilidade por um longo período. A memória de "longo prazo" é armazenada num vetor de células de memória. Embora muitas arquiteturas LSTM que diferem na sua estrutura de conectividade e funções de ativação,

todas as arquiteturas LSTM possuem células de memória explícita para armazenar informações por longos períodos. O LSTM pode decidir substituir a célula de memória, recuperá-la ou mantê-la na próxima etapa (Zaremba et al., 2015). A transição determinística do estado é uma função:

$$LSTM = h_t^{l-1}, h_{t-1}^l, c_{t-1}^l \rightarrow h_{t-1}^l c_t^l \quad (2.1)$$

A Figura 2.5 ilustra as equações de LSTM:

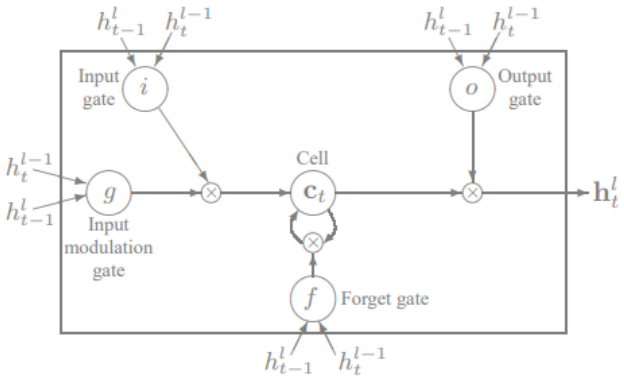


Figura 2.5: Gráfica das células de memória LSTM, Fonte: (Zaremba et al., 2015)

Nas redes de propagação para frente, o fluxo de informação é unidirecional. Neurônios que recebem a informação simultaneamente agrupam-se em camadas. Camadas que não estão ligadas às entradas nem às saídas da rede chamam-se camadas ocultas. Exemplos para esse tipo de rede são o *perceptron*, o *perceptron* multicamada (MLP), e o ADALINE (Haykin, 2001).

2.3.11. ANN – Perceptron

No final da década 1950, Rosenblatt (1958), na Universidade Cornell, deu seguimento às ideias de McCulloch. Criou uma genuína rede de múltiplos neurónios do tipo discriminadores lineares e chamou esta rede de *perceptron*. Um *perceptron* é uma rede com os neurónios dispostos em várias camadas (Figura 2.6). Os neurónios que recebem diretamente as entradas da rede constituem o que se chama de camada de entrada. Os neurónios que recebem como entradas as saídas daquela camada de entrada constituem a segunda camada e assim sucessivamente até a camada final que é a camada de saída. As camadas internas que não são, nem a de entrada, nem a de saída, são geralmente referidas como camadas ocultas.

Fórmula de ***Perceptron***: $d(x) = sgn(\sum_j^D \quad = 0^W j^X j)$ (2.2)

A função calculada $d(x)$ fornece uma resposta de "que lado" esta o objeto $x = (x_1, x_2)^T$, assim permitindo uma classificação linear entre duas classes.

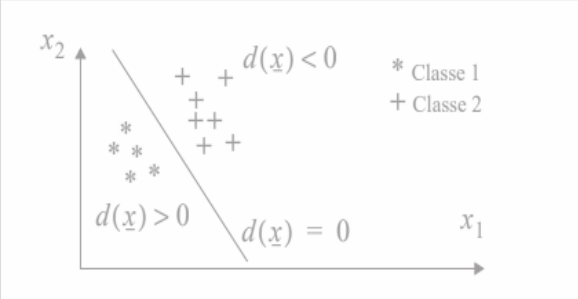


Figura 2.6: Perceptron

Rosenblatt (1958), afirma ainda que o *perceptron* é capaz de classificar entre duas classes que linearmente são separáveis. A Figura 2.6 ilustra como o modelo é extensível facilmente para o caso de várias classes. A função que o *perceptron* implementa é a do neurónio de Mcculloch & Pitts (1943), em que a função de escada é substituída pela função do sinal.

A classificação é uma das aplicações principais do cálculo que as ANNs são capazes de realizar. O objetivo é associar uma categoria de um universo finito a um objeto. Exemplos para classificação são:

- reconhecimento automático de caracteres;
- deteção de falhas em processos;
- identificação de pessoas por impressões digitais, voz, iris do olho;
- diagnóstico médico.

2.3.12. ANN – ADALINE

Segundo Kovács (2002), na mesma época em que Rosenblatt (1958), trabalhava no *perceptron*, Widrow & Hoff (1960), na Universidade de Stanford, desenvolveu um modelo neuronal linear muito simples, que batizou de ADALINE (acrónimo do inglês *Adaptive Linear Element*). O ADALINE é muito parecido com o *perceptron* em termos de arquitetura, porém a diferença está nas saídas contínuas. Permite-se calcular valores de saída do domínio dos números reais (Figura 2.7). A função calculada é simplesmente a combinação linear dos pesos e das entradas ou, equivalentemente, o produto interno do vetor de pesos e o vetor das entradas:

$$\text{Fórmula de ADALINE: } d(x) = \sum_j^D = 0^W j^X j = w^T x \qquad (2.3)$$

Abaixo a ilustração de entrada com um valor constante de 1, que facilita a representação da função calculada:

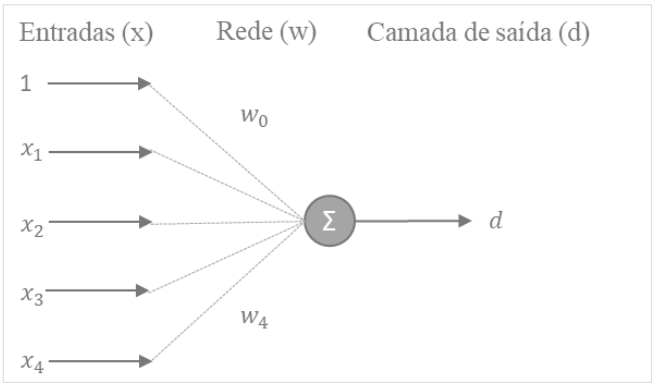


Figura 2.7: Figura de ADALINE com 4 variáveis (X4) de entrada

A contribuição realmente importante do trabalho de Widrow & Hoff (1960), foi a invenção de um princípio de treinamento extremamente poderoso para as redes de ADALINE conhecidos como regra Delta, que foi mais tarde generalizado para redes neuronais mais elaborados (Kovács, 2002).

2.3.13. Multilayer Perceptron (MLP)

Segundo Haykin (2001) e Guilherme et al. (2010), a arquitetura MLP é mais utilizada e tem sido utilizada na solução de diversos problemas da engenharia. A arquitetura MLP consiste de um conjunto de unidade de sensores constituindo a camada de entrada, uma ou mais camadas ocultas, e uma camada de saída. A Figura 2.8 ilustra um exemplo de funcionamento de uma rede MLP.

$$\text{Fórmula de MLP: } d_i(x) = g\left(\sum_{h=0}^{H} w_{ih}\, g\left(\sum_{j}^{D} w_{hj} x_j\right)\right) \qquad (2.4)$$

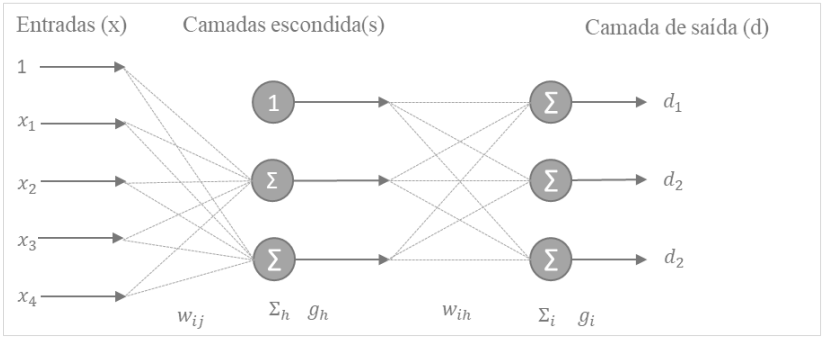


Figura 2.8: Perceptron multi-camada

Uma rede MLP tem três características distintas:

- o modelo de cada neurónio na rede inclui uma função de ativação não linear;
- a rede contém uma ou mais camadas ocultas que não são partes da camada de entrada e saída.
- essas camadas ocultas fazem com que a rede aprenda tarefas mais complexas;

A rede exibe um alto grau de conectividade, determinada pelas sinapses da rede. Uma mudança na conectividade da rede requer uma mudança nas conexões sinápticas ou nos pesos sinápticos.

O modelo MLP utiliza o processo de aprendizagem supervisionada, sendo que o algoritmo utilizado é o *error back-propagation algorithm*. Esse algoritmo é baseado na *error-correction learning rule*[3].

[3] Regras de aprendizagem que corrigem erros são mais frequentemente usadas em simulações cognitivas e nas aplicações tecnológicas de redes neurais. A regra Delta é um exemplo (Widrow & Hoff, 1960)

2.4. Evolução de IA na indústria de E&P

Ao longo dos anos IA tem causado um grande impacto na indústria de E&P e a sua aplicação tem continuado a crescer na indústria de *Oil&Gas* (Bello, Holzmann, Yaqoob, & Teodoriu, 2015). Segundo Bello et al. (2015), a aplicação de IA na indústria de E&P tem mais de 30 anos de história, desde a sua primeira implementação em 1989, sendo voltada para interpretação de *logs* de poços, diagnóstico de brocas, usando Redes Neuronais e interfaces inteligentes de simuladores de reservatórios.

Rable (2017) afirma que os valores mensuráveis da IA incluem: a possibilidade de fazer máquinas resolverem problemas difíceis a uma velocidade mais rápida do que o cérebro humano precisaria para resolver o mesmo problema; a análise de *big data* para entender tendências e fazer previsões de cenários futuros da melhor maneira possível com o menor desperdício de tempo e esforço; e a consequência disso é uma enorme economia no custo operacional. Praticamente todas as partes da cadeia de E&P aplicaram uma técnica de IA ou outra no curso de suas operações (Xia, Xie, Zhang, & Caulfield, 2013).

2.5. Exploração, desenvolvimento e abandono

Antes de perfurar um poço de exploração, a empresa operadora do consórcio terá de obter uma licença de exploração. Antes de solicitar essa licença, os geólogos analisam quaisquer dados sísmicos adquiridos, a geologia regional da área e, finalmente, levam em conta as informações disponíveis sobre campos análogos e eventuais poços perfurados na vizinhança. Na fase de exploração é também realizada uma estimativa de valor esperado do prospeto de exploração com base numa avaliação dos volumes *in place* (HIIP) e numa probabilidade de sucesso (POS) estimados pelas disciplinas de geologia.

Os custos de exploração são estimados no sentido de perceber se o potencial do prospeto combinado com a incerteza geológica é suficiente para se decidir em avançar com a perfuração de um poço de exploração (Shepherd, 2009). No caso de sucesso na exploração (i.e. descoberta de hidrocarbonetos), haverá lugar à perfuração de outros poços no sentido de avaliar a extensão do reservatório e confirmar a comercialização do campo (Shepherd, 2009). Em caso de ser viável, o projeto segue para a fase de desenvolvimento onde serão perfurados os poços de desenvolvimento que irão produzir os hidrocarbonetos ou injetar os fluidos (água ou gás) para manutenção de pressão e gestão do reservatório (Heriot Watt, 2013).

Na decisão de avançar e perfurar um poço de exploração, é preparada uma proposta e os objetivos deste poço serão: determinar a presença de hidrocarbonetos, fornecer dados geológicos (núcleos, *logs*) para avaliação, e testar o poço para determinar o seu potencial de produção e obter amostras de fluido.

O ciclo de vida de um campo de *Oil&Gas* (O&G), de acordo com Heriot Watt (2013), pode ser subdivido nas seguintes fases: exploração, avaliação, desenvolvimento, manutenção e abandono.

2.6. Perfuração de um poço de petróleo

De acordo com Brice (2009), o primeiro poço de petróleo foi perfurado em 1859 por Edwin Laurentine Drake, mais conhecido por coronel Drake em Titusville, Pensilvânia EUA. Este poço tinha menos de 100 pés de profundidade e produzia cerca de 50 bpd (barril de petróleo por dia). O método *cable-tool* de perfuração foi usado para perfurar este primeiro poço. O termo *cable-tool* é usado para descrever a técnica na qual um cinzel[4] é suspenso da extremidade de um fio cabo e é feito para impactar repetidamente a formação na parte inferior do furo (mais detalhes sobre a técnica *cable-tool* pode ser encontrada na literatura Heriot Watt, 2013). Quando a rocha no fundo do poço estiver desintegrada, a água é derramada pelo mesmo e um balde cilíndrico longo (fiador) é escorrido pelo furo para recolher as lascas de pedra. A perfuração com *cable-tool* foi usada até à década de 1930 para alcançar profundidades de 7500 pés.

Na década de 1890, foram introduzidas as primeiras plataformas de perfuração rotativa (Figura 2.9). A perfuração rotativa é a técnica pela qual a ferramenta de corte da rocha é suspensa na extremidade do tubo oco, para que o fluido possa circular continuamente pela face do *drill bit*[5], limpando o material de perfuração (*cuttings*) da face da broca e carregando-a para a superfície. Esse processo é muito mais eficiente do que a técnica *cable-tool* (Heriot Watt, 2013). A perfuração de um poço requer muitas habilidades diferentes e envolve muitas empresas. A empresa de petróleo que gere as operações de perfuração e/ou produção é conhecida como operador. Nas *joint ventures*[6], uma empresa atua como operadora em nome de outros parceiros.

[4] Utensílio ou peça cortante numa das pontas, utilizada essencialmente para esculpir, lavorar, talhar ou facetar metais ou pedras.

[5] *Drill bit* é uma ferramenta conectada na extremidade inferior da coluna de perfuração utilizada para cortar mecanicamente a rocha na perfuração de poços de óleo e gás, composta por elementos cortantes e por um sistema de passagem de fluido, o qual permite que o fluido de perfuração seja injetado através de jatos gerando uma perda de carga, a qual é convertida em potência hidráulica proporcionando a limpeza no fundo do poço.

[6] *Joint venture* é uma expressão de origem inglesa, que significa a união de duas ou mais empresas já existentes com o objetivo de iniciar ou realizar uma atividade econômica comum, por um determinado período de tempo e visando, dentre outras motivações, o lucro (Kogut, 1988).

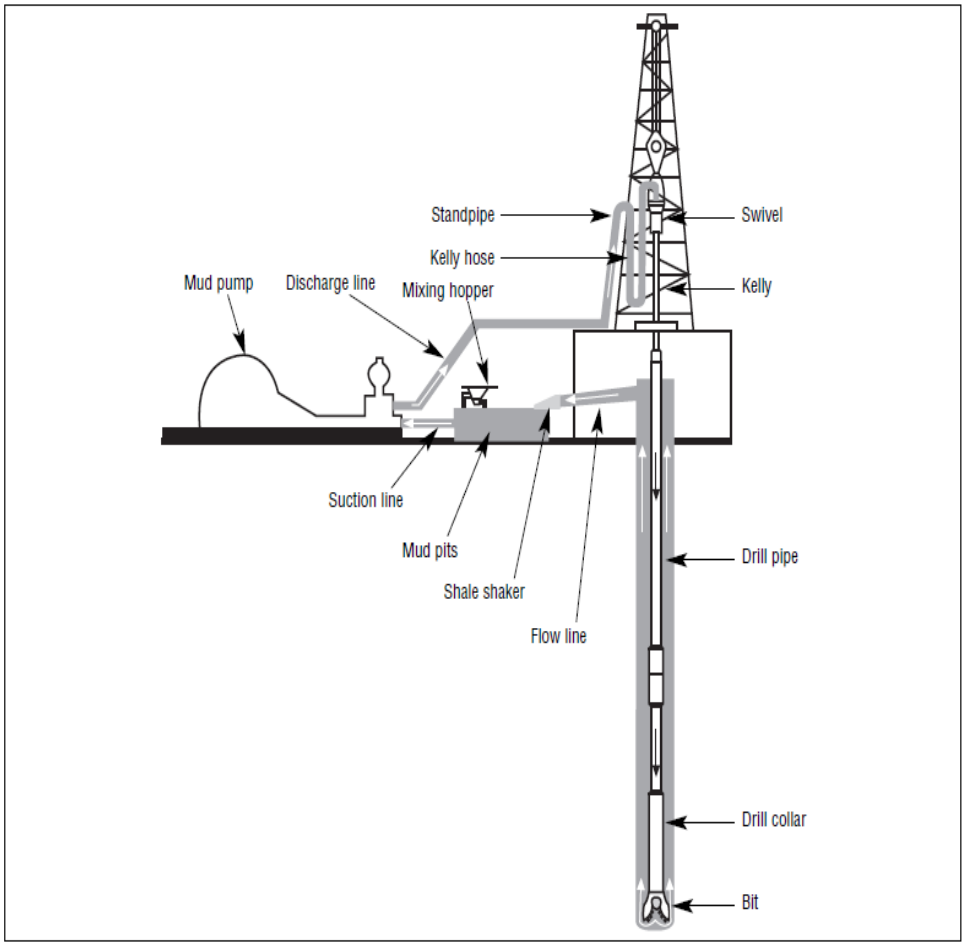


Figura 2.9: equipamento de perfuração rotatória (Thomas, 2001)

Tabela 2.2: Legenda de equipamentos de perfuração (Heriot Watt, 2013)

Função	Definição
Standpipe	É um tubo de parede pesado, preso a uma das pernas da torre. Conduz lama de alta pressão das bombas para a mangueira rotativa.
Kelly	É um tubo de aço quadrado ou hexagonal pesado que atravessa a mesa rotativa e é usado para girar a coluna de perfuração.
Kelly hose	É um tubo flexível reforçado que conduz o fluido de perfuração do tubo vertical à articulação giratória. Também chamado de "*Rotary hose*" ou "*mud hose*".
Swivel	É um componente suspenso do gancho. Permite que a lama flua da mangueira rotativa através do *swivel* para o *kelly* enquanto a coluna de perfuração está girando.
Mud pump	É uma bomba alternativa usada para circular o fluido de perfuração pelo poço. Bombas de lama também são chamados de "*slush pumps*".
Discharge line	É uma linha de descarga é uma seção da tubulação em que a pressão é superior à pressão atmosférica (por exemplo, sistema de bomba).
Mixing hopper	É um equipamento especial para preparar e aumentar o peso dos fluidos de perfuração adicionando e misturando bentónica, altera a densidade do fluido, altera a densidade da lama, viscosidade e desidratação.
Suction line	É a parte do sistema de tubulação de uma unidade de ac/ref. Depois que o refrigerante evapora para um gás na bobina do evaporador, a seção da tubulação dessa bobina para a bobina do condensador é chamada de linha de sucção.
Mud pits	É uma série de tanques abertos nos quais a lama é misturada e condicionada. As plataformas modernas são fornecidas com três ou mais cavidades, geralmente feitas de chapa de aço com tubulação, válvulas e agitadores embutidos.
Shale shaker	É uma série de bandejas com peneiras vibratórias que permitem a passagem da lama, mas retêm as estacas. A malha deve ser escolhida cuidadosamente para corresponder ao tamanho dos sólidos na lama.

Flow line	Uma linha de fluxo, usada em uma sonda de perfuração, é um tubo de grande diâmetro que é conectado ao bocal do sino e se estende ao ventre do gambá e atua como uma linha de retorno aos tanques de lama.
Drill pipe	É um tubo sem costura pesado que é usado para girar a broca e circular o fluido de perfuração.
Drill collar	É um tubo de aço pesado com paredes grossas que fornece peso à broca para obter penetração.
Bit	É um elemento de corte na parte inferior da coluna de perfuração, usado para perfurar a rocha.

2.7. O Processo de perfuração

A perfuração de um poço de petróleo é feita a partir da utilização de uma sonda de perfuração que, por definição, consta como um conjunto de sistemas, equipamentos e ferramentas (ex: torre de perfuração, sistema de cargas, sistema hidráulico, sistema de potência e energia, brocas, tubos de diferentes diâmetros e comprimentos, ferramentas de geonavegação, ferramentas de avaliação) que tem como principal finalidade perfurar diversas camadas de rochas, até encontrar um reservatório de petróleo e, através deste caminho criado através das camadas geológicas (poço), conectar o reservatório à superfície, permitindo assim a correta drenagem dos hidrocarbonetos e futura comercialização destes fluidos (Thomas, 2001).

