UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA DE PETRÓPOLIS
CURSO DE ENGENHARIA DE PRODUÇÃO DA ESCOLA DE ENGENHARIA DE PETRÓPOLIS

VICTOR MENDES DA SILVA

MINERAÇÃO DE DADOS APLICADA A UMA LINHA DE MONTAGEM DE MOTORES AERONÁUTICOS A JATO

IDENTIFICANDO CORRELAÇÕES ENTRE ATRIBUTOS DE MONTAGEM E VIBRAÇÃO

PETRÓPOLIS
2020

VICTOR MENDES DA SILVA

MINERAÇÃO DE DADOS APLICADA A UMA LINHA DE MONTAGEM DE MOTORES AERONÁUTICOS A JATO

IDENTIFICANDO CORRELAÇÕES ENTRE ATRIBUTOS DE MONTAGEM E VIBRAÇÃO

Projeto Final apresentado ao Curso de Engenharia de Produção da Escola de Engenharia de Petrópolis da Universidade Federal Fluminense, como parte dos requisitos necessários à obtenção do título de Engenheira(o) de Produção.

Orientador(a):
Prof. D.Sc. Fabio Ribeiro Cerqueira

Petrópolis, RJ
2020

Ficha catalográfica automática - SDC/BCPE
Gerada com informações fornecidas pelo autor

S586m Silva, Victor Mendes da
MINERAÇÃO DE DADOS APLICADA A UMA LINHA DE MONTAGEM DE MOTORES AERONÁUTICOS A JATO : IDENTIFICANDO CORRELAÇÕES ENTRE ATRIBUTOS DE MONTAGEM E VIBRAÇÃO / Victor Mendes da Silva ; Fabio Cerqueira, orientador. Petrópolis, 2020.
94 f. : il.

Trabalho de Conclusão de Curso (Graduação em Engenharia de Produção)-Universidade Federal Fluminense, Escola de Engenharia de Petrópolis, Petrópolis, 2020.

1. Mineração de dados (Computação). 2. Aprendizado de máquina. 3. Aeronáutica. 4. Montagem industrial. 5. Produção intelectual. I. Cerqueira, Fabio, orientador. II. Universidade Federal Fluminense. Escola de Engenharia de Petrópolis. III. Título.

CDD -

Bibliotecário responsável: Sandra Lopes Coelho - CRB7/3389

VICTOR MENDES DA SILVA

MINERAÇÃO DE DADOS APLICADA A UMA LINHA DE MONTAGEM DE MOTORES AERONÁUTICOS A JATO

IDENTIFICANDO CORRELAÇÕES ENTRE ATRIBUTOS DE MONTAGEM E VIBRAÇÃO

Projeto Final apresentado ao Curso de Engenharia de Produção da Escola de Engenharia de Petrópolis da Universidade Federal Fluminense, como parte dos requisitos necessários à obtenção do título de Engenheira(o) de Produção.

Aprovado em 20 de Agosto de 2020, com nota 9, 3 pela banca examinadora.

BANCA EXAMINADORA

_____

Prof. D.Sc. Fabio Riberio Cerqueira – Orientador

UFF

_____

Prof. D.Sc. Anibal Alberto Vilcapoma Ignacio – Membro Convidado

UFF

_____

Prof. D.Sc. Victor Barreto Braga Mello – Membro Convidado

UFF

Petrópolis

2020

Ao meu pai, Rogério, que não pode me ver formar, mas curiosamente formou-me muito antes de qualquer faculdade em algo muito mais difícil que engenharia.

AGRADECIMENTOS

À minha família, por todo o apoio durante este período de estudos, estágios, concursos e cursos. Sem a base que me proporcionaram, jamais estaria escrevendo estas linhas. Em especial a minha mãe, que sempre me incentivou e me garantiu condições para formar-me profissionalmente. A minha irmã, Bia, que me ensinou a ver o mundo de forma diferente. E a meu tio, Carlos, com quem compartilho do interesse por engenharia e por quem tenho profunda admiração profissional.

Ao professor Fabio, por ter me orientado durante esse trabalho de fim de curso.

À Universidade Federal Fluminense e todo seu corpo de funcionários e docentes, que possibilitaram minha formação, aprendizado e desenvolvimento.

“Mostre-me um homem cem por cento satisfeito e eu mostrar-te-ei um fracassado.”

Thomas A. Edison

RESUMO

O presente trabalho explorou a compreensão e investigação de relações entre atributos da montagem de motores *turbofans* aeronáuticos e suas respectivas assinaturas de vibração quando testados em banco de provas com especial foco em vibração do eixo de alta rotação (N2), relacionada ao eixo da turbina de alta (HPT) com o compressor de alta (HPC). Sabe-se que essa vibração costuma ser o parâmetro mais crítico no que diz respeito a possíveis rejeições pós montagem. Foram utilizadas, portanto, ferramentas e técnicas de mineração de dados para melhor compreender e investigar quais as relações possíveis entre parâmetros de montagem e vibração. Além disto, foram propostos, através dos resultados encontrados, modelos de previsão de vibração. Exploraram-se, nessa ordem, uma vasta revisão bibliográfica no que diz respeito à mecânica e funcionamento de motores aeronáuticos, uma revisão acerca de fenômenos vibratórios, assim como uma revisão bibliográfica referente a técnicas e algoritmos de análise e mineração de dados. Por fim, os conhecimentos discutidos foram aplicados em um caso real, estudando uma linha de montagem de *turbofans*.

Palavras-Chave: *Turbofan*. Motores Aeronáuticos. Motores A Jato. Aviação. Vibração. N2. Mineração de Dados. Aprendizado de Máquina.

ABSTRACT

The aim of this study was to better understand and investigate relations between the assembly process of aeronautical turbofan engines and their vibrations signatures when tested. Especially the N2 vibration, which is related to the high-pressure turbine (HPT) and high-pressure compressor (HPC) axis. It is known that this type of vibration is the most critical and a major cause of rejections in quality tests. Therefore, data mining tools and techniques were used to better understand and investigate the possible relationships between assembly and vibration parameters as well as to develop a prediction model to classify engines before test. This study was conducted according to the following steps: first, a vast bibliographic review of the mechanics and aeronautical engines design were explored. Next, a bibliographic review regarding data mining and data analysis algorithms was performed. Finally, the knowledge discussed were applied on a real case involving a turbofan assembly line.

Keywords: Turbofan. Aircraft Engines. Jet Engines. Aviation. Vibration. N2. Data Mining. Machine Learning.

LISTA DE FIGURAS

Figura 1 – Eixos de rotação (esquema). 23
Figura 2 - Turbojato (esquema). 24
Figura 3 - Turbopropulsor (esquema). 25
Figura 4 - Turboshaft (esquema). 25
Figura 5 - Turbofan (esquema). 27
Figura 6 - Eficiência de propulsão em diferentes motores de acordo com velocidade de operação. 29
Figura 7 - KDD (processos). 34
Figura 8 - KDD (esquema). 35
Figura 9 - Esquema CRISP-DM. 36
Figura 10 - Etapas de Pré-Processamento. 38
Figura 11 - Método de Suavização por Intervalo de Classes ou *Binning*. 41
Figura 12 - Exemplo visual de agrupamento de dados em *clusters*. 48
Figura 13 - Ilustração de Rede Neural. 51
Figura 14 - *Fuzzy*. 52
Figura 15 - Curva ROC. 58
Figura 16 - Curva ROC, Diagonal de Aleatoriedade. 59
Figura 17 - Parte do Código VBA utilizado. 63
Figura 18 - Histograma - Atributo 29. 64
Figura 19 - Histograma – Atributo 62. 65
Figura 20 - Distribuição de motores, Classe vs. Atributo 6. 65
Figura 21 - Distribuição de motores, Classe vs. Atributo 8. 66
Figura 22 - Distribuição de motores, Atributo 77 vs. Atributo 29. 67
Figura 23 - Distribuição de motores, Atributo 19 vs. Atributo 21 vs. Atributo 43. 67
Figura 24 - Resultados do Algoritmo CfsSubset. 69
Figura 25 - Resultados do Algoritmo InfoGain. 70
Figura 26 - Matriz de Resultados dos Algoritmos. 75
Figura 27 - Comparativo de Métricas para o Algoritmo Naive Bayes em 3 cenários. 79
Figura 28 - Comparativo de Métricas para o Algoritmo kNN em 3 cenários. 80
Figura 29 - Comparativo de Métricas para o Algoritmo Adaboost em 3 cenários. 81
Figura 30 - Comparativo de Métricas para o Algoritmo J48 em 3 cenários. 82
Figura 31 - Comparativo de Métricas para o Algoritmo Random Forest em 3 cenários. 83

Figura 32 - Comparativo de Métricas para o Algoritmo Random Forest em 3 cenários. 84
Figura 33 - Resultados do Modelo de Predição Random Forest + CfsSubset + SMOTE. 86
Figura 34 - Resultados do Modelo de Predição InfoGain + SMOTE + J48. 87
Figura 35 - Árvore de Decisão do Modelo de Predição InfoGain + SMOTE + J48. 87

LISTA DE TABELAS

Tabela 1 - Matriz de Confusão. 57
Tabela 2 - Comparação F-valor por Algoritmo de Classificação. Para o kNN, considerou-se os 10 vizinhos mais próximos. 71
Tabela 3 - Comparação de Melhora do F-valor médio pós Seleção de Atributos. 71
Tabela 4 - Composição dos Cenários de Estudo. 74
Tabela 5 - Cenários e suas métricas formatados condicionalmente. 76
Tabela 6 - 10 Melhores Valores Para Cada Métrica. 77
Tabela 7 - Resultados Para Classificação de Instancias Positivas. 78
Tabela 8 - Comparação Final Entre Combinação do Melhor Cenário Para Cada Algoritmo. . 85
Tabela 9 - Tabela Comparativa NaiveBayes em cenário 4 e Random Forest em Cenário 5. ... 85

LISTA DE ABREVIATURAS E SIGLAS

ABEPRO	Associação Brasileira de Engenharia de Produção
APU	*Auxilary Power Units*
AUC	Area *Under Curve*
BPR	*Bypass Ratio*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CART	*Classification and Regression Trees*
Cfs	*Correlation-based feature selection*
CMCS	*Central Maintenance Computer System*
COBREM	Congresso Brasileiro de Educação Médica
CRISP-DM	*Cross-Industry Standard Process of Data Mining*
DWT	*Discrete Wavelet Transform*
EGBT	*Exhaust Gas Temperature*
EGT	*Exhaust Gas Temperature* e empuxo
FMU	*Fuel Metering Unit*
FN	*False Negative*
FP	*False Positive*
FTY	*First Time Yield*
FTY	*First Time Yield*
GE	*General Electric*
GNU	*General Public License*
HPC	*High Pressure Compressor*
HPT	*High Pressure Turbine*
IATA	*International Air Transpor Association*
KDD	*Knowledge Discovery From Data*
kNN	*k-Nearest Neighbors*
LPC	*Low Pressure Compressor*
LPT	*Low Pressure Turbine*
N2	Eixo de alta rotação
PCA	*Principal Components Analysis*
ROC	*Relative Operating Characteristics*
SMO	*Sequential Minimal Optimization*