A perfuração de poços evoluiu muito com o passar dos anos, de acordo com Thomas (2001) mas, de forma geral, consistem na aplicação de peso e rotação numa broca, posicionada no fim de um conjunto de tubos interconectados (coluna de perfuração), utilização de um fluido de perfuração (lama) para arrefecimento da broca, transporte dos *cuttings* para a superfície e balanceamento da pressão das formações. Posteriormente, realiza-se o revestimento da formação perfurada e a cimentação do mesmo. O processo descrito acima é realizado em diferentes fases onde, a cada uma delas, reduz-se o tamanho da broca a ser utilizada e o diâmetro do tubo de revestimento a ser instalado.

A descrição seguinte é apenas uma visão geral do processo de perfuração de um poço (o processo de construção).

2.7.1. Primeira fase - Instalação do revestimento de 30” (Condutor)

O primeiro estágio da operação de acordo com Heriot Watt (2013) é consistente em conduzir um tubo de grande diâmetro a uma profundidade de aproximadamente 100 pés abaixo do nível do solo (ou assoalho oceânico no caso de perfurações marítimas). Este tubo, geralmente chamado de revestimento condutor é instalado para impedir que as formações com sedimentos não consolidados colapsem durante a perfuração. Pode ser assentado por cravação, uso de jatos (no mar) ou perfuração com uma broca de 36”.

2.7.2. Segunda fase – Instalação do revestimento de superfície

A segunda fase é perfurada com uma broca de diâmetro inferior ao diâmetro interno (DI) do condutor (já que o conjunto de perfuração é descido por dentro do revestimento de 30”) e é normalmente realizada

com uma broca de diâmetro externo de 26". Esta fase visa proteger aquíferos superficiais e prevenir desmoronamento de formações não consolidadas (Rowsell & Waller, 1994).

2.7.3. Terceira fase – Instalação do revestimento intermediário

Esta fase tem a principal função de isolar zonas de alta ou baixa pressão, onde podem acontecer fenómenos indesejados como: desmoronamento, corrosão devido a fluidos muito abrasivos, perda de circulação, formações contendo gases sob alta pressão, entre outros (Thomas, 2001).

Para a perfuração desta fase, o sistema de segurança é indispensável. Trata-se de um equipamento de segurança crítico designado por BOP (*Blow Out Preventor*), que tem como principal função impedir que qualquer influxo de fluido que tenha vencido a barreira do fluido de perfuração (devido a pressão), venha para a superfície. O BOP é basicamente um conjunto de válvulas e "gavetas", que atuam no sentido de fecho parcial ou completo do poço, numa situação de emergência (Schlumberger, 2019).

Nesta fase, de acordo Heriot Watt (2013), é perfurada com uma broca de 17 ½" e revestido com tubos de 13 3/8" e a sua cimentação pode ser total, até à superfície, ou parcial.

2.7.4. Quarta fase – Instalação do revestimento de produção

Esta fase, de acordo com Thomas (2001), é a última fase de perfuração e revestimento e tem como principal função permitir a produção do poço, suportando as suas paredes e possibilitando o isolamento entre os vários intervalos de reservatório.

O diâmetro da broca para esta secção é muito variável, tendo em conta o tipo de reservatório e as condições de produção, podendo ser de 10 3/4" ou 8 ½", com um revestimento de 9 5/8" ou 7". A depender do tipo de rocha ou grau de consolidação, este revestimento pode não ser necessário, o que denomina um "poço aberto". Ou seja, sem o revestimento final na zona de reservatório (Heriot Watt, 2013).

2.7.5. Completação de um poço

A completação é um termo genérico usado para descrever a montagem de tubagem e equipamentos de fundo de poço necessários para permitir a produção segura e eficiente de um poço de petróleo ou gás. O ponto em que o processo de completação começa pode depender do tipo e *design* do poço. No entanto, existem muitas opções aplicadas ou ações executadas durante a fase de construção de um poço que têm um impacto significativo na produtividade do mesmo (Schlumberger, 2019).

De um modo geral, a completação e os seus equipamentos têm como finalidade principal permitir a produção de hidrocarbonetos de forma segura, rápida e eficiente.

2.8. Principais Problemas de Perfuração

Conforme mencionado anteriormente, a perfuração de um poço de petróleo é uma operação bastante complexa e está sujeita a ocorrência de uma série de eventos indesejáveis. A evolução da técnica de

perfuração e o uso de boas práticas têm proporcionado redução na frequência de ocorrência de problemas ao longo dos anos, mas, ainda assim, alguns desafios persistem. Essas complicações que podem ocorrer durante a perfuração são altamente indesejáveis e podem causar prejuízos avultados devido a: contratação de tempo adicional de sonda, o atraso no cronograma de execução do projeto e, em casos extremos, a perda do poço e necessidade de perfurar novamente.

Seguidamente são apresentadas algumas das situações apontadas por Heriot Watt (2013), de eventos que podem ocorrer durante a operação de perfuração de um poço de petróleo. As situações apresentadas dizem respeito à execução da operação propriamente dita.

2.8.1. Perda de circulação

A perda de circulação consiste na invasão de fluido de perfuração para a formação através de fraturas existentes ou provocadas em formações com alta permo-porosidade ou em zonas despressurizadas. A perda de circulação pode ser total ou parcial. Total quando não ocorre retorno de fluido para a superfície e parcial quando ocorre retorno parcial do fluido para a superfície.

A perda de circulação total provoca instabilidade mecânica no poço devido à redução da pressão hidrostática com a queda do nível de fluido de perfuração no anular e, consequentemente, desmoronamento das camadas superiores ou inferiores à zona de perda podendo, inclusive, permitir a invasão indesejada dos fluidos da formação para dentro do poço (*kick*), pondo em risco a segurança do mesmo.

2.8.2. Prisão da coluna de perfuração

A prisão da coluna de perfuração pode ocorrer devido a problemas operacionais (como a incapacidade de se identificar os sinais emitidos pelo poço), durante a parada da coluna de perfuração ou durante as manobras. Além disso, em situações de movimento vertical de subida e descida durante a retirada da coluna de perfuração pode ocorrer a prisão da coluna (Bradley et al., 1991). As anormalidades que podem levar à prisão de uma coluna de perfuração podem ser encontradas com mais detalhes no glossário *Oilfield* da Schlumberger (2007).

2.8.3. Desmoronamento de Poço

O desmoronamento do poço é caracterizado pela queda de blocos ou fragmentos que se desprendem após a passagem da broca de perfuração (Schlumberger, 2019).

Uma das principais causas do desmoronamento é a insuficiência de pressão hidrostática no interior do poço. Devido à falta de sustentação, as paredes do poço acabam por colapsar e o interior do mesmo é invadido por esse material que se desprende da formação

2.8.4. Alargamento do Poço

O alargamento do poço é caracterizado por um aumento não desejado do diâmetro do poço. O mesmo pode ser provocado por desmoronamentos localizados e influxos de fluidos para dentro do poço. O alargamento pode ser também causado pela erosão da formação provocada pelo regime de fluxo turbulento ou mesmo por atrito mecânico da coluna com as paredes do poço (Schlumberger, 2007).

2.8.5. Influxo de fluidos indesejados (*Kick)*

O *kick* é o influxo indesejado de fluidos presentes na formação (água, gás ou óleo) para dentro do poço devido à existência de uma pressão hidrostática no interior do mesmo, insuficiente para conter a pressão da formação. Quando o mesmo ocorre de maneira descontrolada e atinge a superfície chama-se *blowout* (Heriot Watt, 2013). As causas mais prováveis de um *kick* podem ser encontrados no glossário *Oilfield* da Schlumberger (2007).

2.8.6. Quebra de BHA (Vibração)

A vibração da coluna de perfuração é a principal causa de falha mecânica do BHA. A vibração no sentido axial, conhecida como "*bit bouncing*", é mais frequente em brocas tricónicas. Essa vibração faz com que a broca seja elevada, perdendo contato com a rocha e, logo de seguida, seja baixada e se choque contra a rocha. Esse ciclo pode repetir-se até nove vezes por revolução. A consequência provocada é a redução na taxa de penetração e possíveis danos ao BHA (Thomas, 2001).

2.8.7. Pack-*off*

O *Pack-off* ocorre devido à entrada de cascalhos nos jatos quando a circulação é interrompida. A presença dos cascalhos pode reduzir o diâmetro útil do jato e até mesmo bloquear por completo a passagem do fluido de perfuração (Schlumberger, 2019).

2.8.8. Bit Balling

O *bit balling* é um problema que ocorre devido à hidratação de argilas. Os cascalhos argilosos aderem à superfície da broca de forma a não poderem ser removidos pela circulação do fluido de perfuração (Schlumberger, 2007).

Com o agravamento dessa condição, a broca é totalmente coberta por cascalhos e perde o contato com a formação. O principal resultado do enceramento pode ser percebido através da queda na taxa de penetração que, em condições severas de enceramento, torna-se praticamente nula (Heriot Watt, 2013).

2.8.9. Washout

De acordo com Schlumberger (2019), *Washout* é o nome dado a um pequeno vazamento no sistema de circulação causado por um furo na coluna de perfuração. Esse furo pode surgir, por exemplo, na

extremidade do tubo devido ao desgaste das conexões, ou mesmo no corpo do tubo de perfuração devido a uma pequena fissura causada por fadiga ou corrosão.

2.9. Problemas onde as técnicas de IA são aplicadas na engenharia de perfuração

A classificação de textos através de técnicas de IA, gerou novos desafios para muitos autores. Estes novos desafios são focados principalmente pela forma como os relatórios são escritos. Em muitos casos não existe um padrão a ser seguido pelas empresas operadores de *Joint Venture*. Assim, e de forma a encontrar o *research gap*, foram analisados cerca de 50 artigos científicos, resultantes da filtragem conforme descritas no capítulo 1.5.1. (estratégia de pesquisa bibliográfica) com diferentes combinações de pesquisa.

De entre todos, decidiu-se optar pelos artigos mais próximos com o presente estudo, escolhendo os catorze mais relevantes. Para tal, foi criada uma tabela, com os critérios "tipo de estudo conduzido", "Autor", "Objetivo", "Método" e "Resultados". Pode-se verificar na tabela 2.3 que vários autores se focaram mais nas técnicas de regressão para resolver problemas ou estimar resultados de perfuração e nenhuma técnica de classificação de operações de perfuração é verificado conforme a proposta para este trabalho.

A Tabela 3, destaca o tipo de estudo realizado, por cada autor, sobre a técnica de IA usado, bem como a avaliação de desempenho da técnica desenvolvida em cada estudo. Isto é, representado em termos do coeficiente de correlação (R^2), erro quadrático médio da raiz (RMSE), erro relativo percentual absoluto médio (AAPE) e erro percentual absoluto médio (MAPE)). Do resumo, o seguinte é observado:

Tabela 2.3: Técnicas de IA aplicadas na engenharia de perfuração, (Fonte: Elaboração do autor)

Tipo de estudo conduzido	**Autor**	**Objetivo**	**Método**	**Resultado** (Critérios de avaliação de performance)
Determinação da alteração da densidade da lama de perfuração com pressão e temperatura simplificadas e precisas	(Osman & Aggour, 2003)	Fornecer previsões precisas da densidade da lama em função do tipo, pressão e temperatura da lama	ANN	R^2 = 0.9998
Abordagem de ANN para estimar propriedades de filtragem de fluidos de perfuração	(Jeirani & Mohebbi, 2006)	Estimar o volume de filtro e a permeabilidade do bolo de filtro usando os dados de filtragem estática.	ANN	R^2 (Volume do filtro) = 0.9815 R^2 (Permeabilidade do cake) = 0.9433
Previsão e Prevenção de Tubos Presos: Uma Abordagem de Rede Neural Convolucional	(Siruvuri, Nagarakanti, & Samuel, 2006)	apresentar uma aplicação de métodos de IA para entender e estimar a ocorrência de tubos diferencialmente presos durante a perfuração.	ANN	R^2 (Tubos presos) = 0.063 R^2 (Tubos não presos) = 0.01619
Estimando padrões de fluxo e perdas de pressão por atrito de fluidos bifásicos em poços horizontais usando ANN	(Ozbayoglu & Ozbayoglu, 2009)	Estimar os padrões de fluxo e as perdas por pressão de atrito de fluidos bifásicos que fluem através de geometrias anulares horizontais usando ANN, em vez de usar modelos mecanicistas convencionais.	ANN	MSE=0.006 [FPL with BP] MSE=0.005 [FPL with J/E] MSE=0.005 [FP with BP] MSE=0.005 [FP with J/E]
Tomada de decisão para redução do tempo improdutivo por meio de uma previsão integrada de circulação perdida	(Moazzeni, Nabaei, & Jegarluei, 2012)	Prever a gravidade da perda de lama durante a perfuração ao longo de diferentes setores do campo petrolífero.	ANN	R^2 = 0.82
Pesquisa de colagem de tubos com pré-aquecimento baseada em rede neural	(Zhu, Liu, & Zhang, 2013)	Propor o uso da tecnologia de IA para realizar o pré-aviso de acidente de tubos presos durante a perfuração.	ANN	-
Novo método para prever e resolver o problema da perfuração e perda de fluidos usando rede neural modular e enxame de partículas algoritmo de otimização	(Toreifi, Habib, Abbas, & Manshad, n.d.)	Prever a perda de circulação durante a perfuração em qualidade e quantidade.	ANN	R^2 = 0.94
Previsão em tempo real de parâmetros reológicos do fluido de perfuração à base de água KCl usando redes neuronais artificiais	(Elkatatny, 2017)	Usar as frequentes de medições de densidade da lama, viscosidade do funil de *Marsh* e percentagem sólida para prever as propriedades reológicas desenvolvendo correlações empíricas	ANN	AAPE < 6% R^2 > 0.90
Previsão e prevenção de aderência de tubulação usando modelagem lógica difusa adaptativa	(Murillo, Neuman, & Samuel, 2009)	Estimar o risco de ocorrência de tubos presos no procedimento de planeamento de poços e durante a perfuração em tempo real	Logica Fuzzy	-

Um modelo preciso para prever a densidade do fluido de perfuração em condições de poço	(Ahmadi, Shadizadeh, Shah, & Bahadori, 2018)	Sugerir um modelo preditivo rigoroso para estimar a densidade do fluido de perfuração (g / cm3) em condições de poço	Logica Fuzzy	R^2 = 0.7237 MSE = 69.0907
Uma abordagem de aprendizagem de máquina para a previsão de *settling*	(Goldstein & Coco, 2014)	Utilizar uma abordagem de aprendizagem de máquina baseada em programação genética para prever a velocidade de assentamento de partículas não coesas.	GA	RMSE = 0.26 R^2 = 0.97
Determinação ideal de parâmetros reológicos para fluidos de perfuração de espigão-sela usando GA	(Rooki et al., 2012)	Determinar comportamento reológico não newtoniano de fluidos de perfuração, a fim de determinar os três parâmetros do modelo de *Herschel-Bulkley* com mais precisão.	GA	R^2 = 0.9972
Uma abordagem de SVM para a previsão da densidade do fluido de perfuração em alta temperatura e pressão	(Wang, Pu, & Tao, 2012)	Prever a densidade do fluido de perfuração em alta temperatura e pressão (HTHP).	SVM	MAPE = 0.872 R^2 = 0.9994
Aplicação do algoritmo SVM para o cálculo da perda por pressão por atrito do fluxo trifásico em anéis inclinados	(Shahdi & Arabloo, 2014)	Uso de *Lease Square* (LS-SVM), para cálculo de perdas por atrito de fluidos de perfuração bifásicos baseados em gás, com a presença de estacas como a terceira fase na seção inclinada do poço	SVM	R^2 = 0.9862
Afinação da viscosidade e densidade de fluidos não newtonianos através do processo de mistura usando sensores multimodais, fusão de sensores e modelos	(Shahdi & Arabloo, 2014)	Estimar a velocidade em operações de perfuração	SVM	MAPE = 0.27
Estimação da densidade do fluido de perfuração na tecnologia de lama: Aplicação em poços de petróleo de alta temperatura e alta pressão	(Kamari, Gharagheizi, Shokrollahi, Arabloo, & Mohammadi, 2017)	Desenvolver um modelo confiável para prever a densidade de quatro fluidos de perfuração, incluindo à base de água, à base de óleo, *Coloidal Gás Aphron* (CGA) e sintético.	SVM	R^2 = 0.999

2.10. Conclusão do estado da arte

Das técnicas de IA acima destacadas, é pertinente questionar se há alguma que possa ser considerada 100% perfeita e adequada para uso em todas as circunstâncias. Luchian et al. (2015) e Agwu et al. (2018) afirmam que é mais benéfico focar na solução do problema do que perder tempo para encontrar o melhor método. No entanto, um teorema proposto por Wolpert & Macready (1997, p.77) denominado *No Free Lunch Theorems for Optimization* (NFLTO) afirma o seguinte: "dado que todos os problemas são considerados iguais em força e independentemente dos critérios usados para julgar o seu desempenho, todos os métodos usados para resolver o problema têm o mesmo desempenho".

Em apoio à NFLTO, Anifowose, Labadin, & Abdulraheem (2016) opinam que não existe uma única abordagem abrangente de IA que abordará efetivamente todos os desafios em todas as condições de dados e computação, uma vez que cada uma das técnicas de IA está presa nos seus pontos fortes e defeitos inevitáveis. A Tabela 2.4 resume os pontos fortes e fracos de algumas das técnicas de IA (Agwu et al., 2018).

Tabela 2.4: Comparação das técnicas de IA (Agwu, et al., 2018)

Benchmark	**ANN**	**FUZZY**	**GA**	**SVM**
robustez contra ruído	Alto	Alto	Alto	Alto
velocidade de convergência	lento	Rápido	lento	-
suscetibilidade a sob reajuste	sim	-	-	No
requisitos de volume de dados	Requer enormes dados	-	-	Requer dados pequenos
capacidade de auto-organização	Sim	-	Não	-
capacidade de generalização	Sim	-	-	Sim

Os pontos fortes e fracos das quatro técnicas de IA (ANN, lógica fuzzy, SVM e GA) foram comparados em seis critérios: robustez contra ruído, velocidade de convergência, suscetibilidade a sob reajuste, requisitos de volume de dados, capacidade de auto-organização e capacidade de generalização. Observa-se que ANN, lógica fuzzy, SVM e GA são todas robustas contra ruído, enquanto a lógica fuzzy tem uma velocidade de convergência melhor em comparação com ANN, SVM e GA. A dupla ANN e SVM tem a capacidade de generalizar enquanto a ANN pode auto organizar-se e requer enormes dados para prever bem fenómenos complexos, enquanto que o SVM exige pequenos volumes de dados.

Focando no conteúdo da literatura disponível e, de acordo com informações sobre várias aplicações da IA na engenharia de perfuração (Tabela 2.3), podemos concluir que cerca de 75% dos estudos para resolver e prever problemas que possam acontecer durante a perfuração de poços de petróleo e gás, estão relacionados com uso de ANN e SVM, sendo que, aproximadamente 50% com ANN e 25% com SVM. O restante recorre à lógica *Fuzzy* e GA com os resultados de 12.5 e 12.5%, respetivamente (Figura 2.10).

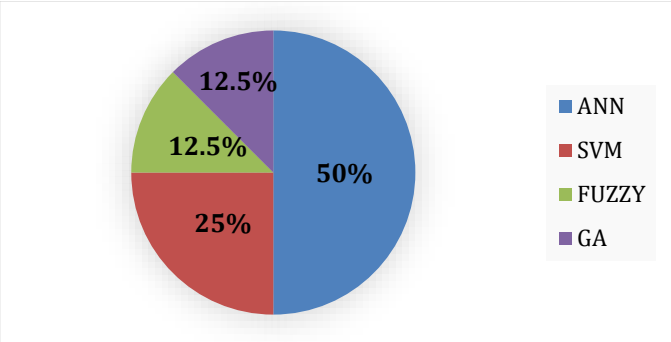


Figura 2.10: Comparação dos resultados dos modelos de IA, (Fonte: Elaboração do autor)

Conforme apresentado ao longo da revisão da literatura, poucos autores apresentaram resultados da topologia de RNA e ainda menos estudos foram realizados com dados reais extraídos de Boletins Diários de Produção (BDP).

Verificou-se também que os estudos apresentados se focam maioritariamente em estimar a velocidade em operações de perfuração e prever problemas que podem acontecer durante a perfuração de um poço de petróleo, não observando a classificação automática das operações de perfuração, de forma a estimar a duração de perfuração em poços futuros. Nesse sentido, e dos resultados obtidos, conclui-se também que os modelos RNA obtiveram melhores resultados face a SVM e, por esta razão, chegou-se à conclusão de que a técnica usada para o presente estudo será RNA. Propõe-se também um novo estudo sustentado na construção do modelo com base na topologia de RNA que apresentar melhor resultado.

Capítulo 3

3. Metodologia

Este capítulo visa descrever a metodologia de análise e *design* de projetos adotada, desde o entendimento dos problemas e do negócio, análise e tratamento de dados, desde aplicação de técnicas preditivas de DM à interpretação dos resultados, representando o tema de dissertação [Aplicação de RNAs para classificação das operações de perfuração], na qual podemos verificar se os objetivos propostos foram alcançados.

O projeto é desenvolvido seguindo a metodologia CRISP-DM, cujo contexto, problema, descrição e tratamento dos dados são explicados no decorrer do capítulo. A escolha deste modelo deve-se ao facto de ser considerado o padrão de maior aceitação e por ter sido usado para problemas semelhantes (Moro, Laureano, & Cortez, 2011). A metodologia CRISP-DM (*Cross-Industry Standard Processo of Data Mining*) é um modelo que possui processos iterativos, com sequências não mandatárias, possuindo também um ciclo de vida, que ocorre nas fases que têm as suas tarefas (Chapman, et al., 1999).

Numa 1ª fase conduziu-se uma revisão da literatura sobre as técnicas de IA para melhor entender as técnicas existentes e, de seguida, foi descrita com detalhe a técnica em estudo (RNA) para melhor entender o seu funcionamento e as suas vantagens. Para melhor compreensão sobre como é feita a perfuração e completação de um poço de petróleo e gás, foi feita uma revisão de literatura sobre o tema, apontando as suas fases, assim como os problemas que podem ocorrer durante a perfuração.

3.1. Equipamento do ambiente de execução

Tabela 3.1: Ambiente de execução do treinamento do modelo (Fonte: Elaboração do autor)

Operational System	64-bit Operating System, x64-based processor
CPU	Intel (R) Xeon (R) CPU E3-1505M v5 @ 2.80Ghz 2.80Ghz
Memory	32-GB de memória DDR3 1866 MHz
HDD	1 disco de 1-TB SSD
Display Adaptor	Intel(R) HD Graphics P530 NVIDA Quadro M2000M

A aplicação do presente trabalho foi desenvolvida em linguagem Python[7], sendo uma ferramenta *Open Source* (gratuita) e multiplataforma (disponível para Windows, Linux e Mac). É uma linguagem de programação de alto nível, interpretada, de *script*, imperativa, orientada a objetos, funcional, de tipo dinâmica e forte. O IDE[8] usado para tornar o código mais fácil de usar e mais produtivo foi o *spyder 4.0* do projeto Anaconda. Foi usado o *Keras* como ferramenta de prototipagem de alto nível para RNA e o *TensorFlow*, executando no *backend* configurado para uso de GPUs e *scikit-learn*, como ferramenta auxiliar nos algoritmos de ML tradicional como o MLP e LSTM, e a validação cruzada.

[7] https://www.python.org/

[8] IDE (*Integrated Development Environment*), ou ambiente de desenvolvimento integrado, é um software que combina ferramentas comuns de desenvolvimento em uma única interface gráfica do usuário (GUI), facilitando o desenvolvimento de aplicações (Kohn & Manaris, 2020).

3.2. Business Understanding

Entender o problema e o seu contexto, bem como os objetivos propostos é crucial para o sucesso de qualquer projeto. É fundamental compreender todos os detalhes do negócio na primeira etapa, desde identificar o problema, determinar os objetivos, avaliar a situação atual, identificar cada critério específico e como se espera que os resultados sejam obtidos, como a influência de cada um para resolver o problema (Azevedo & Santos, 2008).

Um dos objetivos deste estudo é treinar um modelo através de técnicas de RNAs para que, dado um boletim diário de perfuração (BDP), este seja capaz de classificar as operações de perfuração.

No presente trabalho são apresentados procedimentos que visam promover melhorias na operação de perfuração de poços, através da análise e interpretação dos dados de perfuração atualmente disponíveis. Durante a fase de perfuração de um poço de petróleo, compete aos engenheiros de perfuração da Galp E&P emitir um relatório diário de operações de perfuração com objetivo de controlar e monitorizar o processo de perfuração. Por meio da análise dos BDPs, é possível identificar operações que estão a consumir tempo excessivo de sonda e, a partir dessa observação, adotar medidas que melhorem a operação de perfuração. O sistema de classificação proposto identifica qual a operação que está a ser executada, através da interpretação de dados.

Atualmente, o registo das operações realizadas é feito através do BDP, um relatório preenchido diariamente na sonda que descreve as operações executadas nas últimas 24 horas. A descrição das operações é feita em linguagem natural, ou seja, o responsável pelo preenchimento elabora um pequeno texto no qual descreve de forma resumida as atividades executadas. Além da descrição, existe um sistema de codificação que permite uma classificação mais objetiva. A Tabela 3.2 apresenta um trecho do boletim de perfuração.

Tabela 3.2: Trecho de Boletim Diário de Perfuração (Tavares, 2006)

Data do Relatório	Início (h)	Duração (h)	Prof. Inicial (m)	Prof. Final (m)	Descrição	Etapa (Subcode)
08/09	14:30	2	2141	2749	Perfurando orientado	Perfurando orientado
08/09	16:30	0.5	2749	2754	Perfurando com rotação da coluna	Perfurando rotativo
08/09	17:00	1	2754	2762	Perfurando orientado	Perfurando orientado
09/09	18:00	24	2762	2925	Perfurando com rotação da coluna e orientado, fazendo back reaming antes da conexão e após perfurar trecho orientado. (20/35kip, 130/170 rpm, 450gpm pela coluna e 200 gpm p	Perfurando orientado
09/09	18:00	5	2925	2961	Perfurando com rotação da coluna e orientado, fazendo back reaming antes da conexão e após perfurar trecho orientado. (20/35kip, 130/170 rpm, 450gpm pela coluna e 200 gpm pela booster line, 3350/3420 psi, tpm = 7,2 m/h)	Perfurando orientado

Devido ao modo como as informações são registadas no BDP, os registos não são muito precisos do ponto de vista temporal. Uma das dificuldades em obter um registo temporal preciso a partir do BDP é a não uniformidade dos intervalos de classificação. Devido ao formato como os dados são disponibilizados, em PDF, os mesmo são preenchidos manualmente na sonda em linguagem natural e torna-se altamente propensos a erros e dispendioso em termos de tempo para um engenheiro de perfuração classificar as operações.

O modelo de classificação proposto para a Galp visa flexibilizar o tempo de classificação enquanto aumentam as informações e o controlo sobre a execução das operações de perfuração. Para além de reduzir o risco e custos com FTE[9], permite também evitar custos futuros desnecessários na sonda.

Tabela 3.3: Benefícios do modelo de classificação para Galp, (Fonte: Elaboração do autor)

Benefício	Descrição
Eficiência e redução de custo	**Maior capacidade de trabalho com custos mais baixos:** Um modelo de classificação representa um custo baixo comparando com um FTE
Maior qualidade e redução de riscos	**Redução de intervenção humana (i.e processamento manual de dados):** Um modelo de classificação eficiente permite a redução de erros em cerca de 90 a 99%.
Auditoria	**Todas as atividades podem ser registadas,** (tempo de execução, erros), permitindo: Explorar informações analíticas para melhorar o processo
Flexibilidade	**Possibilidade de classificar várias tarefas repetitivas.**
Motivação	**Redução de tarefas repetitivas realizadas por seres humanos:** Realinhamento estratégico de Recursos Humanos, desenvolvimento de novas competências

3.3. Data Understanding

Esta fase envolve a recolha e análise de dados, identificando subconjuntos de dados, problemas de qualidade e muitas outras características que definem seu impacto na obtenção de resultados.

Para construir o algoritmo de classificação é necessário estabelecer quais são os eventos que podem ocorrer durante a operação de perfuração. A perfuração de um poço de petróleo não é um processo contínuo constituído de uma única operação. Se examinarmos numa escala menor, é possível notar que a perfuração de um poço de petróleo é constituída de uma sequência de eventos discretos. Esses eventos serão aqui chamados de operações de perfuração.

3.3.1. Criação do *dataset*

Antes de proceder ao treinamento do modelo, é necessário recolher os dados. Esses dados foram extraídos do BDP (em formato PDF), através de um *script* desenvolvido em linguagem VBA do MS Excel para extração de textos em PDF para Excel.

Para se extrair os dados, seguiu-se o seguinte processo: primeiro abrir o ficheiro em PDF como Word, e depois é copiado o *range* em Word para uma folha em Excel adicionando nova linha de registo, e por último o Excel é gravado contendo um novo registo de entrada. Cada BDP é um PDF e

[9] FTE (*Full-Time Equivalent*) ou equivalente a tempo completo é um método de mensuração do grau de envolvimento de um colaborador nas atividades de uma organização ou unicamente em um determinado projeto, Fonte: http://www.businessdictionary.com/definition/full-time-equivalent-FTE.html

consequentemente é um registo de entrada no ficheiro Excel. A figura a seguir 3.1, ilustra um exemplo do processo de conversão dos BDP em formato PDF para Excel.

Figura 3.1 Processo de conversão de PDF para Excel

Na figura a seguir (Figura 3.2), ilustra o script responsável pela conversão. No campo *input files* é adicionado o caminho onde contém o PDF para conversão e no campo *output files* é adicionado o caminho que contem o Excel onde os dados são copiados. Após adicionar os caminhos, clica-se no botão *start conversion* onde é feito a conversão e é construído o *dataset*.

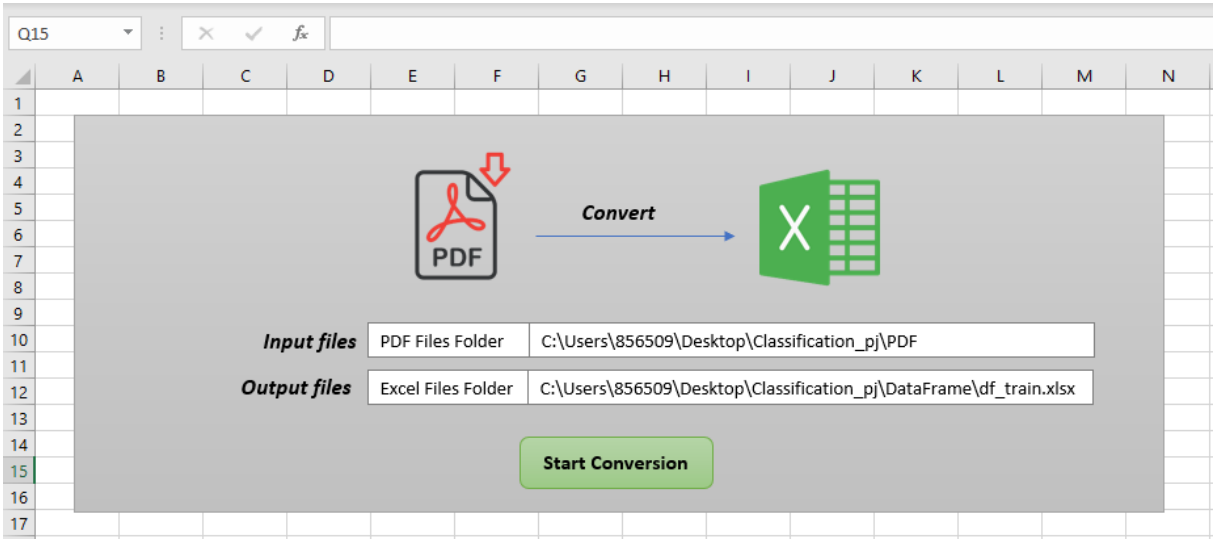


Figura 3.2: Script de conversão

A Tabela 3.4 ilustra uma análise descritiva após a construção do *dataset*.

Tabela 3.4: Análise descritiva do dataset

Nome	Descrição	Tipo	Domínio	Valores em falta
PDF	Descrição da operação	Character	Valores em caracteres	Não

Well	Nome do poço	Character	Valores em caracteres	Não
Date	Data de emissão do relatório	Numeric	Valores contínuos Ex. 31-Dec	Não
Start	Hora de início da atividade	Numeric	Horas no intervalo de 00:00 até 24:00	Não
End	Hora do fim da atividade	Numeric	Horas no intervalo de 00:00 até 24:00	Não
Duration	Duração da atividade	Numeric	Valores decimais superior a zero	Não
Type	Tipo de atividade	Factor	PT e NPT	Não
Rig	Designação da sonda de perfuração	Character	Valores em caracteres	Não
Job	Tipo de atividade	Factor	Valores em factor Ex.: *drilling* e *completion*	Não
NPT_Cause	Causa do Não Produtivo	Factor	8 possibilidades	Não
NPT_Description	Descrição do Não Produtivo	Factor	Valores em caracteres	Sim
Phase	Fase da operação de perfuração	Factor	9 fases	Não
Operation	Tipo de operação	Factor	Valores em caracteres	Não
Sub_Operation		Factor	Valores em caracteres	Sim
Description	Descrição da operação	Characters	Valores em caracteres	Não

Depois de construção do *dataset*, obteve-se 20418 registos de entrada e a próxima etapa é pré-processar o texto, importando os dados para uma lista de Python, através do *IDE Spyder*.

A variável PDF, representa os dados extraídos do BDP, e trata-se de uma variável dependente que contém textos que ajudam a classificar as operações. Cada linha desta variável representa uma atividade, e é com base nestas atividades que o engenheiro de perfuração consegue classificar as operações de perfuração.

Ao trabalhar com análise de textos, é muito comum depararmo-nos com problemas de *encoding*. Palavras com acentos, cedilha (ç) ficam desconfiguradas, comprometendo a apresentação dos resultados. Por esta razão, foi aplicada uma função no código Python para resolver questões desta natureza. A partir do atributo PDF, gerou-se, com a ajuda de um código Python, a seguinte nuvem de palavras:

Figura 3.3: Nuvem de palavras

Pode-se notar na Figura 3.1 que uma das palavas que mais se destacou foi “Navegando” e esta palavra pode ser classificada muitas vezes, dependendo do contexto, como uma atividade de tempo produtivo.

Atributo “*Type*”

O atributo usado para classificar o tipo de operação foi o “Type”. Este campo contém valores NPT (*Non-Productive Time*) e PT (*Productive Time*). Estes atributos identificam o tipo de atividade, como o tempo produtivo e o tempo não produtivo. Uma atividade pode ser considerada como PT quando o trabalho estimado para esta mesma atividade é efetuado de acordo com o tempo previsto. Enquanto que uma atividade NPT ocorre quando esta mesma atividade não é planeada ou concretizada conforme planeada.

A má classificação destas atividades pode gerar gastos enormes para as empresas de *joint ventures*, sendo que a eliminação/minimização do NPT é algo desejável na perfuração de qualquer poço, principalmente em operações de perfuração *offshore*, onde o custo diário da sonda representa uma parcela importante do custo total do poço. Por meio da análise dos relatórios de diversos poços, foi possível estabelecer análises comparativas e identificar casos onde os gastos de tempo estavam a ser excessivos. Uma vez identificado estes casos, esta informação pode ser utilizada no planeamento de novos poços. Assim sendo, durante o planeamento desses novos poços, as ações preventivas podem ser adotadas, visando eliminar o gasto excessivo de tempo em determinada etapa. Estas ações dependerão da natureza do problema observado e podem incluir desde ajustes no programa de perfuração, até modificações no projeto de novos poços.

No atributo *Type*, 74.9% de atividades do *dataset* usado é classificado como PT, enquanto que 25.1% representa a atividade como NPT (Figura 3.2).

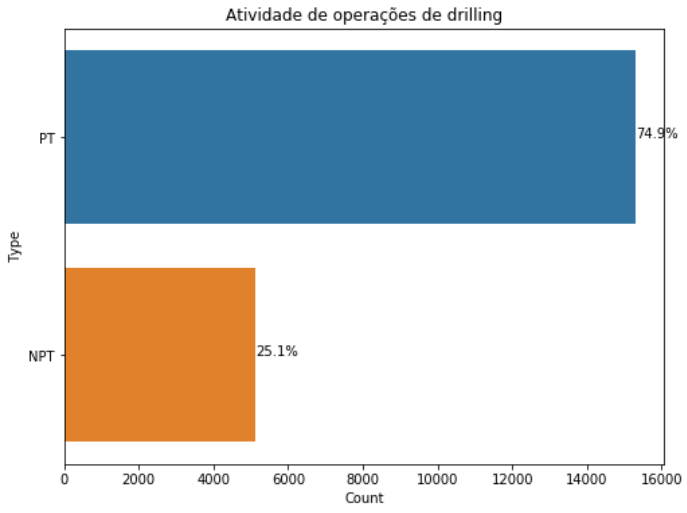


Figura 3.4: Gráfico de distribuição por tipos de atividade

Atributo “*NPT Cause*”

Após a classificação de atividade por tipo, o segundo ponto é identificar as causas de NPT. As atividades são categorizadas em 8 tipos de causas, nomeadamente:

- ***Hole Problems & Practices**:* (problemas com o poço), pode ser por exemplo, durante a perfuração o poço colapsa e a broca fica presa.
- ***Rig Equipments – Other:*** equipamentos da sonda como as bombas de injeção de lama no poço ou *top drive* (transmite rotação a coluna).
- ***3rd Party Services**:* Serviços na sonda que são prestados por terceiros - fornecedor de lama, ou fornecedor de sistema de ferramentas usadas durante a perfuração.
- ***WOW:** Wait On Weather* - condições não favoráveis para perfurar: muito vento ou muita ondulação no caso *offshore.*
- ***Rig Equipments – BOP & Riser**:* equipamento específico da sonda. Cada uma tem o seu BOP e este é o elemento chave de segurança no poço. Em caso de risco de fuga de hidrocarbonetos é possível fechar a comunicação entre o poço e a sonda.
- ***Logistics*:** Trata-se uma causa proveniente de questões logísticas.
- ***Other:*** Outros problemas que não se enquadram nos indicados acima.

De acordo com a Figura 3.3, a causa de NPT que representa a quantidade dos dados é *Hole Problems & Pratices* e *3rd Party Services* com uma percentagem de 28.8% e 28.6% respetivamente.

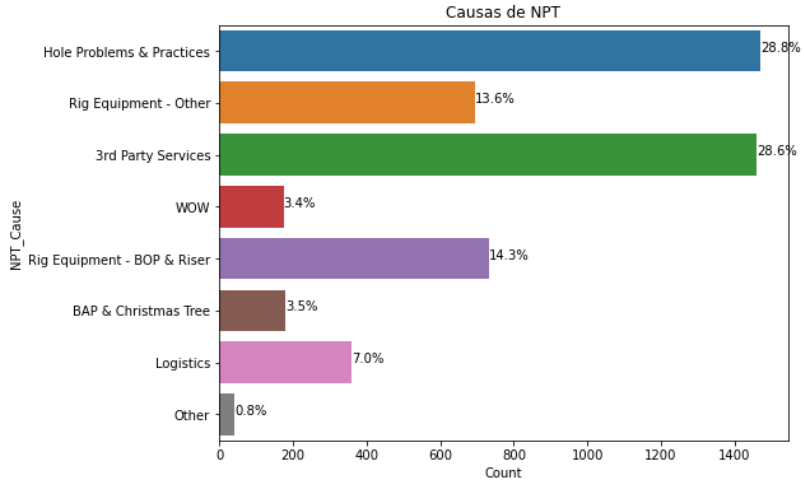


Figura 3.5: Gráfico de distribuição por Causa do NPT

Atributo "*Phase*"

As operações de perfuração estão divididas em 8 fases com uma quantidade de dados em 30.8% mais elevada na fase de completação, quando comparada com os outros dados conforme ilustra a figura 3.4.