SMOTE	*Synthetic Minority Over-sampling Technique*
SVD	*Singular Value Decomposition*
SVM	*Support Vector Machines*
TN	*True Negative*
TP	*True Positive*
TS	*Test Set*

LISTA DE SÍMBOLOS

m	Metro
km	Quilometro (s)
h	Hora
kg	Quilograma (s)
$	Cifrão
%	Porcentagem
N	Newton
kN	Quilonewton (s)
lb	Libra (s)
∞	Infinito
>	Maior que
<	Menor que
β	Importância relativa entre precisão e sensibilidade
k	Número inteiro positivo

SUMÁRIO

Sumário **1**
1 INTRODUÇÃO **17**
1.1 Contextualização do problema de pesquisa **17**
1.2 Objetivos **18**
1.2.1 Objetivo geral 18
1.2.2 Objetivos específicos 18
1.3 Questões **18**
1.3.1 Questão central 18
1.3.2 Questões específicas 19
1.4 Justificativa **19**
1.5 Classificação da pesquisa **19**
1.6 Pesquisa bibliográfica **20**
1.7 Pesquisa de campo **20**
2 REVISÃO BIBLIOGRÁFICA **21**
2.1 Motores aeronáuticos a jato: *design*, mecânica, aplicações e funcionamento **21**
2.1.1 Principais construções e tipos de motores a jato 23
2.1.2 Regimes de operação 28
2.2 Vibração **29**
2.2.1 Vibração em sistemas aeronáuticos 30
2.2.2 Desbalanceamento 31
2.3 KDD – descoberta de conhecimento em bases de dados **32**
2.3.1 Tratamento dos dados 37
2.4 Mineração de dados apoiada por algoritmos de aprendizado de máquina **47**
2.4.1 Problemas típicos 47
2.4.2 Algoritmos de classificação 49
2.4.3 Mineração de dados e desbalanceamento 53
2.5 Softwares **59**
3 ESTUDO DE CASO **60**
3.1 Abordagem inicial **61**
3.2 Preparação **61**
3.2.1 Integração 62
3.3 Mineração de dados **63**

3.3.1 Análise inicial 63
3.3.2 Pré-processamento, seleção de atributos 68
3.3.3 Construção de modelos preditivos para diversos cenários 72
3.4 Análise comparativa dos resultados 75
3.4.1 Análise comparativa da performance para cada algoritmo de classificação 78
3.4.2 Etapa final de comparação 84
4 CONCLUSÃO 88
5 LIMITAÇÕES DA PESQUISA E TRABALHOS FUTUROS 89
5.1 Limitações da pesquisa 89
5.2 Trabalhos futuros 90
REFERÊNCIAS 91

1 INTRODUÇÃO

Segundo dados da (IATA,2019) a Associação Internacional de Transporte Aéreo, em 2019 foram transportados cerca de 4,54 bilhões de passageiros em transportes aéreos, tendo a indústria faturado mais de 87 bilhões de dólares. Empregando cerca de 2.9 milhões de trabalhadores. O objeto desse estudo é um dos componentes centrais dessa indústria. Motores aeronáuticos.

O processo de fabricação de motores aeronáuticos a jato é bastante complexo, envolvendo uma grande quantidade de peças e operações de montagem que são fundamentais para a garantia de eficiência e segurança dos mesmos. Após montagem, todos os motores são testados e devem estar dentro de faixas extremamente rígidas de uma série de parâmetros, entre os quais, destacam-se como mais importantes: vibração, EGT – *Exhaust Gas Temperature* e empuxo.

Caso os motores sejam rejeitados durante teste, a empresa sofrerá grande impacto financeiro e de planejamento uma vez que, em muitos casos, será necessário desmontar e remontar determinados módulos a fim de garantir a aprovação do motor. Uma das métricas mais importante para empresa é o índice de FTY (*First Time Yield),* índice de aprovações diretas no banco de provas. No entanto, não existem modelos de previsão de rejeições ou conhecimento de quais são as medidas de montagem mais relevantes para garantir a aprovação de motores quando testados.

É nesse contexto que se encontra o presente projeto que será fundamentado através de revisão bibliográfica da literatura atual. Inicialmente será feita uma revisão de literatura a respeito do funcionamento, construção e aplicação de motores aeronáuticos a jato. Em seguida, serão abordados os fenômenos de vibração em sistemas mecânicos. Por fim, serão abordados os principais algoritmos e metodologias utilizadas em problemas de mineração de dados.

Ao final do trabalho, serão aplicados conhecimentos desenvolvidos durante a revisão de literatura em um estudo de caso real em uma linha de produção de motores *turbofan*.

1.1 Contextualização do problema de pesquisa

A montagem de motores aeronáuticos a jato é de extrema complexidade dependendo da utilização de equipamentos, seres humanos, instruções de montagem e softwares. Segundo (Pereira, 2017), em alguns motores de aeronaves o número total de peças pode chegar a 12.000 itens. Tamanha complexidade está relacionada às variáveis e riscos que permeiam todos os processos de montagem.

1.2 Objetivos

Serão descritos, a seguir, o conjunto de objetivos gerais e específicos do presente trabalho.

1.2.1 Objetivo geral

O presente projeto visa reduzir a quantidade de rejeições de motores em testes de pós montagem relacionados ao parâmetro de vibração trazendo ganhos para a empresa estudada por meio da compreensão de relações possivelmente desconhecidas e desenvolvimento de modelos de previsão de vibração. Ao mesmo tempo, como estudo acadêmico, pretende fomentar questões e descobertas científicas que possam ser utilizadas ou mesmo incrementadas em oportunidades futuras.

1.2.2 Objetivos específicos

Como objetivos específicos destacam-se: 1- Identificação dos atributos mais críticos durante o processo de montagem de motores a fim de garantir padrões de vibração dentro dos desejados durante teste. 2- Construção de modelos preditivos capazes de gerar a melhor classificação de instâncias no cenário estudado.

1.3 Questões

Nesse subtópico serão descritas as questões centrais e específicas que servirão de suporte para a pesquisa desenvolvida.

1.3.1 Questão central

Quais conjuntos de atributos são mais influentes no excesso de vibração do eixo de alta rotação do motor (N2) e qual conjunto de algoritmos é capaz de gerar o melhor modelo de classificação/predição de instâncias para o cenário estudado?

1.3.2 Questões específicas

- O que é vibração e como ela pode afetar a operação do motor?
- Quais os parâmetros de montagem que mais influenciam a performance de vibração do eixo de alta rotação (N2) dos motores aeronáuticos a jato, do ponto de vista da Engenharia?
- Como utilizar análise de dados e algoritmos para quantificar as correlações entre os parâmetros de montagem e a vibração N2 dos motores, definindo os parâmetros mais críticos?
- Qual algoritmos melhores se relacionam com a questão, de modo que seja possível propor um modelo de previsão de vibração com base em parâmetros de montagem?

1.4 Justificativa

A presente pesquisa apresenta duas contribuições principais. A primeira de ordem teórica, ao sistematizar o conjunto de conhecimentos e informações necessárias para compreensão e resolução do problema abordado no formato de uma revisão de literatura concisa e objetiva que diga respeito ao tema estudado fomentando assim o conhecimento acadêmico e sua aproximação com o setor privado. A segunda justificativa é de ordem prática e diz respeito aos possíveis benefícios obtidos pela empresa parceira. Nesse quesito, destacam-se o maior controle dos parâmetros de montagem e maior previsibilidade dos resultados de vibração do eixo de alta rotação (N2) nos bancos de provas. Com isso, potencialmente tendências serão identificadas e ações preventivas adotadas, gerando melhorias internas que afetarão positivamente o desempenho das operações. Segundo (Turroni; Mello, 2012), é contribuição de uma pesquisa-ação gerar teoria emergente a partir de uma síntese do que emerge dos dados e dos resultados da teoria aplicada.

1.5 Classificação da pesquisa

Ao desenvolver uma pesquisa, é importante, segundo a metodologia proposta por (Turroni; Mello, 2012), situá-la mediante alguns critérios a fim de garantir um melhor desenvolvimento. São eles: natureza, objetivos e abordagem. Quanto à natureza, a pesquisa classifica-se como aplicada, devido ao interesse prático e na solução de problemas reais. Já em relação aos objetivos, classifica-se como exploratória já que proporciona familiaridade com o

problema e é sustentada por revisão bibliográfica adequada. Quanto à abordagem, é combinada, ao envolver elementos qualitativos e quantitativos. Já em relação aos métodos, é Pesquisa-Ação, pois a observação é ativa, envolvendo não apenas o estudo, mas também a aplicação e contato direto com especialistas do setor, havendo intervenção no objeto da pesquisa – direcionado a uma empresa multinacional de fabricação de motores aeronáuticos a jato.

1.6 Pesquisa bibliográfica

A metodologia utilizada será baseada nos estudos de Turrioni e Mello (2012). A busca das referências bibliográficas dar-se-á predominantemente através do Portal da CAPES e suas diversas bases, tais como: Scopus, Scielo e Informs. Também serão utilizadas plataformas como: Google Scholar, bancos de teses e bancos de dissertações de algumas universidades. Além disso, sites de Congressos, como o da ABEPRO e COBEM também serão consultados. A finalidade é coletar artigos, teses, dissertações e livros nacionais e internacionais. As referências bibliográficas dos primeiros arquivos encontrados também servirão de base para a busca de novas fontes e refinamento dos mecanismos de busca e palavras chaves que guiarão a pesquisa.

1.7 Pesquisa de campo

De acordo com a metodologia escolhida, os dados da empresa serão coletados de duas fontes principais: os que Turrioni e Melo, (2012) chamam de “peso-leve”, que serão coletados através de entrevistas, observações e discussões entre mecânicos, engenheiros e gestores da produção. Esse tipo de dados é de grande relevância uma vez que consideram a experiência e visão dos processos por meio da percepção dos agentes, o que é de grande relevância. E os “pesos-pesados”, dados predominantemente quantitativos, que irão ser traduzidos no formato de séries históricas e planilhas, no caso específico, em especial os parâmetros relacionados a montagem dos motores, chamados de “*Data Names*” e das assinaturas de vibração do eixo de alta rotação (N2) durante teste.

2 REVISÃO BIBLIOGRÁFICA

A revisão bibliográfica será dividida em três grandes subtópicos, no que ser refere a motores a aeronáuticos a jato (subtópico 2.1), no que ser refere a vibração (subtópico 2.2) e no que se refere a mineração de dados (subtópico 2.3).