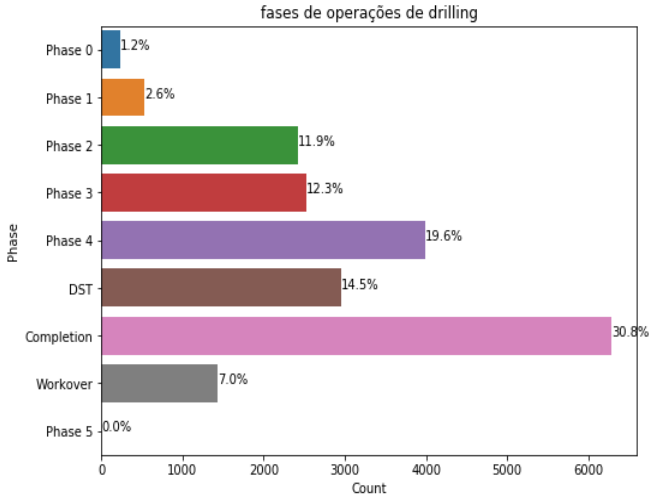


Figura 3.6: Gráfico de distribuição por fase de operações

Conforme visto na figura 3.4, os dados do *dataset* não estão balanceados, o que quer dizer que tem muito mais conteúdos em algumas classes em comparação com outras. A questão de balanceamento é muito importante para performance do classificador, uma vez que classes não balanceadas promovem uma certa vantagem em relação a classes que tem mais conteúdos. Para contornar este problema foram usadas técnicas *Undersampling* durante o treinamento do modelo.

3.4. Data Preparation

Segundo Quintela (2005), a preparação dos dados consiste num conjunto de atividades que tem como objetivo construir o conjunto de dados que foi usado para criação e validação do modelo nas próximas fases. Assim, aquando da validação dos dados, foi possível constatar que, após a extração dos dados dos BDP e a construção do *dataset*, nem todos os dados são relevantes e, por esta razão, foram selecionadas apenas as variáveis necessárias.

Deste modo, a coluna "Nome" na Tabela 3.5 mostra os atributos extraídos dos BDP. O "PDF" contém os atributos usados para criar os modelos, "Incluir (*DataSet*)" contém o tipo de dados do atributo processados e criados no ambiente Python. Diante disso, selecionaram-se as variáveis indicadas.

Tabela 3.5: Conjunto de dados a incluir no dataset

Nome	Tipo	Incluir (*DataSet*)	Justificação
PDF	Character	Sim	Variável principal para o treinamento do modelo
Well	Character	Não	Por questões de confidencialidade dos dados a variável não será apresentada.
Date	Numeric	Não	Não relevante para o nosso estudo
Start		Não	Não relevante para o nosso estudo
End		Não	Não relevante para o nosso estudo
Current_TD		Não	Não relevante para o nosso estudo
From	Numeric	Não	Não relevante para o nosso estudo
Until	Numeric	Não	Não relevante para o nosso estudo
Duration_Text	Numeric	Não	Não relevante para o nosso estudo
Duration	Numeric	Não	Não relevante para o nosso estudo
Type	Factor	Sim	Será usada para treinar o tipo de atividade

Rig	Character	Não	Por questões de confidencialidade dos dados a variável não será apresentada.
Job	Factor	Não	Não relevante para o nosso estudo
AFE		Não	Não relevante para o nosso estudo
NPT_Cause	Factor	Sim	Será usada para treinar a causa de NPT
NPT_Description	Character	Não	Contem valores omissos
Phase	Factor	Sim	Variável importante para o nosso treino
Operation	Factor	Sim	Variável complementar para o treinamento
Sub_Operation	Factor	Não	Contem valores omissos
Description	Charecter	Não	Não relevante para o nosso estudo

Por meio do comando de visualização dos dados obteve-se o resultado apresentado na Tabela 3.6, que ilustra as variáveis relevantes para o estudo:

RangeIndex: 20418 entries, 0 to 20417.

Data columns (total 5 columns):

Tabela 3.6: Sumario dos atributos usados

#	***Column***	***Non-Null Count***
0	*PDF*	*20418 non-null*
1	*Type*	*20418 non-null*
2	*NPT Cause*	*5113 non-null*
3	*Phase*	*20390 non-null*
4	*Operations*	*20417 non-null*

Após uma verificação detalhada do *dataset*, constatou-se que os dados existentes na coluna *PDF*, continham números, caracteres especiais e espaços indesejados. De forma a não comprometer a performance no treino, verificou-se a necessidade de se fazer uma "limpeza" dos dados. A Tabela 3.7 apresenta um exemplo com cinco registos do *dataset* antes da referida "limpeza" de dados.

Tabela 3.7: Dataset antes do tratamento dos dados

Index	PDF	Type	NPT Cause	Phase	Operations
3322	10:30 11:30 1,00 0,00 0,00 45 Verificando vazamento de óleo diesel pela STT, parando bombeio, fechando S-1 na ANM com pressão na cabeça de 410 psi, drenando pressão para planta de well test, ciclando S-1 da STT e testando com 500 psi e 2000 psi, ok.	NPT	BAP & Christmas Tree	Workover	Upper Completion
3323	11:30 13:30 2,00 0,00 0,00 Equalizando pressão com 410 psi e abrindo S-1 na ANM e continuando a bombear diesel @ 3 bpm com pressão subindo de 600 psi até 1670 psi e PDG = 7555 psi total de diesel bombeado = 400 bbl. Pressão após bombeio: na cabeça = 1477 psi, no PDG = 7460 psi.	PT	nan	Workover	Upper Completion
3324	13:30 14:00 0,50 0,00 0,00 Fechando S-1 na ANM, e drenando pressão do DPR para well test.	PT	nan	Workover	Upper Completion
3325	14:00 16:00 2,00 0,00 0,00 Retirando BOP / XO de slick line e instalando XO / tampão da Weatherford no topo da STT.	PT	nan	Workover	Upper Completion
3326	16:00 00:00 8,00 0,00 0,00 46 Aguardando barco de acidificação Blue Marlim.	NPT	Logistics	Workover	Upper Completion

Para a limpeza do *dataset* foi construída uma função e algumas técnicas para remover caracteres especiais, acentos, espaços vazios e números do texto. A Tabela 3.8 apresenta um exemplo do *dataset* após a limpeza dos dados.

Tabela 3.8: Dataframe apos o tratamento dos dados

Index	PDF	Type	NPT_ Cause	Phase	Operations
3322	Verificando vazamento de oleo diesel pela STT parando bombeio fechando S na ANM com pressao na cabeca de psi drenando pressao para planta de well test ciclando S da STT e testando com psi e psi ok	NPT	BAP & Christmas Tree	Workover	Upper Completion
3323	Equalizando pressao com psi e abrindo S na ANM e continuando a bombear diesel bpm com pressao subindo de psi ate psi e PDG psi total de diesel bombeado bbl Pressao apos bombeio na cabeca psi no PDG psi	PT	nan	Workover	Upper Completion
3324	Fechando S na ANM e drenando pressao do DPR para well test	PT	nan	Workover	Upper Completion
3325	Retirando BOP XO de slick line e instalando XO tampao da Weatherford no topo da STT	PT	nan	Workover	Upper Completion
3326	Aguardando barco de acidificacao Blue Marlim	NPT	Logistics	Workover	Upper Completion

O exemplo ilustrado na tabela 3.8 (*index* 3322), apresenta a utilização da classificação como um elemento de auxílio na identificação da natureza do tempo não-produtivo (NPT). Se considerarmos que tempo não-produtivo corresponde ao tempo durante o qual não houve avanço na profundidade do poço, os patamares serão indicativos de NPT. Analisando a classificação correspondente aos patamares para este intervalo, vê-se que o tempo não produtivo "destacado" corresponde sobretudo a aguardando, manobras, testando linhas. Assim sendo, neste intervalo, a classificação observada pode ser considerado NPT planeado. Mas como "verificando o vazamento", trata-se de NPT não planeado, a sua causa deve ser investigada, a fim de evitar a repetição desse evento neste poço e em poços futuros.

Para melhor entender a distribuição dos dados foi preciso recorrer uma técnica de t-SNE, por forma a perceber a similaridade entre elas. Esta técnica agrupa as palavras com significado semelhantes mais próximas entre elas.

O t-SNE (*t-Distributed Stochastic Neighbor Embedding*), de acordo com Van Der Maaten & Hinton (2008), é um algoritmo para redução de dimensionalidade e é adequado para visualizar dados de alta dimensão. O nome significa Incorporação estocástica de vizinhos distribuída em *t*. A ideia é incorporar pontos de alta dimensão em baixas dimensões, de maneira a respeitar as semelhanças entre os pontos. Os pontos próximos no espaço de alta dimensão correspondem a pontos de baixa dimensão incorporados nas proximidades, e os pontos distantes no espaço de alta dimensão correspondem a pontos de baixa dimensão incorporados distantes. Apêndice A, ilustra a forma como as palavras estão agrupadas.

As máquinas, ao contrário dos humanos, não conseguem entender o texto bruto. Máquinas só conseguem ler números. Particularmente, técnicas estatísticas, como ML, podem lidar apenas com números. Portanto, optou-se por converter o texto em números (tabela exemplificativa mais abaixo – tabela 3.9). Cada linha dos arquivos de texto contém uma coluna de texto "PDF", que contém a descrição do relatório extraído dos BDPs.

Existem diferentes abordagens para converter texto na forma numérica correspondente, nomeadamente a *Bag of Words Model* (Modelo de Bolsa de Palavras) e o *Word Embedding Model* (Modelo de Incorporação de Palavras). Neste trabalho, foi usado *Bag of Words Model* para converter o

texto em números. O *script* a seguir usa o modelo de incorporação de palavras para converter textos em recursos numéricos correspondentes.

```
# TF-IDF Train Data
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(max_features=1500, min_df=5, max_df=0.7, stop_words=stopwords.words('Portuguese'))
X = vectorizer.fit_transform(documents).toarray()
```

Figura 3.7: Script de conversão de textos para números

O *script* ilustrado na Figura 3.6, usa a classe *CountVectorizer* da biblioteca *sklearn.feature_extraction.text*. Existem alguns parâmetros importantes que foram passados para construir a classe.

O primeiro parâmetro é o *max_features*. Todos os documentos podem conter dezenas de milhares de palavras únicas. Mas as palavras que têm uma frequência de ocorrência muito baixa não são um bom parâmetro para classificar textos. Portanto, foi definido o parâmetro *max_features* com o valor de 1500, o que significa que foram usadas 1500 palavras mais ocorrentes como recursos para treinar o classificador.

O próximo parâmetro é *min_df* e foi definido como 5. Isso corresponde ao número mínimo de documentos que devem conter esse recurso. Portanto, foram incluídas apenas as palavras que ocorrem em pelo menos cinco ficheiros. Da mesma forma, para o *max_df*, o valor é definido como 0,7; em que a fração corresponde a uma percentagem. Aqui, o valor de 0,7 significa que foram incluídas apenas as palavras que ocorrem o máximo de 70% de todos os ficheiros. As palavras que ocorrem em quase todos os ficheiros geralmente não são adequados para classificação porque não fornecem nenhuma informação exclusiva sobre o ficheiro.

E por último, foram removidos as *stopwords*[10] do texto pois, no caso de classificação de textos, as *stopwords* podem não conter nenhuma informação útil. Para removê-las, foi alterado o objeto *stopwords* da biblioteca *nltk.corpus*, para o parâmetro *stop_words*.

A função *fit_transform* da classe *CountVectorizer* é responsável por converter documentos de texto em recursos numéricos correspondente.

Tabela 3.9: Exemplo de uma lista de texto convertidos em uma sequência de números

Key	Type	Size	Value
direcional	*int*	1	602
derecionando	*Int*	1	1959
direita	*Int*	1	1258
disco	*Int*	1	1238
discutindo	*Int*	1	1481
disparo	*Int*	1	735
distancia	*Int*	1	878
diversas	*int*	1	1555

[10] *stopwords* são palavras que podem ser consideradas irrelevantes para o conjunto de resultados a ser exibido em uma busca realizada em uma *search engine*. Exemplos: as, e, os, de, para, com, sem, foi.
(Fonte: http://www.nltk.org/howto/portuguese_en.html)

A seguir, foi feita uma conversão dos atributos *factor* pelo objeto *factorize* da biblioteca *StringIO* começando pelo "*Type*", em que o resultado passou a ser colocado como '0' para classificação NPT e '1' para classificação PT conforme ilustra a Tabela 3.10.

Tabela 3.10: Conversão do atributo Type para factor

Index	Type	Factor
0	*NPT*	0
77	*PT*	1

O atributo "*NPT_Cause*" também foi convertido para número, conforme ilustra a Tabela 3.11, mas antes foi necessário filtrar todos os dados de NPT e limpar os dados omissos para não criar entropia.

Tabela 3.11: Conversão do atributo NPT Cause para factor

Index	NPT_Cause	Factor
77	*Hole Problem e Pratice*	0
106	*Rig Equipment - Other*	1
495	*3rd Party Services*	2
599	*WOW*	3
744	*Rig Equipment – BOP & Riser*	4
1181	*BAP & Cristmas Tree*	5
1531	*Logistics*	6
1940	*Other*	7

Por fim, o atributo *Phase* também foi categorizado em atributos numéricos e, como forma de evitar entropia, foram elimidados os 28 dados omissos contidos neste atributo. A Tabela 3.12, ilustra um exemplo deste atributo após a conversão para números.

Tabela 3.12: Conversão do atributo Phase para factor

Index	Phase	Factor
0	*Phase 0*	0
17	*Phase 1*	1
43	*Phase 2*	2
171	*Phase 3*	3
427	*Phase 4*	4
738	*DST*	5
1114	*Workover*	6
2126	*Completion*	7
3104	*Phase 5*	8

De acordo com Zhang, Jin, & Zhou (2010), a abordagem *Bag of Words Model* funciona bem para converter texto em números. No entanto, tem uma desvantagem: o modelo atribui uma pontuação a uma palavra com base na sua ocorrência num documento específico. Não leva em conta o fato de que a palavra também pode ter uma alta frequência de ocorrência noutros documentos. É preciso ainda ter em conta que algumas palavras no *corpus* de treino serão muito presentes, como é o caso de preposições e artigos. Estas palavras tendem repetir-se em todo o *dataset* e não costumam carregar informação muito significativa para o objetivo deste estudo.

Assim, utilizou-se a medida de TFIDF[11] para limitar a importância destas palavras que se repetem muito ao longo dos documentos, de maneira a que não causem mais influência do que o necessário. O TFIDF resolve esse problema multiplicando o termo frequência de uma palavra pela frequência inversa

[11] http://www.tfidf.com/

do documento. O TF significa "*Term Frequency*", enquanto o IDF significa "*Inverse Document Frequency*". O termo frequência é calculado da seguinte forma:

$$Term\ Frequeny = \frac{Nº\ de\ ocorrencia\ de\ uma\ palavra}{Total\ de\ palavras\ no\ documento} \quad (3.1)$$

E a Frequência Inversa de documento é calculada:

$$IDF\ (Word) = log\ \left(\frac{Nº\ de\ documentos}{Nº\ de\ documentos\ que\ contêm\ a\ palavra}\right) \quad (3.2)$$

A fórmula geral de TFIDF baseia-se em:

$$W_{x,y} = tf_{x,y} * log\ \left(\frac{N}{df_x}\right) \quad (3.3)$$

Onde:

$tf_{x,y}$ = frequência de x em y
df_x = número de documentos que contém x
N = Número total de documentos

O valor TFIDF para uma palavra num documento específico é maior se a frequência de ocorrência dessa palavra for também maior nesse documento. No entanto, menor em todos os outros documentos. Para converter os valores obtidos usando o modelo de pacote de palavras em valores TFIDF, foi executado o seguinte *script* (Figura 3.7):

```
from sklearn.feature_extraction.text import TfidfTransformer
tfidfconverter = TfidfTransformer()
X = tfidfconverter.fit_transform(X).toarray()
```

Figura 3.8: Script de TFIDF

Como qualquer problema de aprendizagem de máquina supervisionado, foi preciso dividir os dados em conjuntos de treino e teste. Para tal, foi feita uma operação de *split*, de forma aleatória, para dividir os dados que serão utilizados para treinamento e os dados que serão utilizados para a validação do algoritmo classificador.

É importante observar que a separação dos dados em treino e validação é uma etapa essencial e que, caso seja realizada de maneira errada, poderá resultar em problemas no modelo. Se, por exemplo, fossem definidas as primeiras 75% das linhas para treino, deixando as últimas 25% para validação, e a classificação das fases estivessem organizadas em ordem, algumas fases especificas estariam presentes apenas nos dados de treino e outras apenas nos dados de teste. Esta situação levaria a um modelo deficiente e que não aprendeu com todos os tipos de dados, logo, também não será validado de forma correta.

A escolha de aleatoriedade dos dados deveu-se ao facto de não haver padrão no momento da divisão dos dados e cada observação terá a mesma probabilidade de ser selecionada. Assim, o algoritmo foi treinado com um grande volume de dados de treino, sendo validado posteriormente os resultados destes

algoritmos através de validação. Só assim se pode ter a confiança de que o algoritmo consegue prever os dados reais.

3.4. Cenários de experimentação

Foram separadas as experimentações em cenários para que seja facilitada a sua compreensão. O estudo vai obedecer a dois cenários diferentes, com experimentação de cada atributo classificador para cada modelo de treinamento dos dados.

Para se tornar o trabalho mais robusto e, dada a sua natureza, os modelos de redes neuronais escolhidos foram os que obedecem a critérios de multicamadas para os dois tipos de redes neuronais apresentados na revisão da literatura: redes MLP e LSTM. A escolha destes modelos deveu-se ao facto de outros modelos de redes neuronais como ADALINE e *Perceptron* apresentarem resultados bastantes baixos, mostrando-se inadequados para classificar textos da natureza do presente estudo.

Tabela 3.13: Cenários de experimento

Cenários	Modelo	Atributos
1	MLP	Tipo de atividade, Causa do NPT, e Fase de perfuração do poço
2	LSTM	Tipo de atividade, Causa do NPT, e Fase de perfuração do poço

3.4.1. Cenário 1

Para treinar classificadores supervisionados, primeiro transformámos a coluna “PDF” num vetor de números. Foram exploradas representações vetoriais, como vetores ponderados por TF-IDF. Após ter este vetor em representações de texto, foram treinados classificadores supervisionados, treinando a coluna *“PDF”* e prevendo o *“Type”, “NPT Cause” e “Phase”* no quais se enquadram.

Cada uma das 20418 registo de entrada extraído do “PDF” é representada por 15461 recursos, representando a pontuação TF-IDF para diferentes *unigramas* (coleção de palavras isoladas) e *bigramas* (coleção de duas palavras). Foi usado a biblioteca *sklearn.feature_selection.chi2* para encontrar os termos que estão mais correlacionados com cada um dos *“Type”, “NPT Cause” e “Phase”*. Após definidas todos os recursos e rótulos, foram treinados os classificadores obedecendo os modelos de multicamadas proposto para este estudo.

Tabela 3.14: Exemplo de Unigrama e Bigrama

	Unigrama	Bigrama
Palavras	coluna	coluna revestimento

Apos vários testes, o modelo MLP proposto para o estudo é de três camadas ocultas com oito neurónios para cada camada. De acordo com Braga, Ludermir, & De Carvalho (2000), o número de camadas depende da complexidade do problema em estudo. Redes maiores, composta de muitas camadas, conseguem “aprender” mais padrões até um certo linear, sendo que cada camada lida com um certo conjunto de abstrações. Muitas camadas, por outro lado, têm um custo computacional envolvido.

É pertinente lembrar que nem sempre ter uma rede com muitas camadas é necessariamente melhor. Além disso, o *overfitting* [12] pode eventualmente surgir.

Apesar de uma camada oculta ser suficiente para a grande maioria dos problemas, o modelo proposto apresentou melhores resultados com três camadas ocultas. Este resultado foi obtido após vários testes porque, quanto mais camadas ocultas, mais profunda é a rede neuronal. Para o treinamento do modelo MLP foram definidos os seguintes parâmetros de entrada, conforme Tabela 3.15:

Tabela 3.15: Parâmetro de entrada de MLP

Parametro	Descrição	Valores
hidden_layer_sizes	Representa o número de neurónios na enésima camada oculta	(8, 8, 8)
alpha	Constante que multiplica o termo L1.	1e-5
max_iter	O número de iterações	200
solver	Tamanho do mini *bach*	lbfgs (é um optimizador na família de métodos quasi-Newton.)
Activation	Função de ativação para a camada oculta	Relu (a função de unidade linear retificada, retorna f (x) = max (0, x))
Verbose	mostrará os níveis de log WARNING e INFO	True
random_state	Determina a geração de números aleatórios para inicialização de pesos e desvios	40

3.4.2. Cenário 2

Neste cenário, o procedimento usado para converter a coluna "PDF" num vetor de números foi feito através do comando *from keras.preprocessing.text import Tokenizer*, onde foi importada uma classe que permite vetorizar um *corpus* de texto em números, transformando cada texto numa sequência de números inteiros (cada número inteiro é o índice de um *token* num dicionário) ou num vetor em que o coeficiente de cada *token* pode ser binário, com base na contagem de palavras , com base em TF-IDF. Após ter este vetor em representações do texto, foram treinados 75 % dos dados em "PDF" e previu-se o *"Type", "NPT Cause" e "Phase"* no qual se enquadram.

Apos vários experimentos, o modelo LSTM proposto para o estudo é de 64 neurónios e *epochs* igual a 60. Em redes neurais, uma época (*epochs*) é uma única passagem pelo conjunto completo de treinamento. O treinamento pode levar milhares de épocas para o algoritmo de retro propagação convergir para uma combinação de pesos com um nível aceitável de precisão. Para o treinamento do modelo RNN foram definidos os seguintes parâmetros de entrada (Tabela 3.16):

Tabela 3.16: Parâmetro de entrada de LSTM

Parametro	Descrição	Valores
input_length	Tamanho do vetor de sentença de entrada	300
units	Quantidade de células na camada LSTM	64
dropout_rate	Taxa de *dropout* da camada de entrada	0.5
bach_size	número de amostras a serem utilizadas em cada atualização do gradiente	128
optimizer	Tipo de optimizador	*adam*
epochs	Número de épocas	10

[12] *overfitting* ocorre quando o modelo se adaptou muito bem aos dados com os quais está sendo treinado; porém, não generaliza bem para novos dados. Ou seja, o modelo "decorou" o conjunto de dados de treino, mas não aprendeu de fato o que diferencia aqueles dados para quando precisar enfrentar novos testes (Cawley & Talbot, 2010).

word_embedding_dim	dimensionalidade do *word embedding* pré-treinado	50
max_fatures	Reflete a quantidade máxima de palavras mantidas no vocabulário	5000
embed_dim	dimensão de saída da camada *Embedding*	128
loss	calcula a quantidade que um modelo deve procurar minimizar durante o treinamento.	*binary_crossentropy*

3.5. Integração do modelo

O processo através do qual as várias frases dos textos extraídos do BDP são classificadas será abstraído do utilizador, sendo que este apenas terá acesso aos resultados devolvidos pelo modelo. Isto é, para os utilizadores não é visível a interação que tem de ocorrer no IDE Spyder para que seja possível classificar o texto do BDP (bastando, para tal, ter instalado o motor do *Python*). Após o BDP ser classificado, este é retornado ao utilizador de forma estruturada, numa tabela onde, para cada uma das frases (*input*), é apresentada a respetiva classificação em classes. Nas Figura 3.8 e 3.9 são apresentadas as páginas onde o utilizador verá o resultado da classificação das fases de operações efetuado pelo modelo de classificação.

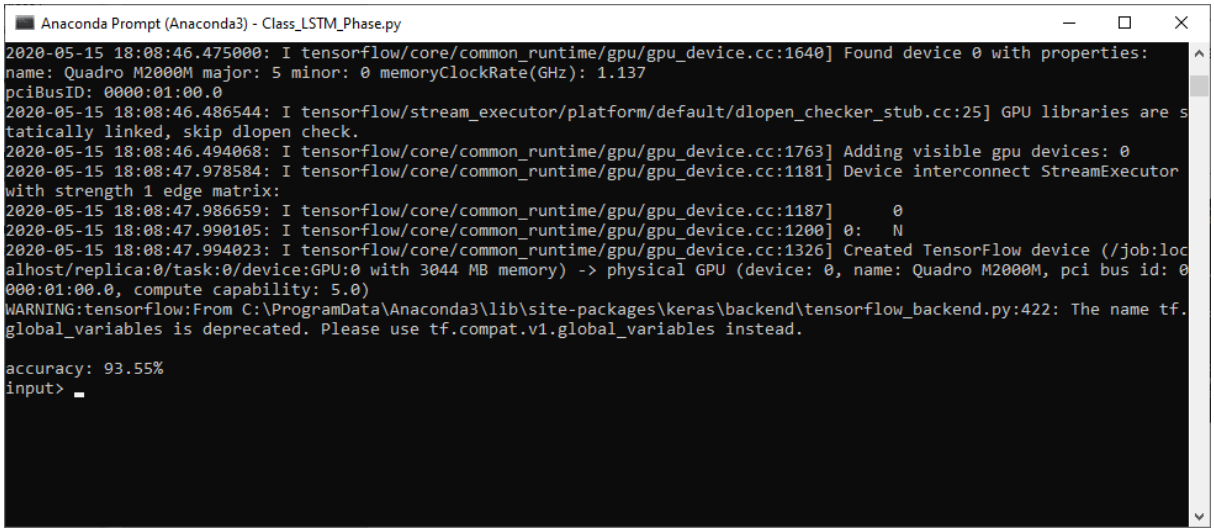


Figura 3.9: Exemplo de ambiente de classificação

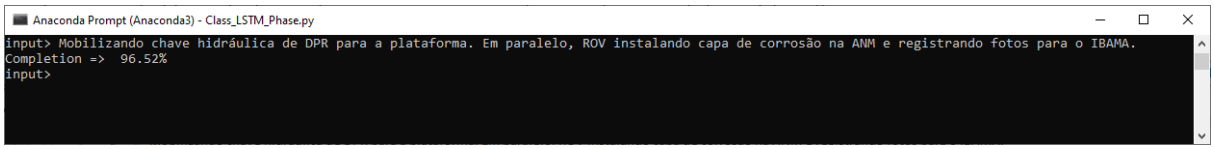


Figura 3.10: Exemplo de um texto classificado

Capítulo 4

4. Resultados

Neste capítulo, numa primeira instância, são detalhadas as métricas de avaliação utilizadas. Subsequentemente, de acordo com essas métricas, são apresentados e discutidos os resultados obtidos com os modelos de classificação de texto e respetivas comparações. São apresentados os resultados para os dois cenários propostos. Além disso, é analisado o comportamento dos dois cenários, e por sua vez, os resultados obtidos.

4.1. Métricas de avaliação e validação dos modelos

Para avaliar os modelos de classificação, optou-se por adotar as seguintes métricas (Ian Witten, Eibe Frank, Mark Hall, 2016): precisão, sensibilidade, F1-Score e curva ROC. Onde TP, FP, e FN corresponde ao número de verdadeiro positivo, falso positivo e falso negativo respetivamente. A precisão é intuitivamente a capacidade do classificador de não rotular como positiva uma amostra negativa.

$$Precisão = \frac{TP}{TP + FP} \tag{4.1}$$

A sensibilidade é intuitivamente a capacidade do classificador de encontrar todas as amostras positivas.

$$Sensibilidade = \frac{TP}{TP + FN} \tag{4.2}$$

Para complementar as métricas de precisão e sensibilidade, foram utilizados o F1-score, que pode ser interpretada como uma média ponderada da precisão e recuperação, onde uma pontuação F1 atinge o seu melhor valor em 1 e pior pontuação em 0. A contribuição relativa da precisão e recuperação para a F1-score é igual. O formulário para F1-score é:

$$F_1 = 2 \times \frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade} \tag{4.3}$$

Geralmente, a sensibilidade e a especificidade são características difíceis de conciliar, é complicado aumentar a sensibilidade e a precisão de um teste ao mesmo tempo. As curvas ROC (*receiver operator characteristic curve*) são uma forma de representar a relação, normalmente antagónica, entre a precisão e a sensibilidade de um teste diagnóstico quantitativo. Outra forma de avaliar o modelo é através do *accuracy* (AUC). É comum interpretar a qualidade dos valores da AUC como: 0,5 - igual a um classificador aleatório; 0,6 - razoável; 0,7 - bom; 0,8 - muito bom; 0,9 - excelente; e 1 – perfeito (Landis & Koch, 1977).

4.1.1. Desempenho do cenário 1

a) Type

O *accuracy* para o atributo *Type* foi 0.98 (98%), com uma precisão de acerto de 99% para o NPT e 97% para o NPT.

Tabela 4.1: Accuracy do modelo MLP no tipo de atividade

	Precisão	Sensibilidade	F1-score	support
NPT	0.98	0.91	0.94	4097
PT	0.97	0.99	0.98	12237
Accuracy			0.98	16334
Macro avg	0.98	0.95	0.96	16334
Weighted avg	0.97	0.97	0.97	16334

O apêndice B, apresenta uma curva de *accuracy* e de perda sobre os dados do treinamento do atributo Tipo, nas métricas de sensibilidade e F1-score.

Como forma de melhor perceber a performance do modelo MLP no cenário 1, foi apresentado uma matriz de confusão (Figura 4.1). Uma matriz de confusão é usada para visualizar a performance de um classificador. As linhas de matriz indicam as classes que se espera obter por um modelo. As colunas indicam as classes que foram obtidas efetivamente. Cada coluna contém o número de predições feitas pelo classificador, relativas ao contexto daquela célula especifica.

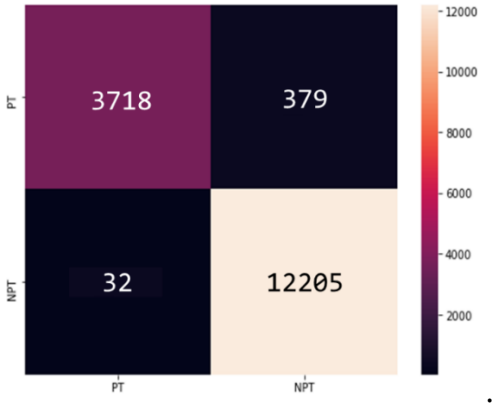


Figura 4.1: Matriz de confusão do modelo MLP no tipo de atividade

b) NPT Cause

O *accuracy* para o atributo NPT_Cause foi de 0.97 (97%)

Tabela 4.2: Accuracy do modelo MLP na Causa de NPT

	Precisão	Sensibilidade	F1-score	support
3rd Party Services	0.91	1.00	0.95	1192
BAP & Cristmas Tree	1.00	1.00	1.00	143
Hole Problem e Pratice	1.00	1.00	1.00	1163
Logistics	1.00	0.95	0.98	292
Other	1.00	0.39	0.57	33
Rig Equipment – BOP & Riser	1.00	0.93	0.96	580
Rig Equipment - Other	1.00	0.93	0.96	557
WOW	1.00	0.94	0.97	130
Accuracy			0.97	4090
Macro avg	0.99	0.89	0.92	4090
Weighted avg	0.97	0.97	0.97	4090

O apêndice C, apresenta uma curva de *accuracy* e de perda sobre os dados do treinamento do atributo NPT_Cause, nas métricas de sensibilidade e F1-score.

É de se observar na Figura 4.2 que o modelo repete inúmeras vezes a classificação *Hole Problems & Pratice* de forma incorreta em várias classificações. Isto deve-se à natureza da própria categoria (problemas com o poço), que facilmente pode ser confundida com outras categorias.

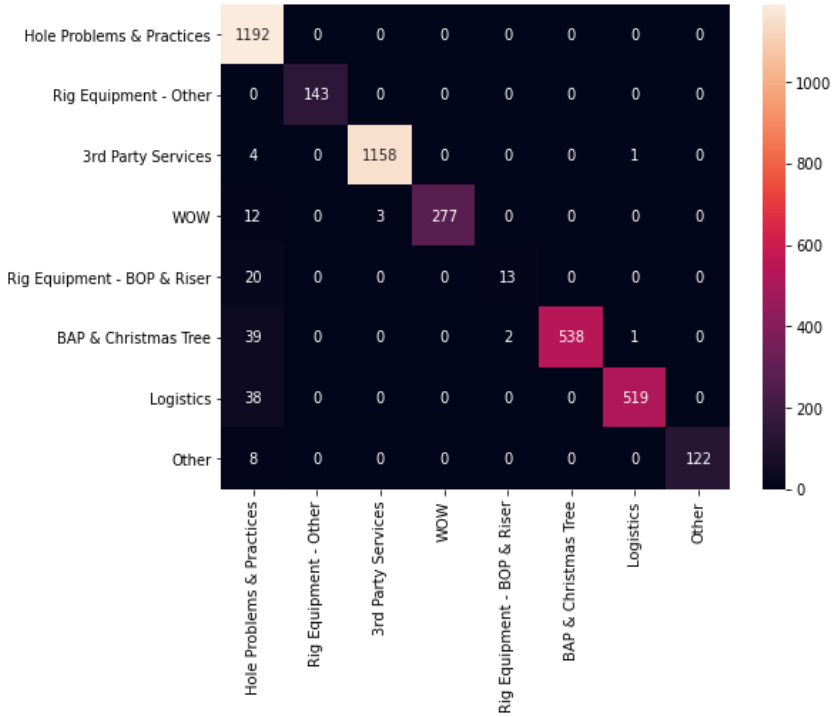


Figura 4.2: Matriz de confusão do modelo MLP na Causa de NPT

c) Phase

O *accuracy* para o atributo *Phase* foi de 0.94 (94%)

Tabela 4.3: Accuracy do modelo MLP na fase de operação

	Precisão	Sensibilidade	F1-score	support
Completion	0.87	0.99	0.93	5010
DST	1.00	0.98	0.99	2366
Phase 0	0.99	0.96	0.92	191
Phase 1	0.93	0.87	0.90	429
Phase 2	0.97	0.93	0.95	1924
Phase 3	0.97	0.89	0.93	2015
Phase 4	0.96	0.85	1.90	3232
Phase 5	1.00	1.00	1.00	1
Workover	0.99	0.99	0.99	1144
Accuracy			0.94	16312
Macro avg	0.97	0.93	0.95	16312
Weighted avg	0.94	0.94	0.94	16312

O apêndice D, apresenta uma curva de *accuracy* e de perda sobre os dados do treinamento do atributo *Phase*. Nas métricas de sensibilidade e F1-score.

A Figura 4.3 ilustra a matriz de confusão do Atributo *Phase*:

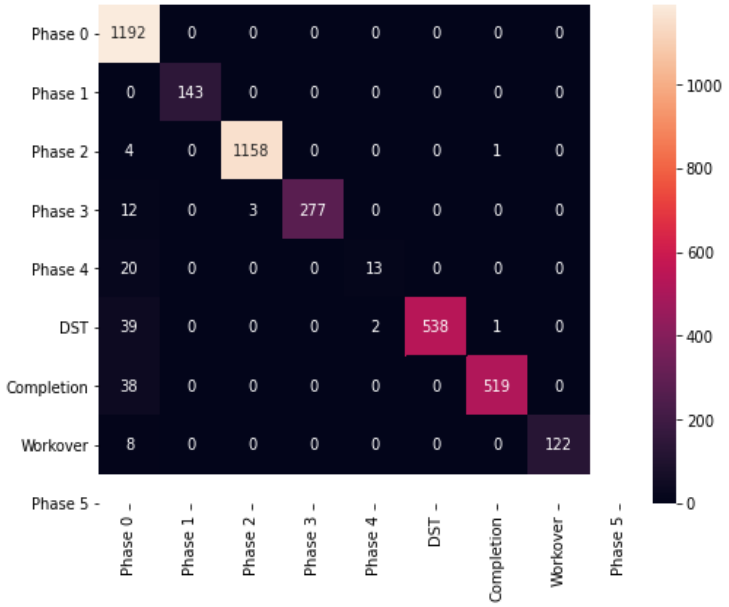


Figura 4.3: Matriz de confusão do modelo MLP na fase de operação

4.1.2. Desempenho do cenário 2

a) Type

O *accuracy* para o atributo *Type* neste cenário foi de 0.91 (91%)

Tabela 4.4: Accuracy do modelo LSTM no tipo de operação

	Precisão	Sensibilidade	F1-score	support
NPT	0.93	0.71	0.80	3718
PT	0.91	0.98	0.94	12205
Accuracy			0.91	16334
Macro avg	0.92	0.85	0.87	16334
Weighted avg	0.92	0.91	0.90	16334

O apêndice E, apresenta uma curva de *accuracy* sobre os dados do treinamento do atributo *Type*. Nas métricas de sensibilidade e *F1-score*, o aumento dos dados do treino teve um impacto positivo.

A matriz de confusão apresentado na Figura 4.4, ilustra um resumo da performance do modelo no atributo *Type* para o modelo LSTM durante o treinamento.

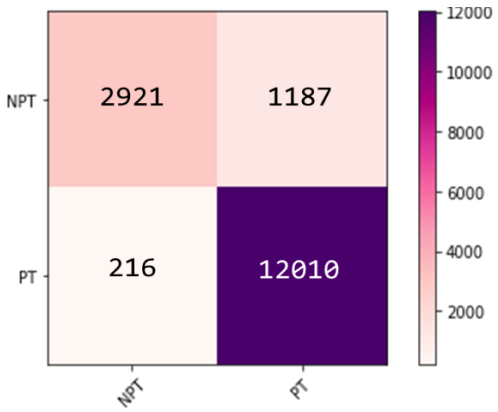


Figura 4.4: Matriz de confusão do modelo LSTM no tipo de atividade

b) NPT_Cause

O *accuracy* para o atributo NPT_Cause é de 0.96 (96%)

Tabela 4.5: Accuracy do modelo LSTM na cause de NPT

	Precisão	Sensibilidade	F1-score	support
3rd Party Services	0.91	1.00	0.95	1192
BAP & Cristmas Tree	1.00	0.91	0.95	143
Hole Problem e Pratice	1.00	0.99	1.00	1163
Logistics	1.00	0.95	0.98	292
Other	1.00	0.12	0.22	33
Rig Equipment – BOP & Riser	1.00	0.92	0.96	580
Rig Equipment - Other	0.95	0.93	0.94	557
WOW	1.00	0.94	0.97	130
Accuracy			0.96	4090
Macro avg	0.98	0.85	0.87	4090
Weighted avg	0.97	0.96	0.96	4090

O apêndice F, apresenta uma curva de *accuracy* sobre o treinamento do atributo NPT_Cause.

A matriz de confusão apresentado na figura 4.5, ilustra um resumo da performance do modelo no atributo NPT_Cause durante o treinamento.

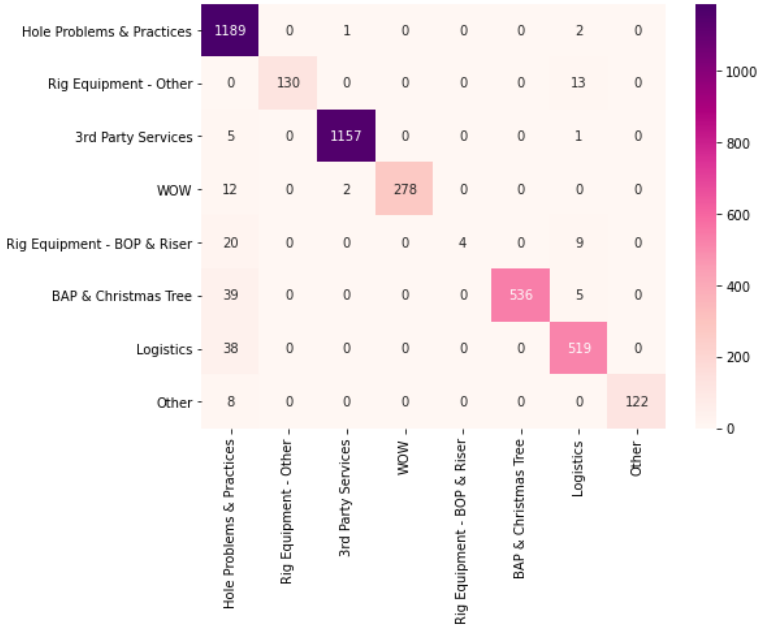


Figura 4.5: Matriz de confusão do modelo LSTM na Causa de NPT

c) Phase

Neste atributo, obteve-se um *accuracy* igual a 0.92 (92%).