2.1 Motores aeronáuticos a jato: *design*, mecânica, aplicações e funcionamento

Para (Ferreira, M. J. B. et al, 2009) "Desde sua origem no início do século XX, com os voos do 14 Bis, de Santos Dumont, e do Flyer, dos irmãos Wright, a evolução da indústria aeronáutica mundial tem se dado através da introdução de inovações tecnológicas."

Um motor aeronáutico é aquele capaz de gerar força de empuxo suficiente para propulsão de aeronaves. De maneira geral existem três tipos possíveis de motores de aviação: motores recíprocos (de pistão), motores a jato (grande maioria) e motores elétricos. Estudaremos unicamente os motores a jato, uma vez que são os motores mais utilizados na atualidade.

Para Hünecke (2003), os motores a jato podem ser classificados de acordo com alguns critérios, sendo os principais: número de *spools*, princípio de compressão, distribuição do fluxo de ar e utilização dos gases de exaustão. Estes motores têm seus regimes de funcionamento e eficiência térmica representados pelo ciclo termodinâmico de Brayton, existindo 4 tipos principais, são eles: *Turbojet* (Figura 2), *Turboprop* (Figura 3), *Turboshaft* (Figura 4) e *Turbofan* (Figura 5).

Todos convertem energia química armazenada como combustível em energia térmica que é convertida em energia mecânica, gerando empuxo. No entanto, cada motor possui uma mecânica e design diferenciada, de modo a melhor operar em determinadas condições.

Via de regra, certos componentes são essenciais para o funcionamento de qualquer um desses motores, são eles os compressores de alta e de baixa pressão, a câmara de combustão e as turbinas de alta e baixa pressão. Cada um desses elementos será definido e abordado para melhor compreensão.

O compressor de baixa pressão, usualmente chamado de *Booster* ou LPC (*Low Pressure Compressor*) é responsável por direcionar e dar início à compressão mecânica do ar através de estágios consecutivos de *blades* rotores, e *vanes* estatores. A cada estágio, as *blades* diminuem de tamanho de maneira que o ar seja comprimido e tenha sua densidade elevada. Uma vez dada início a compressão no LPC, o ar é direcionado para um segundo estágio de

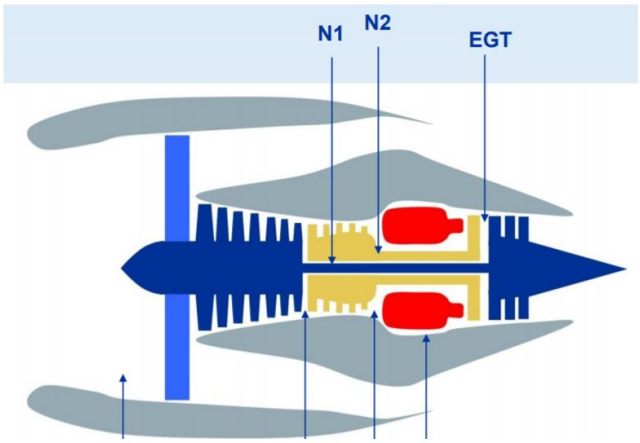
compressão que ocorrerá no HPC. O mecanismo de compressão é o mesmo, no entanto, as taxas de compressão são bem maiores que as atingidas pelo compressor de baixa, sendo elas variáveis de acordo com o tipo e design do motor, normalmente sendo proporcionais à quantidade de estágios em cada compressor. Altamente comprimido, o ar do HPC será direcionado à câmara de combustão, onde bicos injetores, também chamados de *Fuel Nozzles* injetarão o combustível necessário, de acordo com parâmetros de compressão, temperatura e empuxo demandado. A forma com que a combustão e a injeção de combustível ocorrem é bem complexa, sendo comandada pelo sistema FMU, *Fuel Metering Unit.*

Uma vez que a queima ocorre na câmara de combustão, a energia química armazenada no combustível é convertida, em sua maior parte, em energia térmica que irá ser convertida em energia mecânica por meio das turbinas e a passagem das descargas de exaustão. Em primeiro lugar, tem-se a atuação do HPT, *High Pressure Turbine*, que irá converter a energia da combustão e seus gases de exaustão em energia mecânica rotacional por meio de estágios rotores de *blades* e estatores de *Nozzles*, cujo objetivo é direcionar o ar. Um segundo estágio de descompressão, LPT, irá guiar os gases resultantes da queima para a exaustão, descomprimindo-os por meio do sistema de *blades* e *vanes* que aumentam de tamanho a cada estágio, forçando uma descompressão gradual.

De maneira geral, a maioria dos motores a jato, com algumas exceções, possuem dois principais eixos de rotação, o chamado eixo “N1” de baixa rotação, conectado ao compressor e turbina de baixa pressão. E o chamado eixo “N2”, que se conecta ao compressor e turbina de alta pressão. Cada eixo opera de forma independente, sendo “N1” e “N2” também utilizados para se referir ao número de rotações por minuto de seus respectivos eixos, servindo como parâmetros de controle relevantes para ambos pilotos, quando motores estão em operação, como para as montadoras, quando os motores estão em fase de teste.

O eixo de alta rotação induz, por arrasto, a rotação do eixo de baixa rotação e, por isso, quanto menor a rotação de N2 necessária para gerar o empuxo estipulado, melhor a eficiência do motor estudado. A Figura 1 ilustra os diferentes eixos de rotação de um motor *turbofan*.

Figura 1 – Eixos de rotação (esquema).



Fonte: (NASA, 2020)

Esse é o sistema simplificado de funcionamento de um motor a jato, baseado em Hünecke (2003). A seguir, serão abordadas as peculiaridades de cada tipo de design.

2.1.1 Principais construções e tipos de motores a jato

2.1.1.1 Turbojato

Os motores *Turbojet*, ou Turbojato, foram os primeiros motores a turbo-propulsão. O funcionamento desse tipo de motor pode ser resumido da seguinte forma. O ar avança pelo *intake* que tem por objetivo destinar um fluxo de ar constante e contínuo para o compressor. O compressor é um componente mecânico composto por um conjunto de estágios compostos por pares de discos, rotores e *vanes* estatoras, que têm por objetivo aumentar a pressão do ar elevando sua densidade e temperatura.

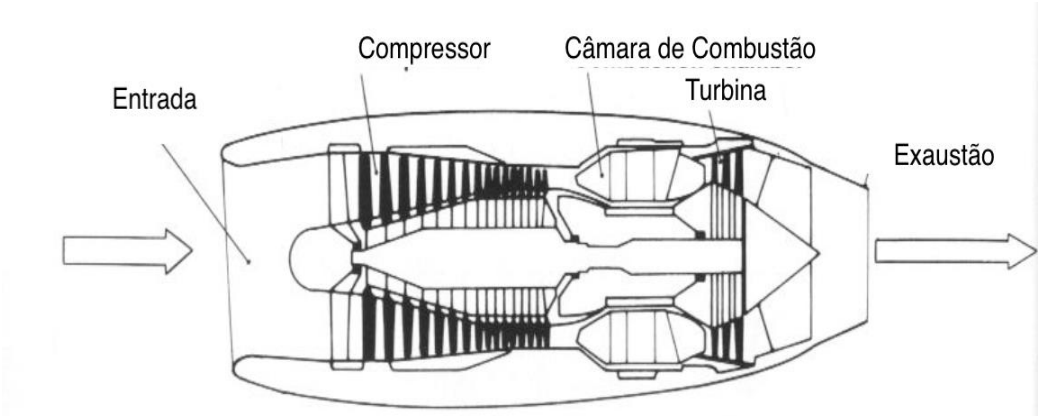
O ar pressurizado entra na *combustion chamber* (câmara de combustão), onde o combustível é injetado através de *fuel nozzles*, dando início à queima da mistura e aumento da temperatura dos gases. A energia dessa queima é então transferida para mecanismos que irão transformar essa energia em trabalho mecânico. Um desses componentes é a *gas turbine* (Turbina) que está atrelada ao mesmo eixo de rotação do Compressor. O objetivo da turbina é converter a energia dos gases em rotação do próprio compressor, assim como outros mecanismos necessários para a operação do motor.

Mesmo após gerar energia mecânica os gases de exaustão ainda possuem energia, nesse momento os gases são direcionados para a *Exhaust Nozzle*, que é responsável por

converter energia térmica e pressão em velocidade. Alta velocidade de exaustão é um pré-requisito para gerar empuxo nessa construção.

A velocidade de exaustão pode ainda ser aumentada através da utilização de *afterburning* ou *thrust augmentation*, uma segunda ingestão de combustível aumentando ainda mais a temperatura de exaustão e, portanto, a velocidade de expulsão dos gases. Um exemplo de motor que obedece a esses regimes é o General Electric J79, utilizado em aeronaves de combate como o *Starfighter* e *Phantom*. A Figura 2 ilustra o funcionamento de um turbojato.

Figura 2 - Turbojato (esquema).

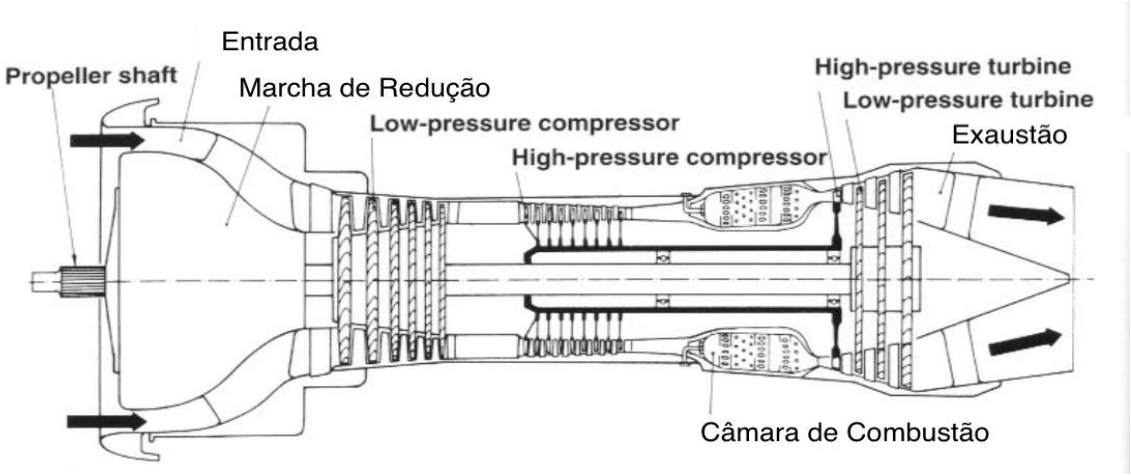


Fonte: Adaptado de Hünecke (1997, p.4)

2.1.1.2 Turbopropulsor

Como todo turbo-motor, os *turboprops* também se utilizam do conjunto: Compressor, Combustor e Turbina - conjunto conhecido como "*gas generator*". Embora o processo de funcionamento seja bastante semelhante ao dos motores *turbojet*, os *turboprops* se distinguem pela existência de uma turbina adicional que irá comandar um *propeller*, um conjunto de dois *spools* de rotação mecânica e uma marcha de redução mecânica que irá converter altas velocidades de rotação da turbina em velocidades mais moderadas de rotação do *propeller*. A Figura 3 ilustra o funcionamento de um turbopropulsor.