Tabela 4.6: Accuracy do modelo LSTM na fase de operação

	Precisão	Sensibilidade	F1-score	support
Completion	0.87	0.98	0.92	5010
DST	0.97	0.98	0.97	2366
Phase 0	0.95	0.83	0.89	191
Phase 1	0.87	0.80	0.84	429
Phase 2	0.94	0.92	0.93	1924
Phase 3	0.96	0.85	0.90	2015
Phase 4	0.93	0.84	0.88	3232
Phase 5	0.00	0.00	0.00	1
Workover	0.99	0.98	0.98	1144
Accuracy			0.92	16312
Macro avg	0.83	0.80	0.81	16312
Weighted avg	0.92	0.92	0.92	16312

O apêndice G, apresenta uma curva de *accuracy* sobre o treinamento do atributo *Type.*

A matriz de confusão apresentado na figura 4.6, ilustra um resumo da performance do modelo durante o treinamento.

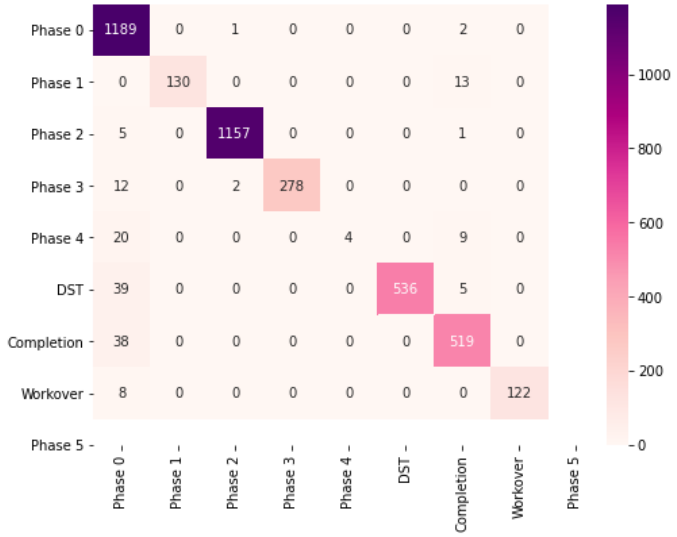


Figura 4.6: Matriz de confusão do modelo LSTM na Fase de operação

4.2. Análise dos modelos

Face aos resultados obtidos, constatou-se que o modelo que apresentou a melhor performance foi o modelo MLP do cenário 1 (Tabela 4.7). Por esta razão, escolheu-se o cenário 1 como proposta para o presente estudo.

Tabela 4.7: Resumo de análise do modelo

Cenário	Atributo	Accuracy	Precisão	Sensibilidade	F1-score
Cenário 1 (MLP)	*Type*	0.97	0.98	0.95	0.96
	NPT Cause	0.97	0.99	0.98	0.92
	Phase	0.94	0.96	0.93	0.95
Cenário 2 (LSTM)	*Type*	0.91	0.86	0.85	0.87
	NPT Cause	0.96	0.95	0.96	0.96
	Phase	0.92	096	0.83	0.86

4.3. Interpretação *do modelo*

Neste trabalho foi proposto o uso de algoritmos para fornecer previsões individuais como uma solução para "confiar numa previsão" e selecionar várias previsões (e explicações) como uma solução para o problema “confiar no modelo”. Para tal, existem algumas técnicas para interpretar modelos do tipo *black-box*. Um exemplo disso é o package lime. LIME (*Local Interpretable Model-agnostic Explanations*) é um algoritmo que pode explicar as previsões de qualquer problema de classificação ou regressão de maneira fiel, através de um modelo interpretável (Ribeiro, Singh, & Guestrin, 2016).

A Figura 4.7 mostra o resultado obtido para classificação do NPT e PT, no modelo MLP, incluindo as respetivas previsões atribuídas pelo modelo. Neste cenário, o LIME previu com uma certeza de 100% a classificação de NPT. As palavras que dominaram a esta escolha foram: “vazamento”, “apresentou”, “localizado”, “indicativo” e “queda”. Estas palavras, permitem a equipa de perfuração segregar os

problemas e entender onde os mesmos ocorrem e se aparecem nas mesmas fases. Caso, se perceba que estas palavras estão associadas sempre a mesma fase podem permitir a equipa rever o programa de perfuração e promover melhorias para futuros poços, de forma a olhar para o programa do poço e ver onde se pode melhorar.

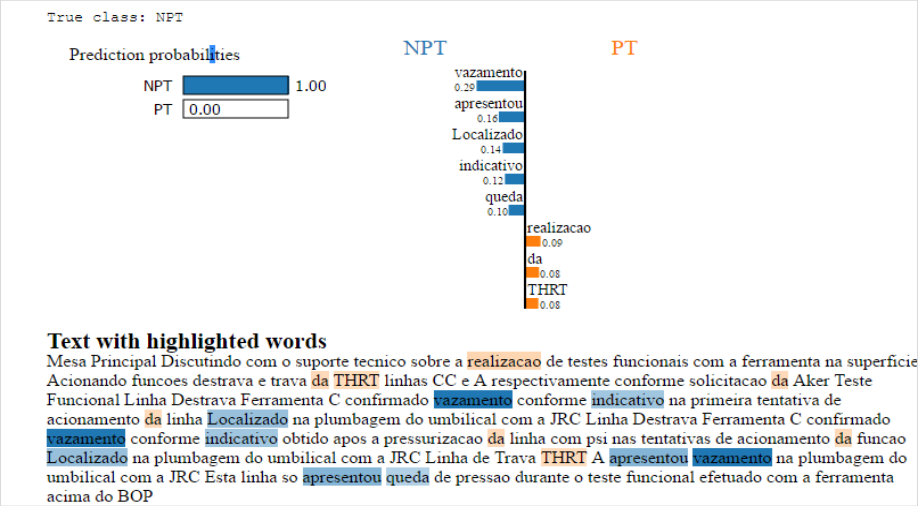


Figura 4.7: Validação do modelo para tipo de atividade

Para sustentar a capacidade do modelo, foi possível fazer um *predict* do texto como forma de verificar se o modelo era capaz de prever o resultado (Figura 4.9):

```
In [13]: print(mlp.predict(count_vect.transform(["Testando ferramenta CaTS Obtido uma amostra
de cada sensor de pressaotemperatura Leitura positiva Obs Para obtencao de cada dado e
necessario aguardar min para envio do sinal comprimento da informacao Em seguida a ferramenta
de teste do CaTS repete o mesmo sinal em min Desta forma e lido o sinal de fundo e o sinal
transmitido pela ferramenta de teste"])))
['NPT']
```

Figura 4.8: Predict do tipo de atividade, fonte: Elaboração do autor

No seguinte exemplo, a azul, estão destacados os termos que mais contribuem para a classificação "*Rig Equipment – Other*", sendo o principal destaque para os termos "Reparo" e "Manipulador". A laranja, são destacados os termos que levam o modelo a considerar que a Causa de NPT poderá pertencer a uma classe que não é "*Rig Equipment – Other*".

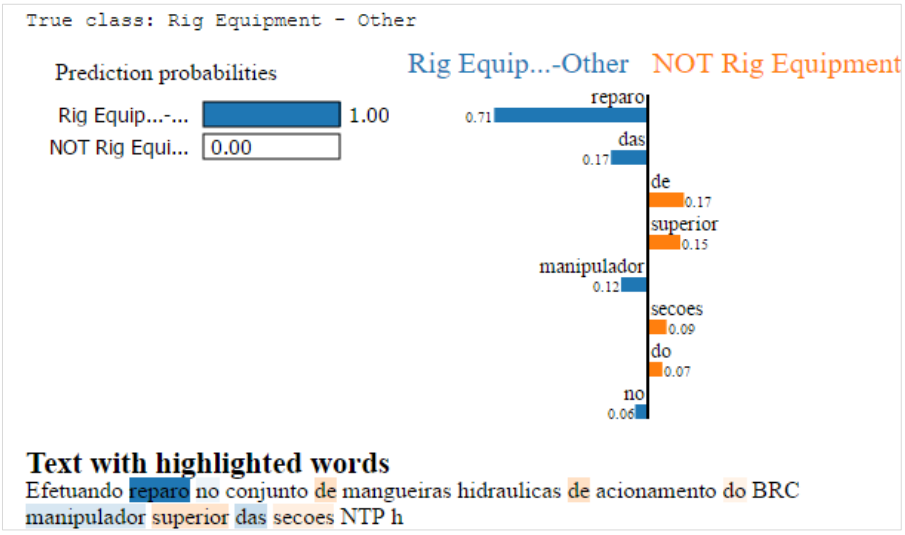


Figura 4.9: Validação do modelo para Causa de NPT

A figura 4.10, apresenta um *predict* de texto como forma de verificar se o modelo era capaz de prever o resultado:

```
In [30]: print(mlp_npt.predict(count_vect_npt.transform(["Testando ferramenta CaTS Obtido uma
amostra de cada sensor de pressaotemperatura Leitura positiva Obs Para obtencao de cada dado e
necessario aguardar min para envio do sinal comprimento da informacao Em seguida a ferramenta
de teste do CaTS repete o mesmo sinal em min Desta forma e lido o sinal de fundo e o sinal
transmitido pela ferramenta de teste"])))
['BAP & Christmas Tree']
```

Figura 4.10: Predict do modelo MLP na Causa de NPT

A azul (Figura 4.11), estão destacados os termos que mais contribuem para a classificação "*Completion*", sendo o principal destaque para os termos "*overpull*". A laranja, são destacados os termos que levam o modelo a considerar que a fase de operação poderá pertencer a uma classe que não é "*Completion*".

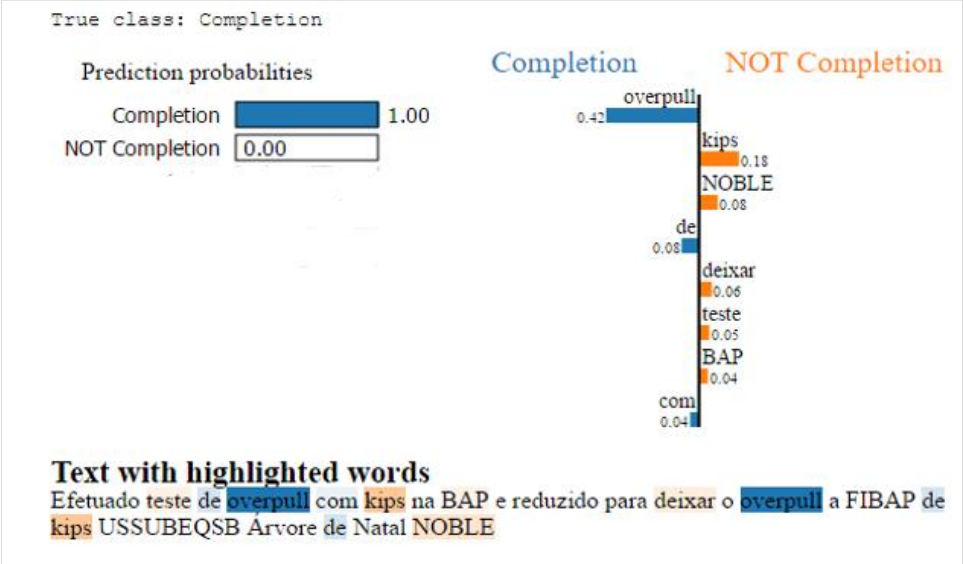


Figura 4.11: Validação do modelo para fase de operação

Conforme apresentado nas visualizações alcançadas através do método de validação LIME, percebeu-se que o modelo apresenta um bom desempenho de classificação. A figura 40 apresenta um *predict* de texto como forma de verificar se o modelo era capaz de prever o resultado (Figura 4.12):

```
In [40]: print(mlp_phase.predict(count_vect_phase.transform(["Testando ferramenta CaTS Obtido
uma amostra de cada sensor de pressaotemperatura Leitura positiva Obs Para obtencao de cada
dado e necessario aguardar min para envio do sinal comprimento da informacao Em seguida a
ferramenta de teste do CaTS repete o mesmo sinal em min Desta forma e lido o sinal de fundo e o
sinal transmitido pela ferramenta de teste"])))
['Completion']
```

Figura 4.12: Predict do modelo MLP na fase de operação

O apêndice H, ilustra a curva ROC para o modelo MLP, nos atributos Type, NPT_Cause e Phase, o que mostra que a curva apresenta resultados bastante próximos de 1 para os 3 modelos.

Capítulo 5

5. Considerações Finais

Machine Learning tem sido bastante utilizado em várias áreas de estudo, porém raramente são exploradas as suas capacidades na área de E&P. Neste trabalho, foram aplicadas técnicas de redes neuronais num *dataset* com 20418 registos com os dados exportados dos BDP, através de uma macro desenvolvida em linguagem VB.NET. Este estudo tem contribuições académicas bastante interessantes, pois fornece resultados concretos sobre o comportamento dos modelos de RNA provenientes de treinamento, com base em dados reais.

O objetivo geral deste estudo era construir um modelo de IA usando RNA para classificar as operações de perfuração de um poço de petróleo. Esse objetivo foi desmembrado noutros três e, para que fosse alcançado, construiu-se um modelo baseado em modelo MLP para classificação das 3 etapas: tipo de atividade, causa de atividade não produtiva e a fase de operação de perfuração.

Submetidas as análises de validação do modelo por meio de gráficos de *accuracy* e de perda e uma validação por meio de *package lime* para garantir interpretabilidade do modelo, o modelo foi considerado aceite para classificação de operações de perfuração. No que diz respeito ao desempenho no cenário 1, os resultados obtidos nas métricas de precisão, sensibilidade e *F1-score* são satisfatórios.

Além disso, foram comparados os modelos MLP e LSTM e, apesar de não ser escolhido o modelo LSTM, os resultados obtidos também foram satisfatórios, com as áreas abaixo da curva ROC bastante próxima do valor 1. Assim, o modelo desenvolvido no cenário 1, baseada na arquitetura MLP, revela um desempenho de *accuracy* superior ao modelo LSTM.

É possível afirmar que os resultados obtidos confirmam que o estudo feito na revisão da literatura para os modelos de RNA, apresentam melhores resultados em resposta a outros modelos na área de E&P. Este estudo transforma-se numa vantagem competitiva para uma empresa que opera na área de E&P, pois consegue facilmente classificar as operações de perfuração.

O sistema de classificação de etapas forneceu uma classificação precisa e detalhada das atividades realizadas durante a perfuração. A descrição fornecida pelo modelo permite identificar eventos que estão a consumir tempo excessivo de sonda e contribui para o processo de minimização do tempo não-produtivo na perfuração do poço.

Este modelo não só poderá trazer benefícios na redução de horas de trabalho à equipa de D&C da Galp E&P, como também poderá economizar recursos financeiros. Segundo números fornecidos pela equipa de D&C, em média, uma classificação manual é feita em cerca de 45 minutos, divididos em 15 minutos para leitura do relatório e 30 minutos para classificar as operações.

Com o sistema de IA proposto, prevê-se uma poupança de 30 minutos por relatório. E o tempo gasto passará de 45 para 15 minutos apenas para validar a classificação e rever as possíveis falhas. O que quer dizer que, para cada 5 relatórios por dia, o sistema consegue poupar em média 2h30min de trabalho.

É de realçar que este tempo poderá diminuir com a evolução de aprendizagem contínua do modelo e, consoante a confiança e quantidade de dados que o modelo pode aprender, maior será a sua performance.

5.1. Limitações do estudo e propostas de investigações futuras

O presente trabalho deparou-se com limitações, sendo que a maior foi o facto de não ser possível identificar, através do modelo, os problemas que podem ocorrer durante a perfuração. O procedimento de identificação das fases de operação proposto neste trabalho poderia ser estendido à identificação de problemas de perfuração. No entanto, a limitação referente à confidencialidade dos dados, não permitiu incluir estes problemas no *dataset.*

Com base nos dados do mesmo campo de exploração e, onde a geologia é semelhante, é possível, através de histórico, prever o tempo de duração de novos poços. Assim sendo, para trabalhos futuros, será apresentado um modelo de regressão capaz de prever o tempo de perfuração de um poço, com objetivo de diminuir o tempo de sonda e otimizar a perfuração de novos poços.

Através do procedimento de identificação de problemas, será possível identificar o comportamento dos parâmetros para um caso real de *Pack-off*. O procedimento pode ser utilizado para evitar a ocorrência de problemas durante a perfuração do poço. Considerando novas pesquisas, propõe-se o aprofundamento dos estudos sobre a identificação de problemas de perfuração.

Referências bibliográficas

Ab Wahab, M. N., Nefti-Meziani, S., & Atyabi, A. (2015). A Comprehensive Review of Swarm Optimization Algorithms. *PLOS ONE, 10*(5), e0122827. https://doi.org/10.1371/journal.pone.0122827

Agwu, O. E., Akpabio, J. U., Alabi, S. B., & Dosunmu, A. (2018, August 1). Artificial intelligence techniques and their applications in drilling fluid engineering: A review. *Journal of Petroleum Science and Engineering*, Vol. 167, pp. 300–315. https://doi.org/10.1016/j.petrol.2018.04.019

Ahmadi, M. A., Shadizadeh, S. R., Shah, K., & Bahadori, A. (2018). An accurate model to predict drilling fluid density at wellbore conditions. *Egyptian Journal of Petroleum*, *27*(1), 1–10. https://doi.org/10.1016/j.ejpe.2016.12.002

Anifowose, F. A., Labadin, J., & Abdulraheem, • Abdulazeez. (2016). Hybrid intelligent systems in petroleum reservoir characterization and modeling: the journey so far and the challenges ahead. *Journal of Petroleum Exploration and Production Technology*, *7*. https://doi.org/10.1007/s13202-016-0257-3

Azevedo, A., & Santos, M. F. (2008). *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*.

B. Rable. (2017). *The future is here: 3 ways AI roots itself in O&G in the surge Magazine*. Retrieved from http://thesurge.com/stories/future-artificial-intelligence-roots-oil-gas-industry

Barreto, J. M. (2002). Introdução às Redes Neurais Artificiais. *V Escola Regional de Informática. Sociedade Brasileira de Computaçao, Regional Sul, Santa Maria, Florianópolis, Maringá*, 5–10. https://doi.org/10.4028/www.scientific.net/AMR.716.240

Bello, O., Holzmann, J., Yaqoob, T., & Teodoriu, C. (2015). APPLICATION OF ARTIFICIAL INTELLIGENCE METHODS IN DRILLING SYSTEM DESIGN AND OPERATIONS: A REVIEW OF THE STATE OF THE ART. *JAISCR*, *5*(2), 121. https://doi.org/10.1515/jaiscr-2015-0024

Bishop, C. M. (1996). *Neural Networks: A Pattern Recognition Perspective*. Retrieved from http://www.ncrg.aston.ac.uk/

Bradley, W. B., Jarman, D., Auflick, R. A., Plott, R. S., Wood, R. D., Schofield, T. R., & Cocking, D. (1991). Task force reduces stuck-pipe costs. *Oil and Gas Journal; (United States)*, *89:21*.

Braga, A. de P., Ludermir, T. B., & De Carvalho, A. P. de L. F. (2000). *Redes Neurais Artificiais - Teoria e Prática*. Retrieved from http://www.livrariasaraiva.com.br/produto/1975236/redes-neurais-artificiais-teoria-e-pratica-2-ed-2011

Brice, W. R. (2009). *Edwin L. Drake (1819-1880) His Life and Legacy*.

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. In *Journal of Machine Learning Research* (Vol. 11).

Chapman, Pete and Clinton, Julian and Kerber, Randy and Khabaza, Thomas and Reinartz, Thomas and Shearer, Colin and Wirth, R. (1999). *CRISP-DM 1.0 Step-by-step data mining guide*. DaimlerChrysler.

Cheeseman, P., & Stutz, J. (1990). *Bayesian Classification (AutoClass): Theory and Results*.