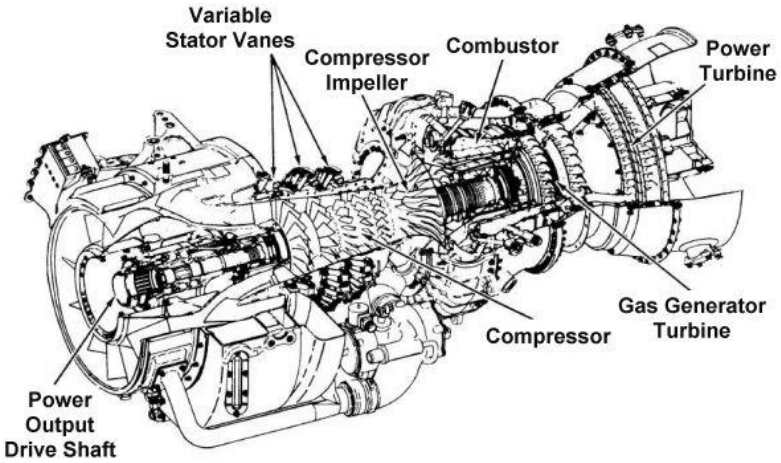
Figura 3 - Turbopropulsor (esquema).



Fonte: Adaptado de Hünecke (1997, p.8)

2.1.1.3 Turboshafts

Figura 4 - Turboshaft (esquema).



Fonte: (Aerospaceweb, 2005)

Os Motores *Turboshaft* (Figura 4) são similares ao *turboprop*. A principal diferença está na função da segunda turbina. Uma vez que os gases saem do conjunto compressor-combustor, são direcionados para duas turbinas distintas. A primeira alimenta o compressor, a segunda alimenta um *shaft horsepower* que irá fornecer energia para as hélices através de um sistema de transmissão, ao invés de fornecer energia para o *propeller* como ocorre nos *turboprops*. Muitos *turboshafts* são utilizados em sistemas chamados de APU (*Auxilary Power*

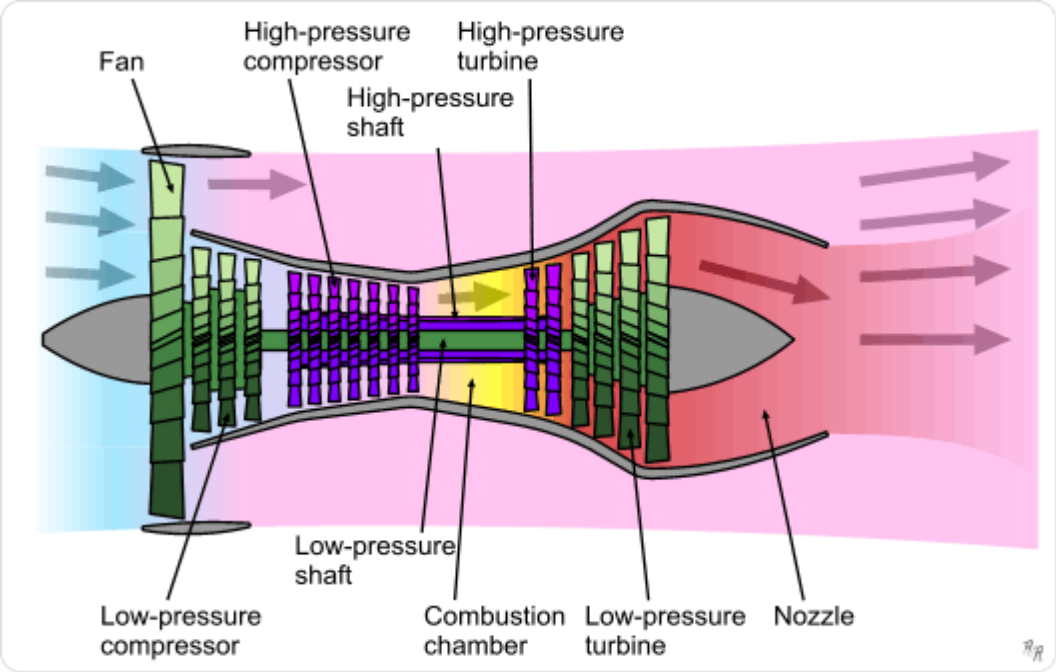
Units). Servindo como fontes secundárias de energia provendo sistemas pneumáticos e, ou, elétricos durante o voo ou quando a aeronave está no solo. Além da aviação, também são utilizadas em geradores estacionários de energia, navios e tanques militares.

2.1.1.4 Turbofans

Os *turbofans* são motores a reação altamente eficazes, utilizados largamente nas aeronaves comerciais onde baixos níveis de ruído e alta eficiência e confiabilidade são de extrema importância. São projetados para atuar em altitudes elevadas, entre 10.000 metros e 15.000 metros, operando entre velocidades de 700 km/h até 1000km/h (NASA, 2015).

Diferente dos motores *turbojet* em que todo o ar ingerido percorre a turbina até a câmara de combustão, em um *turbofan* parte do ar ingerido é desviada, passando apenas pelo *fan* ou *low compressor*. A relação entre a massa de ar deslocada para fora da turbina e massa de ar que de fato percorre a turbina é chamada de "*bypass ratio*" ou BPR. Um BPR de 10:1 significa, por exemplo, que enquanto 10 kg de ar passam pelo *fan*, 1 kg de ar percorre o *core* do motor. O empuxo gerado por esses motores é uma combinação dos dois processos, tanto do ar que passa pelas turbinas quanto do ar que escapa pelo *fan*. Motores em que o empuxo é gerado em sua maior parte pelo ar que percorre a turbina são chamados de *low-bypass turbofans*, ao passo que os motores em que a maior parte do empuxo é gerada pelo ar que passa pelo *fan* são chamados de *high-bypass*. A grande maioria dos motores utilizados hoje na aviação comercial são do segundo tipo. Em contrapartida, a maior parte dos motores de uso militar são do tipo *low-bypass*.

Figura 5 - Turbofan (esquema).



Fonte: (Wikipedia, 2020)

2.1.1.4.1 Turbofans de high bypass-ratio

Os motores desse tipo trouxeram em sua construção e design modificações que possibilitaram grande economia de combustível e, rapidamente, tomaram conta do mercado da aviação comercial. Os primeiros motores desse tipo foram utilizados no setor militar para o transporte de cargas. A tecnologia rapidamente ganhou força e aplicabilidade no fim da década de 60 entrando em operação em aeronaves como *Boeing* 747, Lockheed L-1011 TriStar e McDonnell Douglas DC-10. Motores com essa construção tem seu empuxo gerado em maior parte pela massa de ar que não é queimada, mas deslocada pelos *fans*. Além da economia de combustível esses motores apresentam baixos níveis de ruídos quando em operação uma vez que a velocidade de exaustão dos gases não é tão grande, como no caso de motores turbojato.

A tecnologia foi desenvolvida durante os anos de 1940, mas foi por volta de 1960 que foi aperfeiçoado pelos Estados Unidos com o icônico GE TF39, o primeiro motor *high bypass* da história. No início da década, o governo americano sediava as duas maiores fabricantes de motores aeronáuticos do mundo, General Electric e Pratt & Whitney. A força área americana, com objetivo de fomentar a indústria, deu início a uma competição para o desenvolvimento de motores com altas taxas de fuga, isso é, "*high bypass-ratio*". Em agosto de 1965 a General Electric ganhou a competição e por consequência um dos maiores contratos da história no valor

de $459 milhões de dólares para desenvolver e fornecer motores capazes de gerar 183kN (41,000 lb) de empuxo á taxas de 8:1 BPR.

A partir de então, *Prat & Whitney*, assim como outras companhias como a *Rolls-Royce* desenvolveram também motores competitivos para o mercado, inclusive novos contratos foram disputados para o desenvolvimento de novas aeronaves como foi o caso do Airbus A300. Em 1968, novamente a General Electric ganhou mais uma disputa contratual, fornecendo motores para o DC-10 *trijet* por meio do motor CF6/34.

Em 1971, a GE aliou-se a empresa francesa *Snecma* para desenvolver o CFM56, que tornou-se um grande sucesso de vendas, em especial pela opinião pública acerca da poluição atmosférica que chegou a fazer com que em 1985 muitos aeroportos americanos viessem a banir motores com altos consumos de combustível e emissão de gases poluentes. A família CFM56 é, ainda hoje, muito popular.

Podemos citar ainda alguns motores relevantes que representaram grandes avanços tecnológicos dos *turbofans*. O GE 90-11B era, até então, o maior e mais potente motor do mundo sendo capaz de gerar incríveis 567kN de empuxo (127.900 lbs). Mas em 2017 teve seu posto ganho pelo GE9X, capaz de gerar 134.300 lbs de empuxo. Outras duas menções importantes são os motores GEnx e LEAP.

O GEnx foi o motor mais rápido em número de vendas, tendo atingido mais de 2700 motores em serviço e em ordem. (GE Aviation, 2020). O projeto teve como base a arquitetura do GE90. O motor oferece até 15% de mais eficiência e 15% menos emissões de CO2 quando comparados aos CF6. O GEnx representa um enorme salto em termos de eficiência e adoção de novas tecnologias através da adoção de novos materiais e compósitos estruturais. Em Outubro de 2019, uma aeronave 787-9 equipada com motores GEnx bateu o recorde, até então do GE90, de voo mais longo da história percorrendo 10.200 milhas (GE Aviation, 2019).

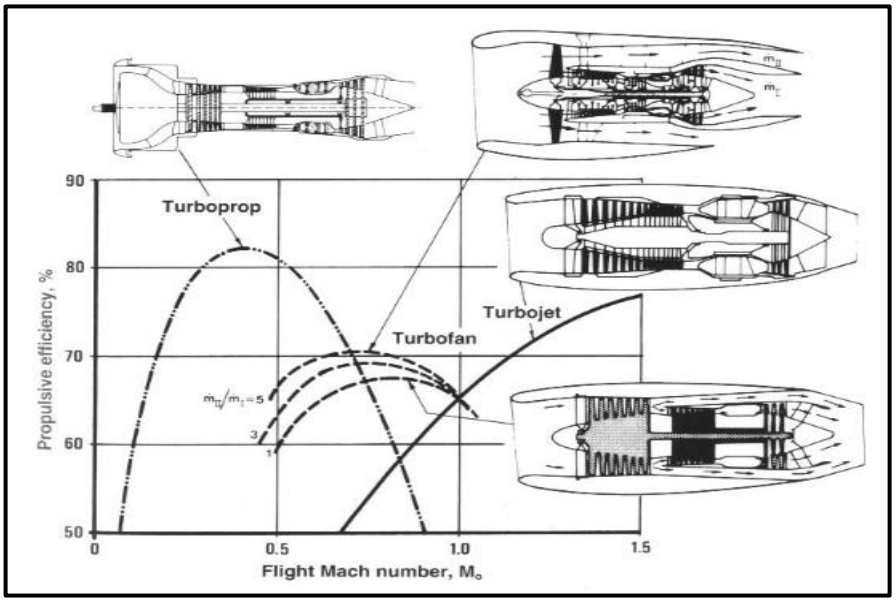
O LEAP é o motor mais novo produzido pela CFM, a joint venture entre GE *Aviation* e *Snecma*. Tem sua arquitetura baseada em muitos aspectos do GEnx, trazendo ainda mais inovações no que diz respeito a utilização de materiais e manufatura aditiva. Alguns componentes, como as *fuel nozzles*, são 25% mais leves e 5 vezes mais resistentes do que peças fabricadas de modo convencional.

2.1.2 Regimes de operação

É importante observar que, pelas diferentes construções, cada tipo de motor se mostra mais ou menos eficiente em determinados regimes de operação. De maneira geral, pode-se

relacionar a eficiência de cada tipo de motor dada a sua média de velocidade de operação. Por exemplo, motores turbojato são desenvolvidos para acelerar baixas massas de ar a grandes velocidades de exaustão, sendo ineficientes quando submetidos a regimes de voo subsônicos, isso é, quando em velocidades de cruzeiro inferiores a velocidade do som (1.235 km/h). Por sua vez, motores *turbopop* são desenvolvidos para acelerar grandes massas de ar a baixas velocidades de exaustão, não conseguindo atingir velocidades próximas de *Mach* 0.8 (1000km/h), em altitude de 40,000 pés. É nessa faixa de trabalho que surgiu a necessidade do desenvolvimento dos *turbofans*. A Figura 6 abaixo compara a eficiência do propulsor (eixo Y) pela velocidade de cruzeiro na escala de *Machs* (eixo X) para 4 diferentes tipos de motores a jato.

Figura 6 - Eficiência de propulsão em diferentes motores de acordo com velocidade de operação.



Fonte: Adaptado de Hünecke (1997, p.9)

2.2 Vibração

> "Vibração é uma resposta repetitiva, periódica ou oscilatória de um sistema mecânico. A taxa dos ciclos de vibração é chamada de "frequência". Movimentações repetitivas que são em algum ponto regulares, e que ocorrem a relativamente baixas frequências, são comumente chamadas de oscilações, enquanto que movimentações repetitivas que ocorram mesmo em altas frequências com baixas amplitudes e comportamento irregular ou randômico são chamados de vibração" (SILVA, 2000, p.13)

Em sistemas mecânicos vibrações ocorrem de forma natural, sendo inerentes à própria condição do movimento, resultantes do intercâmbio entre energias cinéticas e potenciais entre os componentes do sistema. As vibrações naturais estão presentes não apenas em sistemas mecânicos, como em sistemas elétricos e sistemas de fluidos, onde existem conversões entre diferentes tipos de energia e diversas fontes de dissipação de energia. Também podem ser derivadas de interferências de ordem interna, presentes no próprio sistema, ou externas.

Quando a interferência gera frequências que coincidem com a frequência natural de vibração do sistema, a amplitude do movimento aumenta de forma vigorosa. A esse fenômeno dá-se o nome de ressonância, e a sua frequência, o termo "frequência de ressonância". De maneira geral, o fenômeno de ressonância é extremamente indesejável do ponto de vista de engenharia, podendo inclusive ser um fenômeno destrutivo em alguns sistemas.

A vibração está presente em diversos ramos da engenharia como engenharia aeroespacial, civil, mecânica e até elétrica. Modelos e sistemas de controle são imprescindíveis para o controle e estudos das vibrações.

2.2.1 Vibração em sistemas aeronáuticos

A análise de vibrações é de extrema importância em sistemas aeronáuticos, especialmente em motores. Todos os motores apresentam algum grau de vibração produzida, por exemplo, pelo excesso ou falta de torque nos parafusos que fixam componentes rotativos do motor, folgas entre peças, falta de lubrificação entre componentes e rolamentos, desalinhamentos, desequilíbrios, torções, contato entre peças e demais fenômenos de natureza mecânica. De maneira geral, a vibração excessiva em motores está associada ao desgaste prematuro ou excessivo de componentes podendo resultar em perdas de eficiência, trincas e falhas mecânicas.

Para garantir o bom funcionamento do motor, bem como evitar possíveis falhas mecânicas, a vibração é um dos principais parâmetros de controle utilizados para estimar a vida útil de uma peça e estimar quando a mesma deve ser reparada ou ajustada para evitar desgastes. Nesse contexto:

> "A análise de vibrações tornou-se, inequivocamente, o método mais relevante para o controlo de condição dos motores turbo *fan*, bem como para garantir a aeronavegabilidade dos mesmos, permitindo aferir com maior confiança as intervenções de manutenção preventivas." (FERNANDES, 2019, p.15)

Atualmente sensores presentes nos motores e aeronaves são capazes de estimar e capturar com precisão os níveis de vibração de diferentes componentes do motor a fim de estimar possíveis manutenções preventivas e prever possíveis tendências através do recorte histórico das vibrações de operação. Vale ressaltar que para companhias aéreas o processo de manutenção de motores é extremamente crítico e custoso uma vez que durante esse processo o mesmo está impedido de funcionar e, portanto, se traduz em perdas financeiras para o operador. Por isso, para (Nunes, 2005), os sistemas de monitoramento são concebidos com o objetivo de prever os níveis de degradação do motor com a máxima antecedência possível para que seja possível identificar e isolar possíveis danos antes que os mesmos comecem a se manifestar.

Um exemplo de sistema de controle de vibração é o CMCS (*Central Maintenance Computer System*) que equipa os Boeing 747-400, cujo principal objetivo é recolher e analisar toda informação de manutenção. O CMCS, segundo (Aslin, 1990), interpreta e integra dados relativos ao funcionamento de diversos equipamentos e seus correspondentes históricos de manutenção emitindo alertas relativos a possíveis perdas de desempenho em consequência de anomalias que caso tenham suas correções ignoradas podem resultar em eventuais falhas durante a operação.

As amplitudes, portanto, devem permanecer dentro de faixas e regimes específicos para garantir a preservação do sistema e baixos níveis de fadiga e desgaste dos componentes envolvidos. Segundo (GE AVIATION, 2018), esses limites são definidos através de extensivos testes de fadiga dos componentes. No caso de estruturas rotativas, sabemos, por exemplo, que os eixos são postos à prova através de testes onde são forçados a atuar por 10 milhões de ciclos submetidos a um desbalanceamento conhecido. Caso a peça seja capaz de resistir aos níveis de vibração, pode-se dizer que o material é suficientemente forte para operar por toda sua vida. Ainda assim, são estipulados limites da vida útil desses componentes de forma a garantir a não ocorrência de falhas durante a operação.

2.2.2 Desbalanceamento

O balanceamento das turbinas, discos e rotores é de extrema importância para a garantia de performance de um motor. O desbalanceamento é um fenômeno natural dos processos de fabricação de peças rotativas. No caso de *blades*, por exemplo, é comum que cada peça possua pequenas variações de massa que, ao final, resultam em forças resultantes diversas que contribuem para a vibração do conjunto. Segundo (Mason, 1997), o balanceamento de turbinas, por exemplo, é um problema matemático do tipo NP-difícil. Análises matemáticas e

experimentais comprovam que o problema de balanceamento é extremamente complexo. De maneira geral, são utilizados *softwares* como "*Blade Plot*" que calculam o peso de cada *blade* e fazem a distribuição das *blades* de maneira estratégica de modo a reduzir as forças resultantes divergentes, tentando promover o balanceamento ótimo através do remanejamento das *blades* garantindo a maior anulação de forças resultantes individuais. Esse processo, assim como métodos heurísticos, são alternativas interessantes para reduzir o desbalanceamento, no entanto, matematicamente é impossível provar que a solução encontrada é ótima assim como na maioria dos casos é impossível garantir um balanceamento perfeito dos conjuntos rotores.

Quando falamos em motores, existem diferentes sistemas rotativos que serão exemplificados no decorrer desse trabalho. Quando os conjuntos rotativos são conectados os erros de desbalanceamento são propagados no processo de junção dos módulos que compõem o motor, processo esse chamado de *stack-up*, termo em inglês para "empilhamento".

2.3 KDD – descoberta de conhecimento em bases de dados

Segundo Han; Kamber; Pei (2011) enganam-se os que pensam que vivemos hoje na era da informação, vivemos na "Era dos dados". *Terabytes* e *Pentabytes* de dados compõe as redes de computadores e *devices* mundial. O crescimento extraordinário do volume de dados disponíveis na atualidade se traduz na introdução de novas tecnologias e grandes avanços.

Negócios em todo o mundo geram e têm acesso a quantidades gigantescas de dados, referentes a transações, preferências e comportamento dos seus clientes. Estima-se que as redes de comunicações atuais lidem com dezenas de *Pentabytes* de dados diariamente. Em meio a esse contexto de alto volume de dados disponíveis em diversos setores da sociedade, a existência de ferramentas e metodologias para lidar com o alto volume de informações tornou-se imprescindível na identificação de padrões, soluções e relações entres os dados. As técnicas tradicionais de exploração de dados tornaram-se obsoletas e pouco adequadas para tratar a grande maioria dos repositórios, segundo (De Amo, 2004).

Nesse contexto, na década de 80 popularizou-se se o termo "Mineração de Dados", em inglês "*Data Mining*". A expressão faz referência a tradicional atividade de mineração de materiais preciosos, tais como ouro. Os metais preciosos se encontram, normalmente, incrustados e escondidos em meio a rochas e areia. De maneira análoga, a mineração de dados se concentra em extrair dos dados relações e dados "preciosos". Aparentemente o conceito é simples e de fácil compreensão, no entanto, os autores parecem divergir bastante quanto ao que

“mineração de dados” compreende. Para (De Amo, 2004) a “Mineração de Dados” é compreendida como uma área de pesquisa e atuação. Observe sua definição:

> “Mineração de Dados é uma área de pesquisa multidisciplinar, incluindo tecnologia de bancos de dados, inteligência artificial, aprendizado de máquina, redes neurais, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados.”