Elkatatny, S. (2017). Real-Time Prediction of Rheological Parameters of KCl Water-Based Drilling Fluid Using Artificial Neural Networks. *Arabian Journal for Science and Engineering*, *42*(4), 1655–1665. https://doi.org/10.1007/s13369-016-2409-7

Elman, J. (1990). Finding Structure in Time. In *COGNITIVE SCIENCE* (Vol. 14).

Empressa de Pesquisa Energética. (2018). *Desafios do Pré-Sal Estudos de Longo Prazo*.

Fayyad, U. M., Piatetsky-Shapiro, G. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*.

Galp. *Exploration Drilling - Functional Design*. , (2017).

George, A. L., & Bennett, A. (2005). *Case Studies and Theory Development in the Social Sciences* (Belfer Cen).

Gil, A. C. (2002). *Como elaborar projetos de pesquisa*.

Goldstein, E. B., & Coco, G. (2014). A machine learning approach for the prediction of settling velocity. *Water Resources Research*, *50*(4), 3595–3601. https://doi.org/10.1002/2013WR015116

Gove, R., & Faytong, J. (2012). Machine Learning and Event-Based Software Testing: Classifiers for Identifying Infeasible GUI Event Sequences. In *Advances in Computers* (Vol. 86, pp. 109–135). https://doi.org/10.1016/B978-0-12-396535-6.00004-1

Guilherme, I., Queiroz, J., Urgal, P., & Chavarette, F. (2010). Sistema Inteligente Web Para Diagnóstico De Anormalidade Da Perfuração De Poços De Petroleo. *Sbmac.Org.Br*, (November 2014), 423–427. Retrieved from https://sbmac.org.br/dincon/trabalhos/PDF/control/68314.pdf

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.

Haykin, S. (2001). *Redes Neurais - Principios e prática* (2ª). Retrieved from https://books.google.com/books?id=lBp0X5qfyjUC&pgis=1

Heriot Watt. (2013). Drilling Engineering. In *Heriot Watt University*. Material de apoio ao curso de MSc Petroleum Engineering.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holland, J. H. . (1992). Genetic Algorithms understand Genetic Algorithms. *Scientific American*, *267*(1), 66–73. https://doi.org/10.2307/24939139

Ian Witten, Eibe Frank, Mark Hall, C. P. (2016). Data mining: practical machine learning tools and techniques. 4th edn. Retrieved July 12, 2020, from Morgan Kaufmann website: https://www.elsevier.com/books/data-mining/witten/978-0-12-804291-5

Jeirani, Z., & Mohebbi, A. (2006). Artificial Neural Networks Approach for Estimating Filtration Properties of Drilling Fluids. *Journal of the Japan Petroleum Institute*, *49*(2), 65–70. https://doi.org/10.1627/jpi.49.65

Kamari, A., Gharagheizi, F., Shokrollahi, A., Arabloo, M., & Mohammadi, A. H. (2017). Estimating the drilling fluid density in the mud technology: Application in high temperature and high pressure petroleum wells. In *Heavy Oil: Characteristics, Production and Emerging Technologies* (pp. 285–295). Nova Science Publishers, Inc.

Kohn, T., & Manaris, B. (2020). Tell me what'swrong: A python ide with error messages. *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, 1054–1060. https://doi.org/10.1145/3328778.3366920

Konar, A. (1999). Artificial Intelligence and Soft Computing: Behavioral and Cognitive ... - Amit Konar - Google Livros.

Kovács, Z. L. (2002). *Redes neurais artificiais : fundamentos e aplicações*. Ed. Livraria da Física.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Legg, S., & Hutter, M. (2007). *A Collection of Definitions of Intelligence*. Retrieved from http://arxiv.org/abs/0706.3639

Lúcia da Silva, E., & Muszkat Menezes, E. (2005). *Metodologia da Pesquisa e Elaboração de Dissertação*. Retrieved from http://www.ufsc.xn--brctccentrotecnolgicohttp-6pc//www.ctc.ufsc.brhttp://www.ced.ufsc.brhttp://www.cin.ufsc.br

Mathworks. (2016). *Introducing Machine Learning*.

Mcculloch, W. S., & Pitts, W. (1943). A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY* n. In *Bulletin of Mothemnticnl Biology* (Vol. 52).

Moazzeni, A., Nabaei, M., & Jegarluei, S. G. (2012). Decision Making for Reduction of Nonproductive Time through an Integrated Lost Circulation Prediction. *Petroleum Science and Technology*, *30*(20), 2097–2107. https://doi.org/10.1080/10916466.2010.495961

Monard, M. C., & Baranauskas, J. A. (2003). *Capítulo 4 Conceitos sobre Aprendizado de Máquina*.

Moro, S., Laureano, R. M. S., & Cortez, P. (2011). *USING DATA MINING FOR BANK DIRECT MARKETING: AN APPLICATION OF THE CRISP-DM METHODOLOGY*.

Murillo, A., Neuman, J., & Samuel, R. (2009). Pipe sticking prediction and avoidance using adaptive fuzzy logic and neural network modeling. *SPE Production and Operations Symposium, Proceedings*, 244–258. https://doi.org/10.2118/120128-ms

Mushtaq, M.-S., & Mellouk, A. (2017). *Quality of experience paradigm in multimedia services : application to OTT video streaming and VoIP services*. ISTEP Press Ltd.

Neves, J. L. (2002). *PESQUISA QUALITATIVA-CARACTERÍSTICAS, USOS E POSSIBILIDADES*.

Nilsson, N. J. (2009). *THE QUEST FOR ARTIFICIAL INTELLIGENCE A HISTORY OF IDEAS AND ACHIEVEMENTS*. Retrieved from http://www.cambridge.org/us/0521122937http://www.cambridge.org/us/0521122937http://www.cambridge.org/us/0521122937

Osman, E. A., & Aggour, M. A. (2003). Determination of Drilling Mud Density Change with Pressure and Temperature Made Simple and Accurate by ANN. *Middle East Oil Show*. https://doi.org/10.2118/81422-MS

Ozbayoglu, E. M., & Ozbayoglu, M. A. (2009). Estimating flow patterns and frictional pressure losses of two-phase fluids in horizontal wellbores using artificial neural networks. *Petroleum Science and Technology*, *27*(2), 135–149. https://doi.org/10.1080/10916460701700203

Quintela, H. (2005). *Sistemas de Conhecimentos Baseados em Data Mining - Aplicação à Análise da Estabilidade de Estruturas Metálicas*.

Rauber, T. W. (2014). Redes neurais artificiais. Retrieved December 3, 2019, from https://www.researchgate.net/publication/228686464_Redes_neurais_artificiais

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *ACM SIGSOFT Software Engineering Notes*. https://doi.org/10.1145/2939672.2939778

Rich, E., & Knight, K. (1991). *Artificial Intelligence*. Retrieved from https://books.google.pt/books?id=6P6jPwAACAAJ

Rooki, R., Ardejani, F. D., Moradzadeh, A., Mirzaei, H., Kelessidis, V., Maglione, R., & Norouzi, M. (2012). Optimal determination of rheological parameters for herschel-bulkley drilling fluids using genetic algorithms (GAs). *Korea Australia Rheology Journal*, *24*(3), 163–170. https://doi.org/10.1007/s13367-012-0020-3

Rosenblatt, F. (1958). THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN 1. In *Psychological Review* (Vol. 65).

Rowsell, P. J., & Waller, M. D. (1994). *Drilling process and apparatus*.

Schlumberger. (2007). *Oilfield Review Winter 2007/2008*.

Schlumberger. (2019). Oilfield Glossary. Retrieved December 10, 2019, from https://www.glossary.oilfield.slb.com/en/Terms.aspx?filter=r

Shahdi, A., & Arabloo, M. (2014). Application of SVM Algorithm for Frictional Pressure Loss Calculation of Three Phase Flow in Inclined Annuli INVESTIGATION OF WAX DEPOSITION PROCESSES UNDER TWO-PHASE GAS-OIL SLUG FLOW View project Application of SVM Algorithm for Frictional Pressure Loss Calculation of Three Phase Flow in Inclined Annuli. *J Pet Environ Biotechnol*, *5*, 3. https://doi.org/10.13140/RG.2.1.2297.8162

Shepherd, M. (2009). Oil Field Production Geology. In *American Association of Petroleum Geologists*. Retrieved from https://books.google.pt/books?id=a2ErAgAAQBAJ&pg=PA17&lpg=PA17&dq=volume+evaluation+in+place+HIIP&source=bl&ots=1nei2SKBDe&sig=ACfU3U2Q1WimoAwfsmsA3A28gZ4BgpX0zA&hl=pt-BR&sa=X&ved=2ahUKEwjMsqrX_8vmAhWoxYUKHaAAArQQ6AEwBXoECAkQAQ#v=onepage&q=volume evaluati

Shobha, G., & Rangaswamy, S. (2018). Machine Learning. In *Handbook of Statistics* (Vol. 38, pp. 197–228). https://doi.org/10.1016/bs.host.2018.07.004

Siruvuri, C., Nagarakanti, S., & Samuel, R. (2006). Stuck Pipe Prediction and Avoidance: A Convolutional Neural Network Approach. *IADC/SPE Drilling Conference*. https://doi.org/10.2118/98378-MS

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Toreifi, H., Habib, Abbas, R., & Manshad, K. (n.d.). *New method for prediction and solving the problem of drilling fluid loss using modular neural network and particle swarm optimization*

algorithm. https://doi.org/10.1007/s13202-014-0102-5

Van Der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. In *Journal of Machine Learning Research* (Vol. 9).

Wang, G., Pu, X.-L., & Tao, H.-Z. (2012). A Support Vector Machine Approach for the Prediction of Drilling Fluid Density at High Temperature and High Pressure. *Petroleum Science and Technology*, *30*(5), 435–442. https://doi.org/10.1080/10916466.2011.578095

Widrow, B., & Hoff, M. E. (1960). Adaptive Switching Circuits. *1960 IRE WESCON Convention Record, Part 4*, 96–104. Retrieved from http://www-isl.stanford.edu/~widrow/papers/c1960adaptiveswitching.pdf

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Xia, J., Xie, F., Zhang, Y., & Caulfield, C. (2013). Artificial intelligence and data mining: Algorithms and applications. *Abstract and Applied Analysis*, *2013*(December). https://doi.org/10.1155/2013/524720

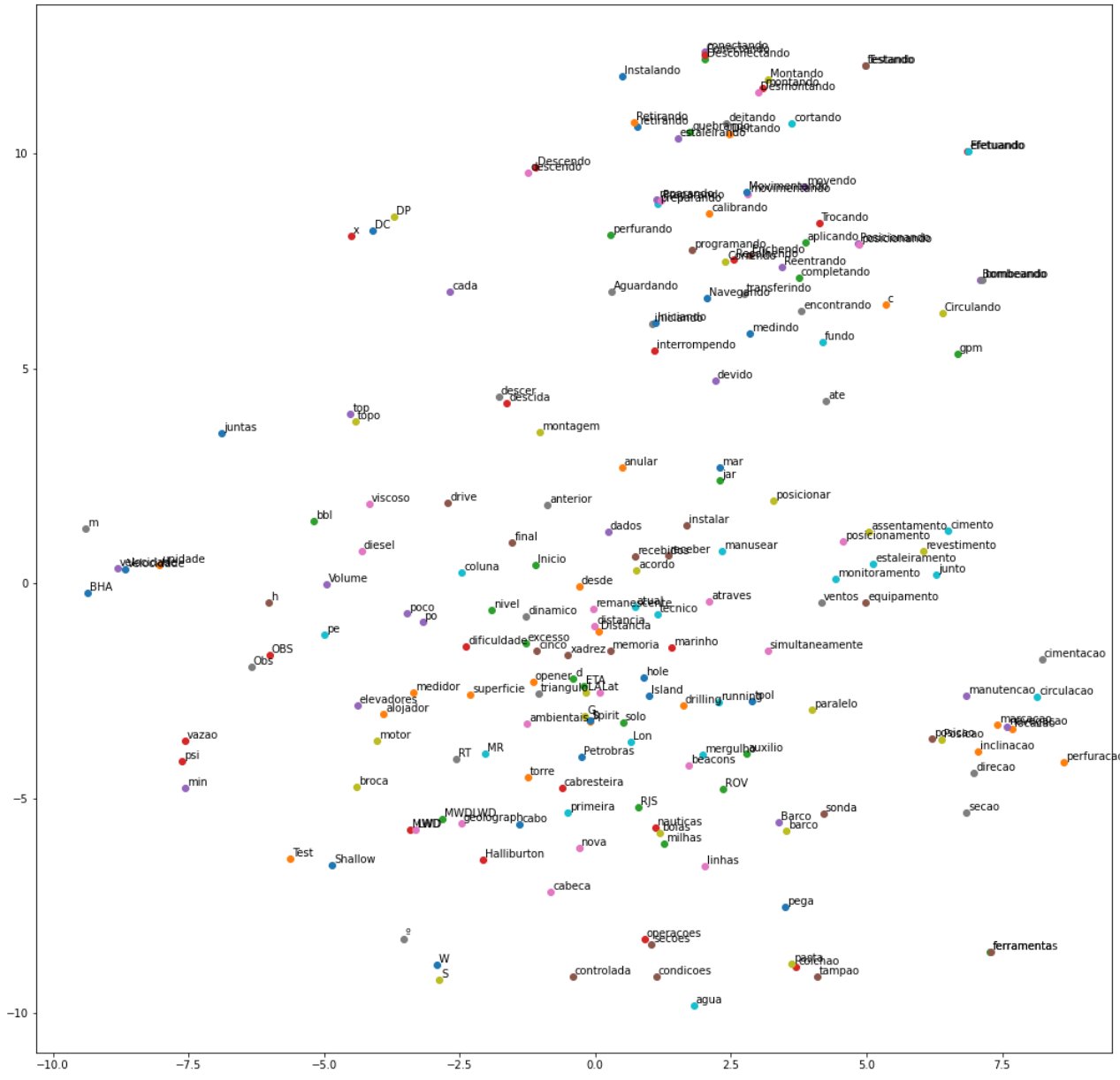
Zaremba, W., Sutskever, I., Vinyals, O., & Brain, G. (2015). *RECURRENT NEURAL NETWORK REGULARIZATION*. Retrieved from https://github.com/wojzaremba/lstm.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). *Understanding Bag-of-Words Model: A Statistical Framework*.

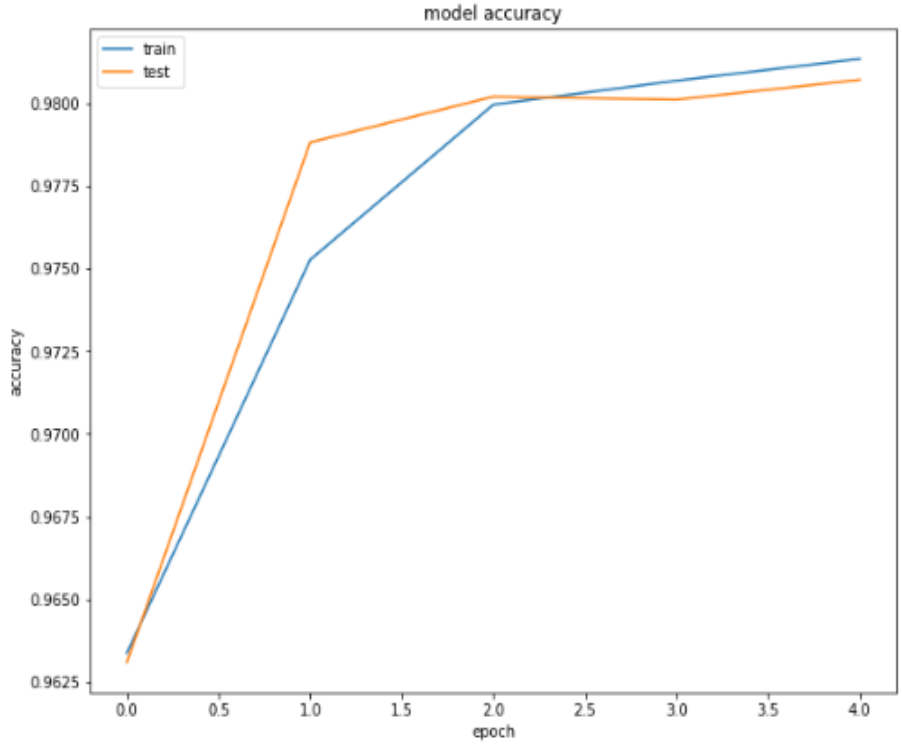
Zhu, D., Liu, G. X., & Zhang, Q. Z. (2013). Research of prewarning pipe-sticking based on neural network. *Applied Mechanics and Materials*, *327*, 1734–1737. https://doi.org/10.4028/www.scientific.net/AMM.325-326.1734