No entanto, para outros Autores a “mineração de dados” não seria uma área de pesquisa, mas uma etapa importante de um processo conhecido *como Knowledge Discovery in Databases*” ou “Descoberta de Conhecimento em Bases de Dados”. Esse é, por exemplo, o entendimento de Fayyad; Piatetsky; Smyth (1996). Para os autores, a mineração corresponde a um processo em particular da KDD, que irá compreender diferentes etapas, entre elas, a mineração. Para Fayyad; Piatetsky; Smyth (1996), sem as etapas prévias que compreendem o KDD (Preparação de Dados, Seleção de Dados, Limpeza de Dados e Transformação), a mineração de dados se torna um processo perigoso, indicando falsas relações e resultados. Essa visão também é compartilhada por Han; Kamber; Pei (2011). Há ainda autores que irão compreender que KDD e Mineração de Dados são sinônimos, como é o caso de Rezende (2005).

A título de melhor compreensão, tomar-se-ão as visões de Han; Kamber; Pei (2011) e Fayyad; Piatetsky; Smyth (1996) como os fios condutores desse trabalho. Sendo assim, a mineração de dados será considerada um subprocesso do KDD, através de uma visão adaptada desses autores. Segundo Han; Kamber; Pei (2011) no Livro “Data Mining: Concepts and Techniques”, o KDD pode ser dividido nos seguintes processos (Figura 7):

1- Limpeza de Dados - Nessa fase são eliminadas inconsistências e redundâncias de maneira a garantir uma fonte clara e consistente de dados.

2 - Integração de Dados - Em análises complexas, é muito comum que os dados interpretados sejam provenientes de diferentes sistemas e fontes de dados, por isso a integração deles é fundamental. É nessa etapa que as diferentes fontes de dados são combinadas em uma única base de dados.

3- Seleção - Nesse momento cabe ao analista selecionar quais atributos serão utilizados para o desenvolvimento de análises. Em muitos casos, grande parte dos dados disponíveis não são relevantes para as análises desejadas.

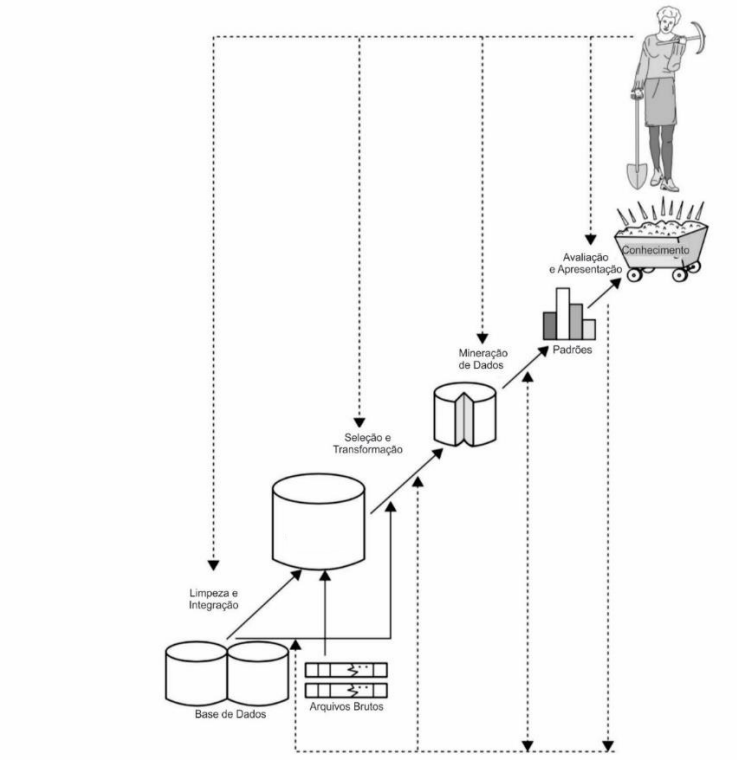
4 - Transformação dos dados - Nesse momento, podem ocorrer transformações tais como: normalização, agregação e criação de novos atributos. Adicionalmente, os dados são formatados para a aplicação de algoritmos de mineração.

5 - Mineração - Nessa etapa são utilizadas ferramentas e técnicas de mineração de dados para encontrar padrões de interesse. A vasta maioria das técnicas utilizadas engloba a aplicação de algoritmos de aprendizado de máquina.

6 - Avaliação ou Pós-processamento: Nesse momento, são identificados os padrões extraídos e os resultados encontrados.

7 - Visualização dos Resultados (Conhecimento) - Essa etapa consiste em utilizar ferramentas para representar as relações encontradas de forma visual.

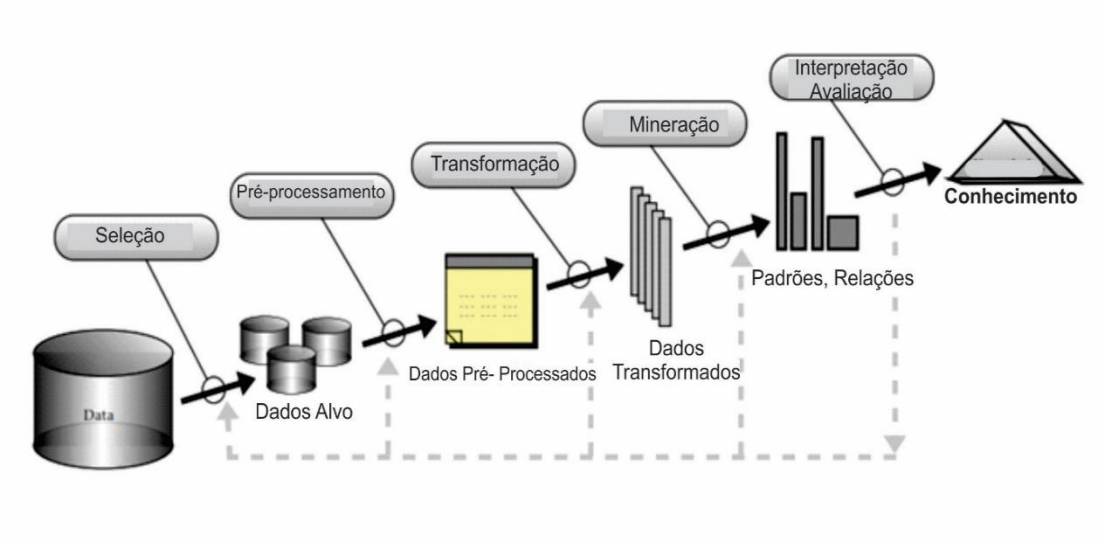
Figura 7 - KDD (processos).



Fonte: Adaptado de Han; Kamber; Pei (2011, p.7)

Fayyad; Piatetsky; Smyth (1996) propõem uma abordagem metodológica um pouco mais prática, divergindo um pouco da proposta por Han; Kamber; Pei (2011). Observe a proposta de Fayyad; Piatetsky; Smyth (1996) ilustrada pela Figura 8.

Figura 8 - KDD (esquema).



Fonte: Adaptado de Fayyad; Piatetsky; Smyth (1996, p.84)

As etapas, de maneira resumida são:

1 - Definir “dados alvo”, isso é, definir as amostras, conjuntos ou subconjuntos de dados que serão estudados. Tendo em vista o conjunto universo de dados disponíveis e as intenções da análise.

2 - Pré-processamento, durante essa etapa os dados são submetidos a uma série de algoritmos de pré-processamento para que, posteriormente, os resultados de mineração sejam mais precisos. Como etapas de pré-processamento, são algumas: limpeza (eliminação de ruídos e inconsistências da amostra), seleção de atributos e balanceamento.

3 – Transformação, durante essa etapa o conjunto de dados iniciais pode ser transformado, seja através de processo de redução, seja através de processos de agrupamento, em outro conjunto.

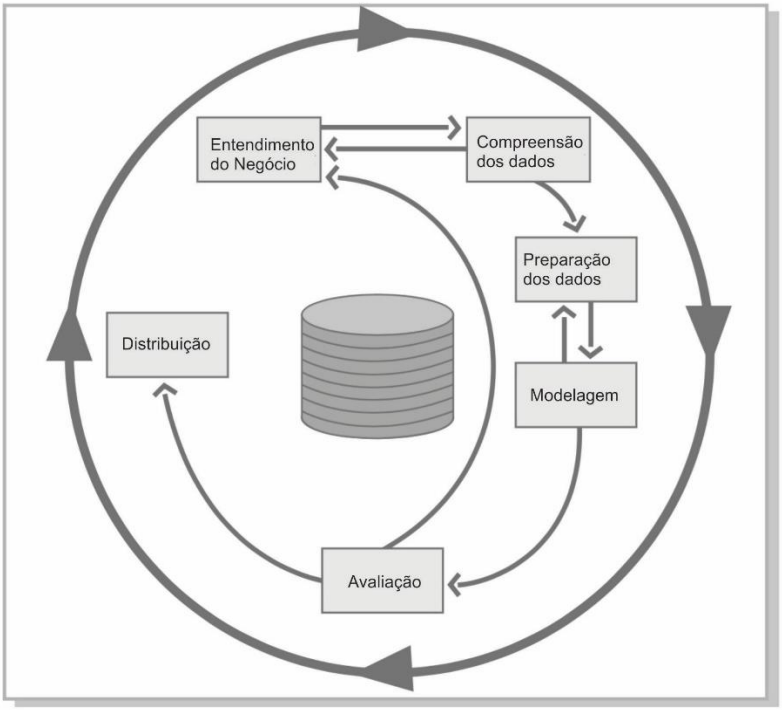
4 – Mineração: nesse momento serão aplicados os algoritmos, majoritariamente baseados em aprendizado de máquina, escolhidos durante o passo anterior para encontrar padrões, relações e representações.

5 – Interpretação: essa etapa consiste em interpretar os resultados da mineração sendo possível recorrer a técnicas de visualização para melhor compreensão e entendimento dos resultados. Também é encorajado que, caso necessário, o analista possa revisitar qualquer dos passos anteriores a fim de testar novas possibilidades e ganhos incrementais de acuracidade.

Como processo final, recomenda-se que com as descobertas sejam traçados planos de ação para impactar o cenário estudado ou, caso não seja necessária ou possível a intervenção no problema estudado, recomenda-se que o processo e descobertos sejam registrados.

Conforme dito anteriormente, ambas as abordagens de Fayyad; Piatetsky; Smyth (1996) e Han; Kamber; Pei (2011) são muito similares. No entanto, como alternativa à metodologia KDD existem outros protocolos para a análise de dados. (Chapman, P. et al, 2000), por exemplo, propõem o método CRISP-DM (*Cross-Industry Standard Process of Data Mining*). Os processos da abordagem CRISP-DM podem ser divididos em seis fases, não necessariamente cronológicas, isto é, o fluxo de processos não é unidirecional sendo bastante comum a ida e volta entre os mesmos como é possível observar através da Figura 9 no formato de um círculo.

Figura 9 - Esquema CRISP-DM.



Fonte: Adaptado de CHAPMAN, P. et al (2000, p.10)

As etapas do processo CRISP-DM podem ser divididas em:

1. Entendimento do Negócio: Nessa etapa, é importante definir com clareza quais os objetivos da análise, o que se pretende atingir com o processo de mineração. Uma vez bem definidos os objetivos, é necessário compreender as restrições da questão e formular um problema de mineração de dados, assim como uma estratégia preliminar de abordagem.

2. Compreensão dos Dados: Devido à natureza multifatorial dos problemas que envolvem dados, é comum que os dados tenham sua origem em diferentes sistemas e bancos de

dados. Uma vez coletados, recomenda-se uma breve análise exploratória para avaliar a qualidade dos dados e interpretar possíveis padrões através de explorações visuais.

3. Preparação dos Dados: Nessa etapa, os dados são formatados e limpos. É comum que os dados estudados sejam provenientes de diferentes fontes e possuam diferentes formatações. Essa etapa costuma demandar bastante tempo dos pesquisadores que devem garantir que todas as fontes sejam consistentes.

4. Modelagem: Nessa fase os dados são submetidos a diferentes técnicas e ferramentas de mineração de acordo com os objetivos desejados. É comum que diferentes técnicas sejam aplicadas a fim de compreender quais melhor se adaptam a questão estudada. A calibração dos modelos também costumar ocorrer nessa etapa, onde certos parâmetros podem ser alterados para garantir melhores resultados e adequação ao problema analisado.

5. Avaliação: Essa etapa consiste na avaliação dos resultados encontrados. É muito importante que os resultados sejam criticados e interpretados por ambos especialistas de dados e conhecedores do negócio. Para (Witten; Frank, 2016), os resultados devem passar por testes e validações visando garantir e questionar a confiabilidade dos modelos. Alguns métodos de validação sugeridos são: *Cross Validation*, *Suplied Test Set*, *Use Training Set* e *Percentage Split*. Também são importantes indicadores para suportar a análise de resultados, como: matriz de confusão, índice de correção e incorreção de instâncias mineradas, estatística *kappa*, erro médio absoluto, erro relativo médio, precisão e *F-measure*.

6. Distribuição: Nesse momento, os modelos comprovados e suficientemente confiáveis são aplicados a todos os dados disponíveis e os resultados são compartilhados com as partes interessadas.

A metodologia CRISP-DM é também muito parecida com a abordagem de KDD, tendo ambas fundamentalmente etapas muito similares se não idênticas. Os três modelos apresentados, são, portanto, formas de abordar problemas de análise de dados. Esse trabalho apresentará, ao seu final, um case prático seguindo a metodologia de abordagem de Fayyad; Piatetsky; Smyth (1996). Também é importante ressaltar que convencionou-se, durante a elaboração desse trabalho, a utilização do termo “atributo”, termo comum no campo das ciências de dados, para representar variáveis, termo comum no meio estatístico.

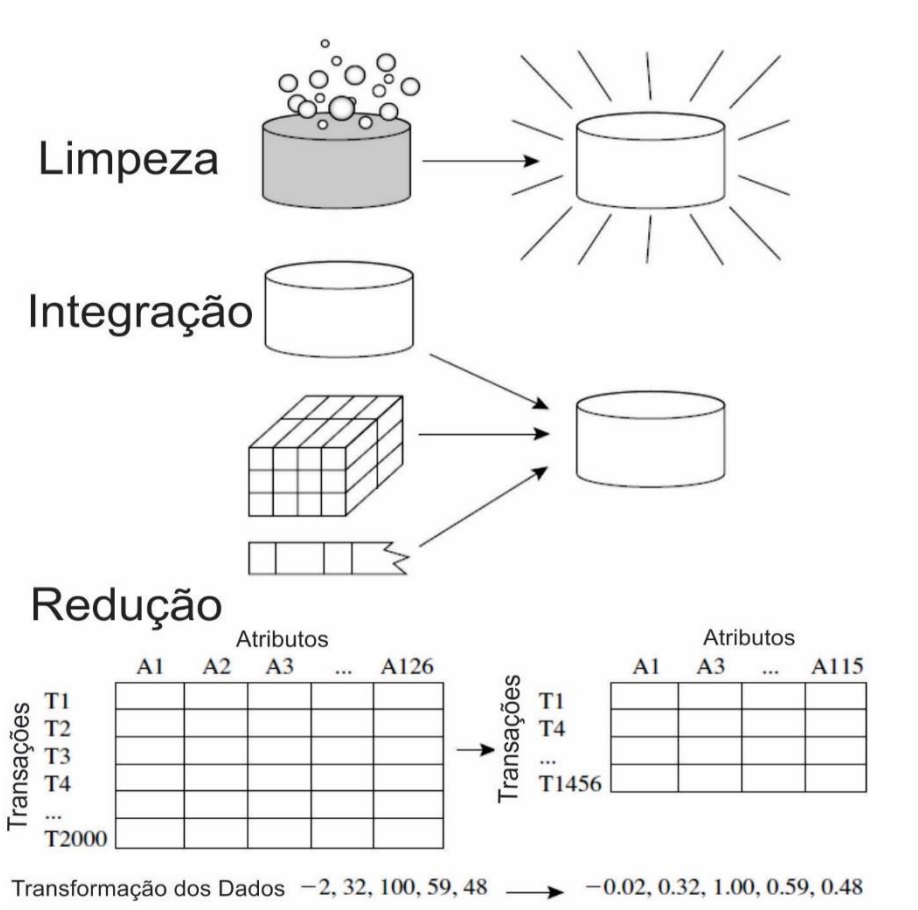
2.3.1 Tratamento dos dados

O tratamento dos dados, isso é, a fase de pré-processamento, é comum a todas as metodologias de análise de dados abordadas. Podemos dividir dados em dois tipos, são eles: os

dados numéricos e categóricos. Enquanto o primeiro tipo diz respeito a dados em valores numéricos, o segundo tipo é de natureza nominal. De forma geral, muitos algoritmos só trabalham com um tipo de dados sendo, por isso, vital a compreensão dos dados estudados e dos métodos aplicados na análise.

Uma vez que a natureza dos dados estudados é definida e compreendida, o próximo passo é dar início, como descrevem os autores Han; Kamber; Pei (2011), à preparação e limpeza dos dados de acordo com as seguintes etapas metodológicas. (Figura 10)

Figura 10 - Etapas de Pré-Processamento.



Fonte: Adaptado de Han; Kamber; Pei (2011, p.87)

1 - Limpeza: Nessa etapa são eliminados valores inconsistentes que poderiam comprometer o resultado da análise. Também são removidos dados duplicados, entradas sem valor e são padronizadas as unidades utilizadas em dados numéricos.

2 - Integração: Como dito anteriormente, é comum que os dados provenham de diferentes fontes. Nessa etapa os dados são integrados em uma única data-base.

3 - Transformação: Certos algoritmos só funcionam com valores numéricos, enquanto outros só são aplicáveis a valores categóricos e, por isso, em muitos casos a transformação de dados é necessária. Existem diversas técnicas para converter dados numéricos em categóricos e vice-versa. Algumas técnicas empregadas são: Agrupamento (são definidas faixas de valores em que os dados são agrupados), Normalização (quando os valores são adequados a diferentes escalas) e a criação de novos atributos. Uma vez formatados, os dados devem ser submetidos a análises estatísticas iniciais e exploração. São ferramentas e técnicas comumente utilizadas na exploração inicial: análise de dispersão (quartil, percentil e variância), média, mediana, moda, histogramas, gráficos de barra, frequência, dispersão e *BoxPlot*.

4 - Redução: Essa etapa costuma ser aplicada apenas em casos onde o volume de dados é tão grande que o processo de mineração se torna impraticável. Nesses casos os dados analisados são reduzidos de forma a preservar a representatividade da amostra, o que permite que algoritmos de mineração sejam aplicados com maior eficiência e velocidade sem renunciar à qualidade e acuracidade dos resultados.

2.3.1.1 *Limpeza de dados, pré-processamento*

Como dito anteriormente, ao lidar com problemas reais é muito comum que os dados precisem ser “limpos” para garantir a consistência e acurácia da análise. Serão abordados os principais problemas enfrentados durante essa tarefa e boas práticas para lidar e contornar essas dificuldades.

1 - Valores Faltantes: A falta de alguns parâmetros e variáveis para entradas individuais é comum. Para contornar essa questão soluções recorrentes são:

A - Ignorar o conjunto/vetor atributos, ou seja, excluir da análise a tupla em questão. Esse é um método pouco eficiente uma vez que excluímos da amostra dados que possivelmente seriam relevantes, ainda mais em um cenário de escassez de dados.

B - Preenchimento manual, isso é, preencher manualmente as informações faltantes com base em uma análise individual dos casos e tentativa de rastreamento dos valores que estão faltando. É uma alternativa que demanda muito tempo e, portanto, impraticável na maior parte dos problemas que envolvam grandes quantidades de dados.

C - Preenchimento com constantes, essa alternativa propõe que os dados nulos, faltantes, sejam substituídos por constantes como “Desconhecido”, nesse caso todos os dados omissos terão em comum essa variável fictícia “Desconhecido”. O objetivo é que também sejam

relacionadas durante análise a relação das variáveis desconhecidas, assumindo-se que todas as variáveis faltantes pertencem a uma mesma classe, ou grupo.

D - Preenchimento com medidas de tendências, isso é, compor os dados omissos com valores como média ou mediana. Para dados que se relacionam de ordem simétrica seria indicado a utilização da média como boa prática. Caso contrário, uma boa opção seria a utilização da mediana.

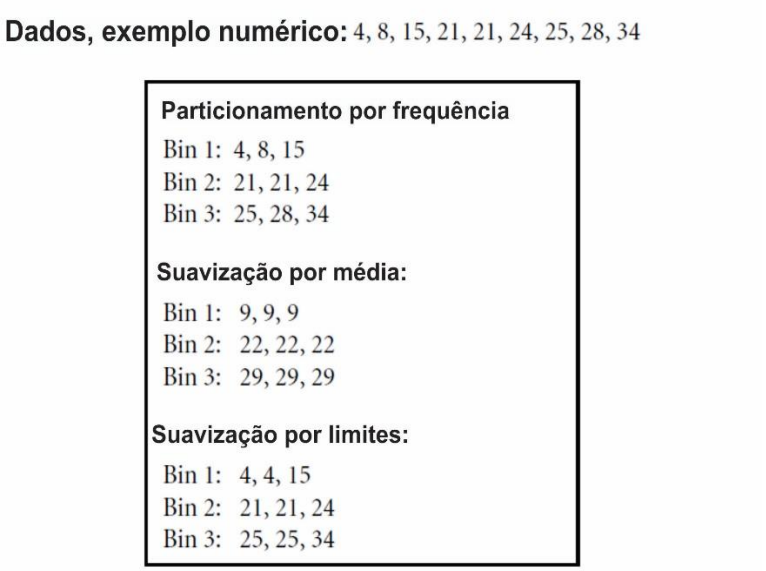
E - Utilizar a média ou mediana de determinada classe para tupla correspondente, essa é uma proposta muito interessante. A ideia é encontrar a média ou mediana para tuplas de uma mesma categoria para cada atributo, dessa forma utilizar-se-íam as tendências para os parâmetros também relacionando a categoria ou classe daqueles dados.