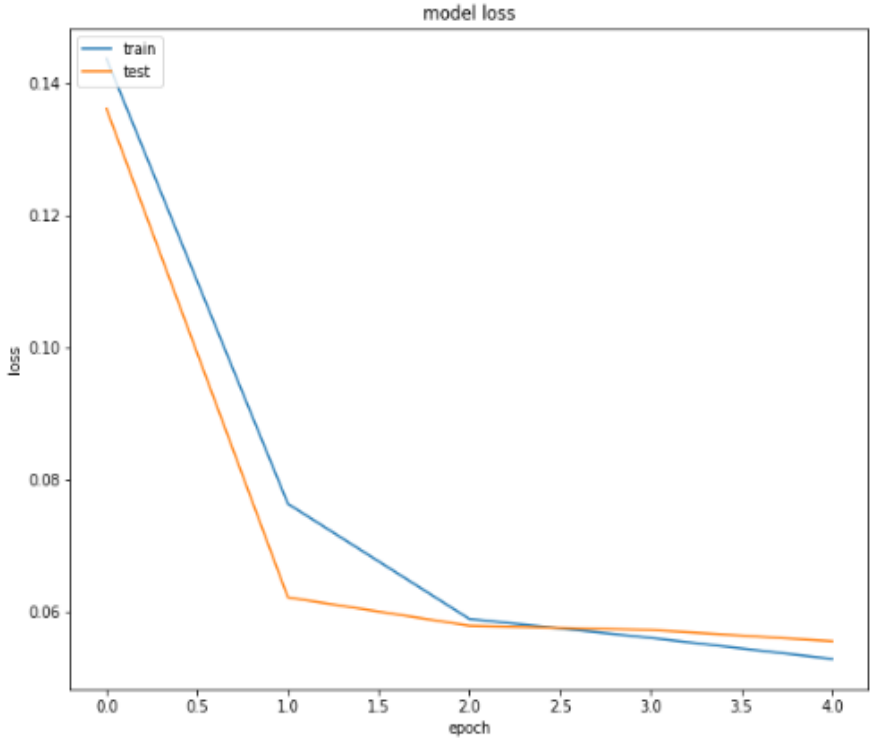
Apêndices

Apêndice A – Plot do TSNE

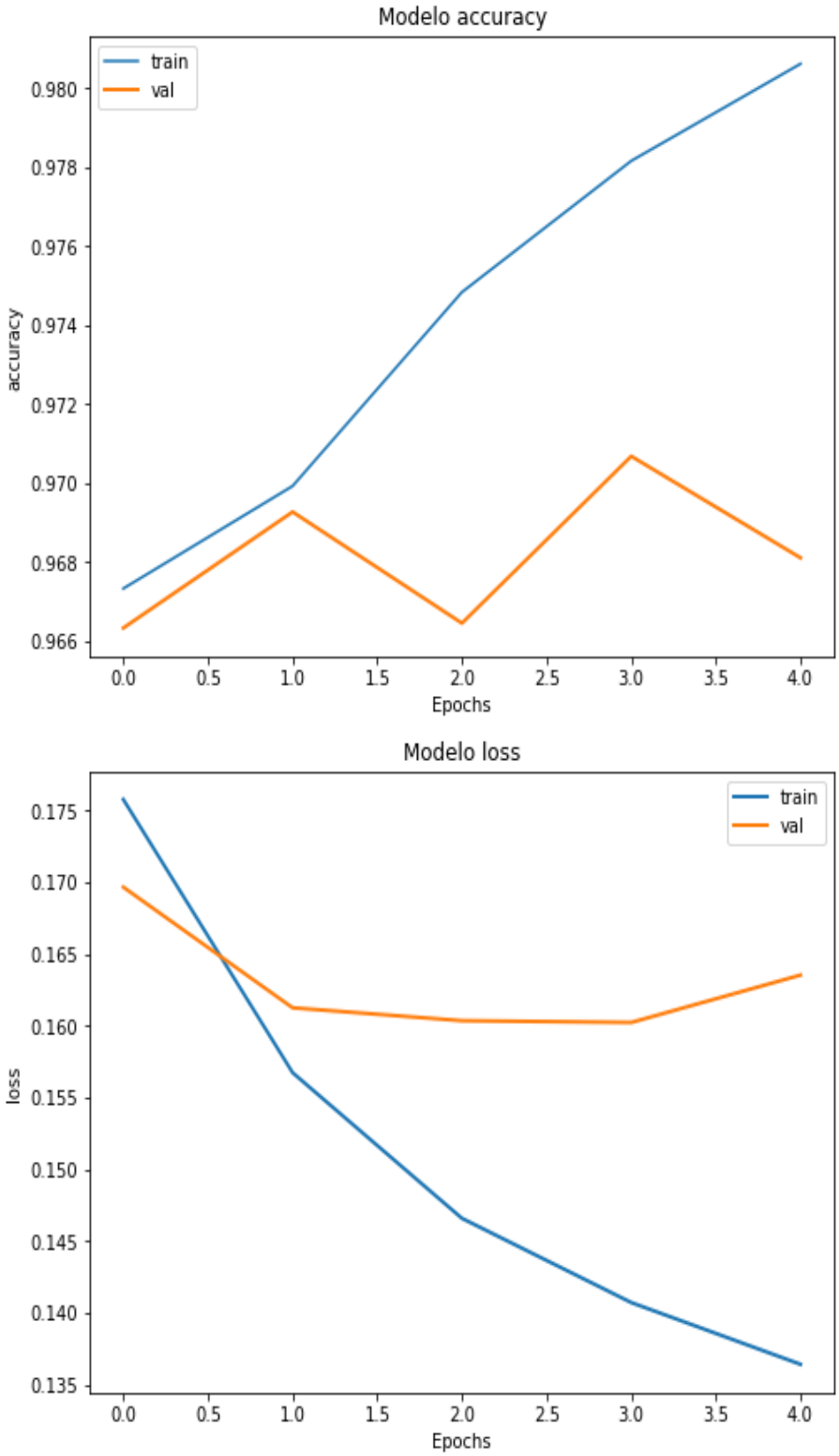


Apêndice B – Accuracy e perda do atributo type

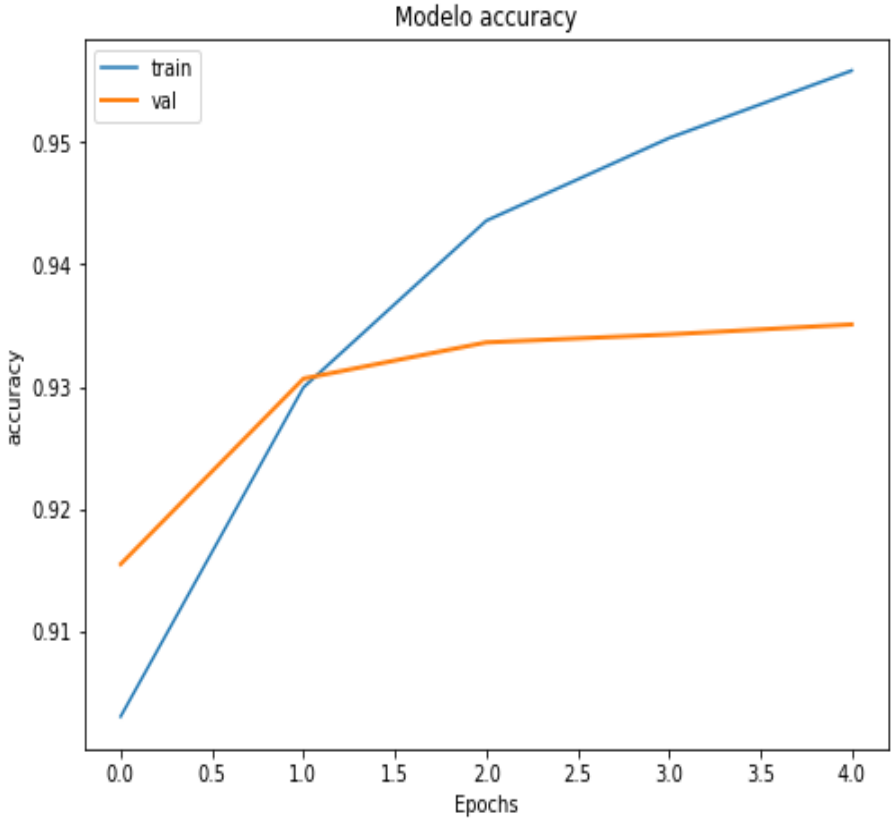


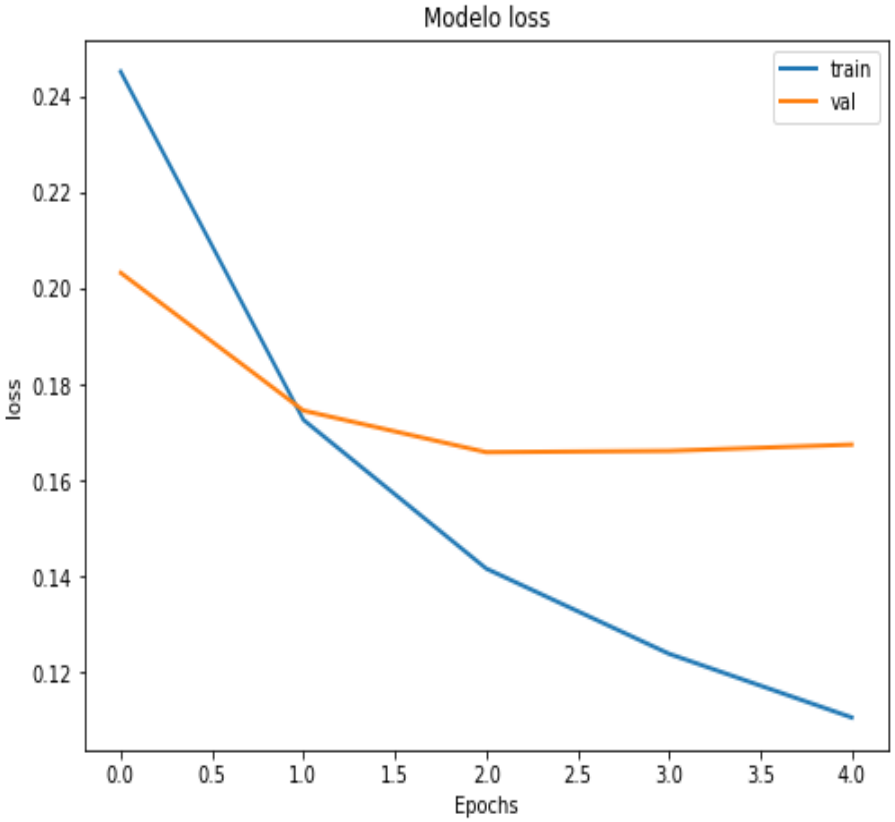


Apêndice C - Accuracy e perda de atributo NPT_Cause

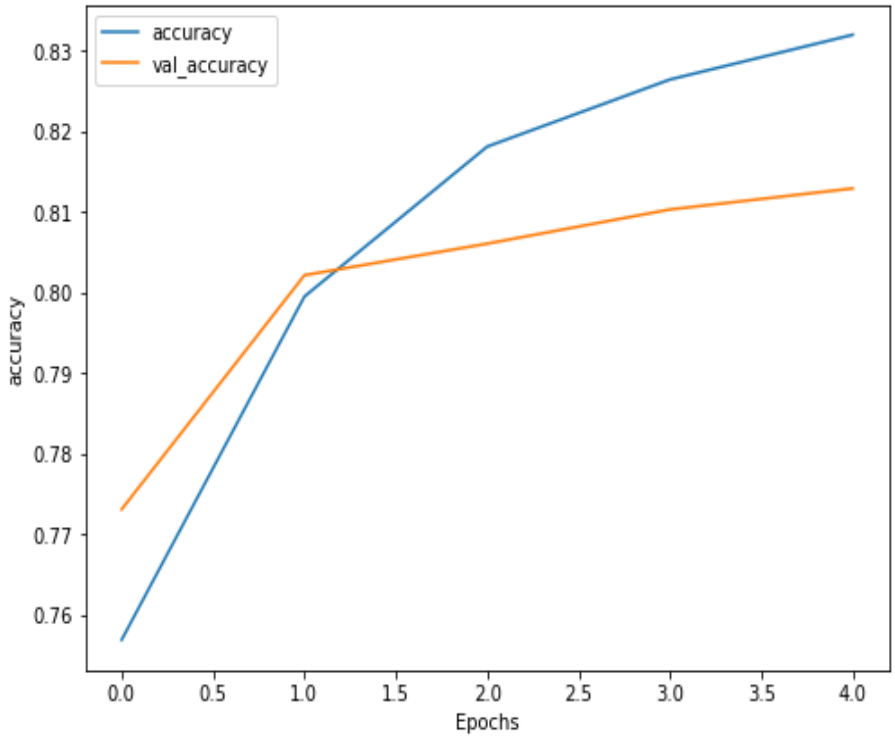


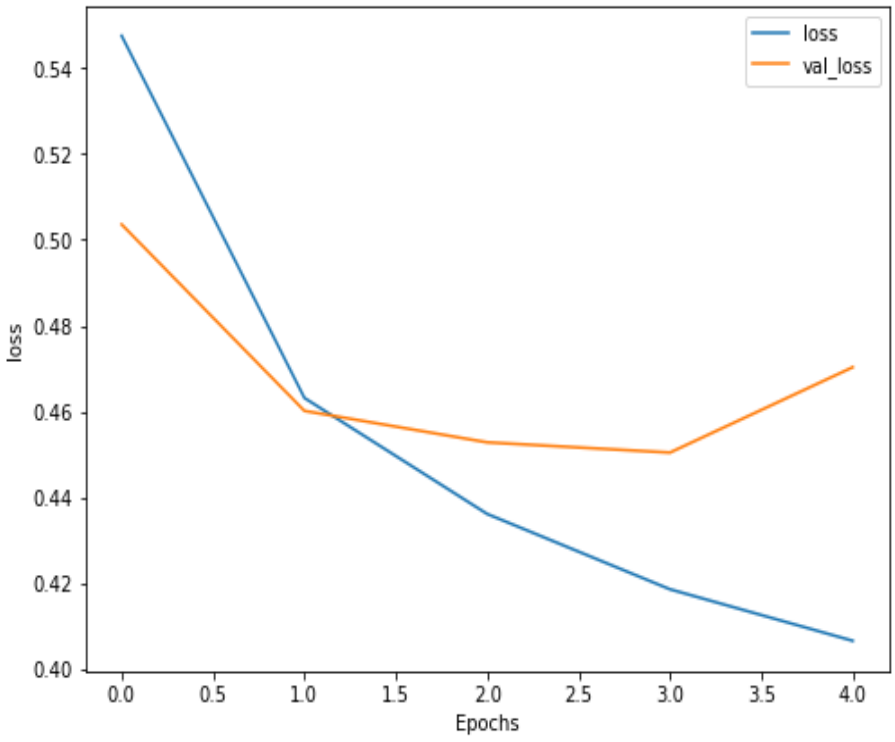
Apêndice D - Accuracy e perda de atributo Phase



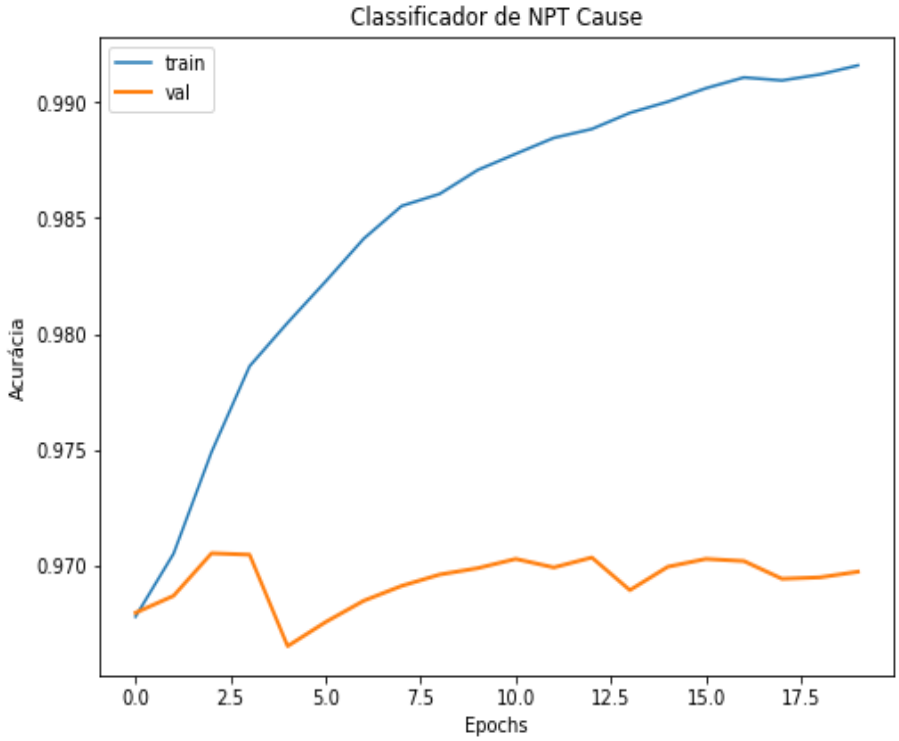


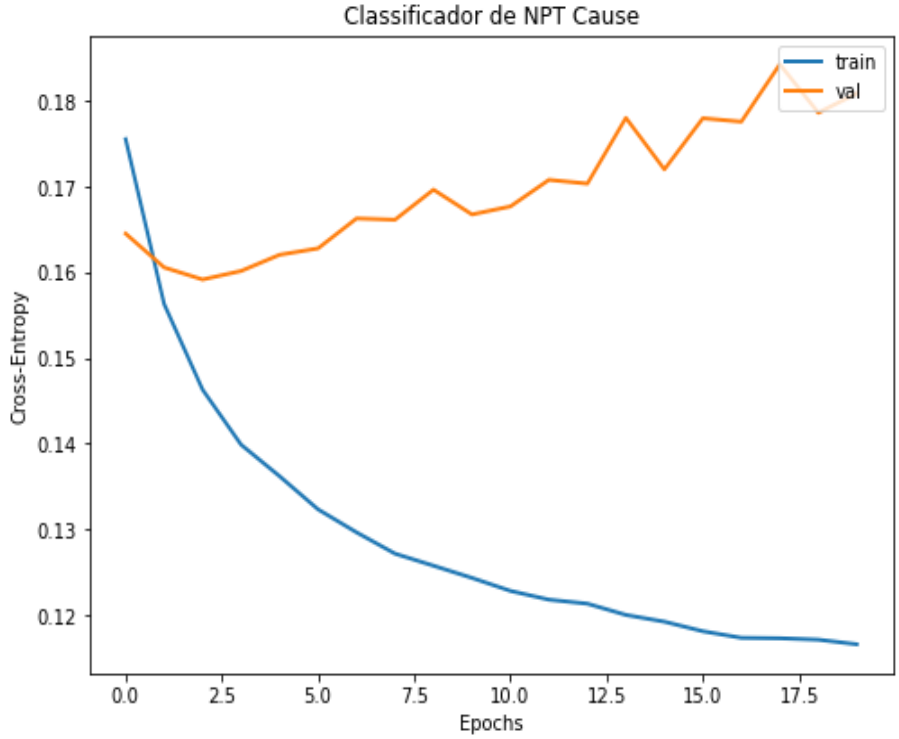
Apêndice E - Accuracy e perda do modelo LSTM no Type



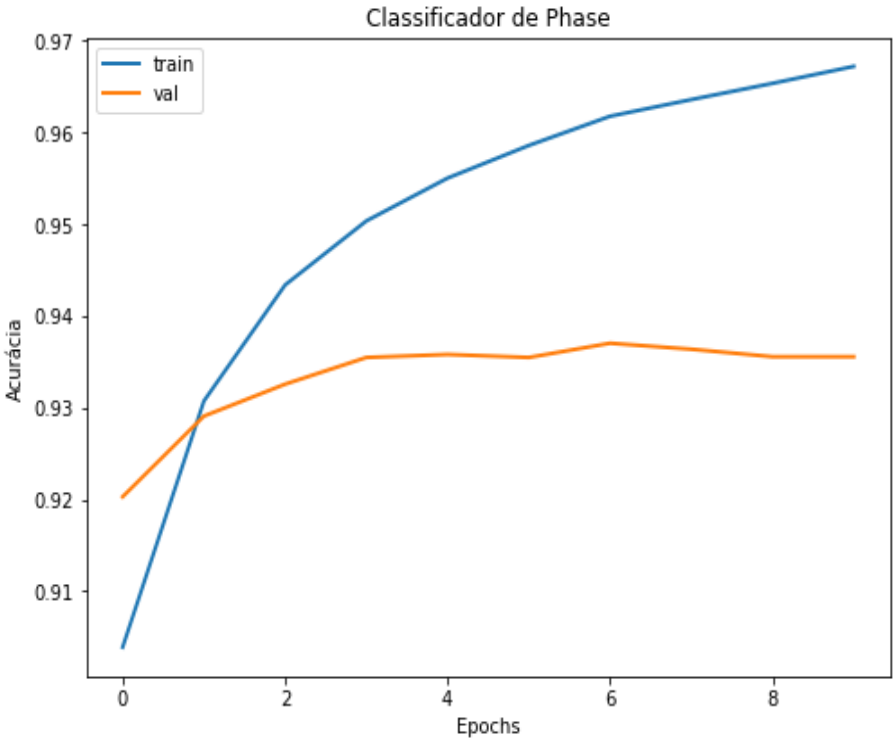


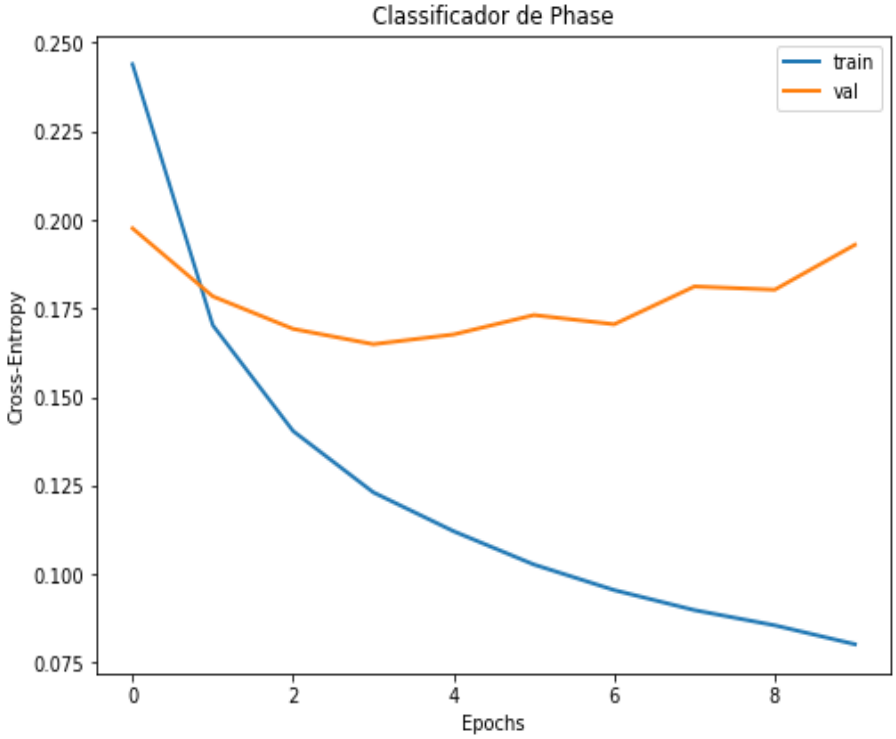
Apêndice F - Accuracy do modelo LSTM na causa de NPT





Apêndice G - Accuracy do modelo LSTM na Phase





Apêndice H – Curva ROC