F - Estimar o valor mais provável, essa é a alternativa que melhor preserva as características dos dados, evitando possíveis vieses. A iniciativa é determinar por meio de ferramentas como: regressão, inferência bayesiana ou árvores de decisão quais os valores mais prováveis para aquela dupla. Muitos algoritmos de classificação fazem esse tratamento automaticamente.

2 - Ruído: É comum que em certas amostras existam ruídos, isso é, a presença de dados pouco relevantes ao problema e que, quando não tratados, podem levar a conclusões e análises incorretas. Três métodos se destacam para mitigar ruídos presentes em dados, são eles:

A - *Binning*, ou, intervalo de classes: nesse método os dados são separados por conjuntos chamados de *bins*, que se estabelecem pela proximidade entre si, ou seja, pela vizinhança. Por isso, esse método promove uma suavização local. O método pode utilizar-se, mais uma vez, da média ou mediana, de acordo com o caso analisado. Para ilustrar o método vejamos a Figura 11.

Figura 11 - Método de Suavização por Intervalo de Classes ou *Binning.*



Fonte: Adaptado de Han; Kamber; Pei (2011, p.90)

Nesse exemplo, observam-se dados (fictícios) numéricos. Os dados são, de início, particionados localmente em *bins*. No exemplo citado, cada *bin* possui três valores. Na suavização com base na média da vizinhança a proposta é substituir cada valor de um bin pela média dos valores do conjunto daquele bin, isso é, pela média de sua vizinhança. Já na suavização com base em mediana, são utilizadas como valor para a substituição a mediana da vizinhança. Há também a possibilidade de fazer uma suavização com base em limites, isso é, são levados em contas os valores de mínimo e máximo em cada vizinhança. São calculadas a distância de cada valor, dentro de um bin, entre o mínimo e máximo de sua vizinhança, limite superior e inferior. Os dados são, então, substituídos pelo valor do limite mais próximo, seja o mínimo ou máximo da sua vizinhança.

B - Regressão: Técnicas estatísticas como regressão linear e multidimensional também podem ser utilizadas para minimizar ruídos deixando as relações mais claras. No caso da regressão linear, o objetivo é, com base em um parâmetro, estipular a linha de regressão que melhor se encaixa na distribuição de dados em relação a outras variáveis (independentes). Nesse caso a linha de tendência é utilizada, substituindo os dados reais encontrados pelos dados que se encaixam na linha de tendência. Também é possível a utilização da regressão multidimensional. Nesse caso, de maneira parecida, são traçados planos de referência, ao invés de linhas de tendência, como ocorre na análise linear.

C - Análise de *Outliers*: A remoção de *outliers*, seja por meio de partições estatísticas da amostra dos parâmetros em percentis seja pela utilização de *"clusters"* espaciais ou geométricos que se separem em vizinhanças bem definidas, os pontos que fugirem da vizinhança são, nesse caso, considerados outliers e podem removidos da amostra mediante interpretação dos dados.

Vale ressaltar que os problemas citados são os mais comuns. No entanto, existem sob o arsenal de um cientista de dados uma gama ainda maior de ferramentas para analisar e remover ruídos. E, por isso, é vital que o analista consiga compreender bem a natureza dos dados estudados para entender quais os tratamentos necessários e quais inconsistências nos dados podem ser mais ou menos importantes no que diz respeito a influência dos resultados e relações encontradas. Em casos reais, como o abordado ao fim desse trabalho, são aplicadas também certas regras que se estabelecem de maneira empírica. Imagine que se esteja analisandas idades de pessoas em determinada região a fim de estabelecer predições eleitorais. Sabe-se que, de maneira empírica, a idade assume uma distribuição próxima à normal e que idades superiores a, por exemplo, 100 anos são raras. Pode-se estabelecer, portanto, regras que excluam conjuntos de dados que tenham idades superiores a 100 anos, uma vez que esses valores estão a muitos desvios padrão da média. Pode-se também excluir, de acordo com a precisão necessária e interpretando o problema, todos aqueles que possuem mais de 65 anos e não são obrigados a votar.

2.3.1.2 Integração

A integração de dados é um problema muito comum uma vez que na maior parte dos casos lida-se com dados de diferentes fontes. A integração tem por objetivo combinar diferentes dados em apenas um conjunto. É comum que o mesmo parâmetro possua diferentes nomes para diferentes fontes e, por isso é imprescindível que sejam compreendidos os parâmetros e, sobretudo, o que se chama de metadados. Metadados correspondem a todos os saberes que extrapolam os dados em si, ou seja, dizem respeito ao conhecimento e compreensão do que aquele parâmetro representa, seja seu nome, significado, tipo, faixa e regras sobre as quais aquele parâmetro está submetido. Uma boa compreensão dos metadados é fundamental em problemas de integração onde antes de combinados os parâmetros de diferentes fontes devem ser confrontados e comparados. Outro ponto no qual deve-se prestar bastante atenção é sobre a estrutura dos dados observados, isso é, como eles são gerados. Certos parâmetros podem surgir de diferentes fórmulas ou algoritmos.

Imagine um supermercado que possui uma promoção de frutas e o sistema do caixa pode trabalhar de diferentes formas para aplicar o desconto. Em um cenário hipotético 1, poder-se-ia dizer que uma vez identificada a classe do produto, caso o mesmo seja uma fruta, será aplicado um desconto de, digamos, 15% para aquele item. Ao fim, todos os itens seriam somados e teriam sido registrados os seguintes parâmetros: “valor_frutas” (soma do valor total de frutas individuais já com o desconto de 15%), “valor_nao_frutas” (soma do valor total dos itens que não são frutas) e “valor_total” (soma dos parâmetros “valor_frutas” e “valor_não_frutas”).

Em um cenário hipotético 2, poder-se-ia dizer que o sistema de caixa atuaria de forma diferente. Ao invés de aplicar o desconto às frutas individualmente, o sistema aplicará 15% ao valor da soma de todas as frutas compradas. Nesse caso, haveria, por fim, os seguintes parâmetros: “valor_frutas” (soma do total de frutas, sem desconto), o parâmetro “valor_desconto_frutas” (soma do total de descontos, 15% do valor das frutas), “valor_nao_frutas” (soma do valor total dos itens que não são frutas) e o parâmetro “valor_total” (soma dos parâmetros “valor_frutas” e “valor_não_frutas”) subtraídos do parâmetro “valor_desconto_frutas”.

Observe que, a título de utilização nos dois cenários, o sistema atende às necessidades do operador apresentando o valor total das compras realizadas. No entanto, os mesmos parâmetros “valor_fruta” e “valor_total” apresentam diferentes naturezas, isso é, são gerados por diferentes regras. Nesse caso, não seria possível, de imediato, realizar a integração dos valores sem antes adequar os dados.

Deve-se também prestar bastante atenção na questão da redundância. Certos atributos são, muitas vezes, composição de outros atributos, ou têm significado idêntico ou próximo, embora tenha sido gerado por outros caminhos. Ou seja, são derivados de outros e, portanto, dependentes. É muito comum que a eliminação de tais parâmetros melhore o desempenho dos algoritmos que serão aplicados posteriormente, tanto no que se refere a tempo de execução quando a poder preditivo.

No exemplo do supermercado citado, pode-se observar (caso hipotético 2) que o parâmero “valor_total” é uma composição matemática de outros parâmetros (“valor_total” = “valor_frutas” + “valor_nao_frutas” – “valor_desconto_frutas”). Ou seja, nesse caso, para reduzir o tamanho dos dados e remover redundâncias, pode-se remover o parâmetro “valor_total”, exceto, é claro, que relações com essa composição em específico sejam importantes. Para identificar redundâncias entre parâmetros é interessante a utilização de testes de correlação.

2.3.1.3 Transformação

O processo de transformação é essencial em casos em que é preciso converter dados numéricos em categóricos ou o contrário. Também pode ser utilizado para a redução de ruído, garantindo maior consistência dos dados. Os tipos mais comuns de transformação são:

1 – Suavização: é um método para reduzir ruídos em uma amostra. Podem ser utilizadas técnicas como binning, regressão e agrupamento.

2 - Construção de Atributo: nesse caso, são utilizados dois ou mais atributos para a construção de um novo.

3 – Agregação: através dese processo os parâmetros de determinados vetores são respectivamente agrupados com seus pares. Podem ser utilizados diferentes critérios para o agrupamento. Por exemplo, imagine que se tenha acesso a dados referentes aos gastos mensais de determinados artistas. Pode-se agrupar os gastos mensais em anuais para confrontá-los nesse formato e comparar, por exemplo, os gastos anuais de cantores de country com gastos anuais de cantores de rock. Esse processo é também comum à técnica chamada de cubo de dados.

4 – Normalização: a normalização é um processo matemático bem conhecido. Assim como feito nas aplicações cotidianas da matemática, a normalização tem por objetivo alterar as escalas de medições, trazendo os valores da amostra para uma nova escala, mantendo, é claro, suas proporções. É muito comum que dados sejam normalizados para ocupar uma escala de -1.0 até +1.0 ou no intervalo [0, 1].

5 – Discretização: essa técnica é muito utilizada na transformação de valores numéricos para categóricos. O objetivo é realocar os dados em diferentes faixas ou categorias. Por exemplo, ao analisar a idade de certa população, poder-se-ia definir as idades em categorias como “Criança”, “Jovem”, “Adulto” e “Idoso”. Ou ainda, definir faixas de idades como “0-12” , “13-18”, “19-65” , “65,+∞”. Pode-se, ainda, adicionar outras camadas dentro das categorias criadas, de maneira a estabelecer relações de pertinência hierárquica. A técnica de árvores de decisão também pode ser utilizada para auxiliar nesse processo.

6 - Generalização Hierárquica: utiliza-se quando é possível identificar a natureza de dados locais de maneira a compor dados globais, isso é, estabelecer relações de pertencimento e hierarquia. Imagine, por exemplo, que se tenha acesso ao banco de dados de espécimes de animais em risco avistados na Amazônia. Poder-se-ia compor com os vetores individuais para cada animal avistado um vetor cujo atributo espécie seja substituído pelo atributo família, uma

vez que existe uma hierarquia predefinida e de fácil compreensão que diz respeito à classificação dos seres vivos (reino > filo > classe > ordem > família > gênero > espécie).

2.3.1.4 Redução

Como abordado anteriormente, pode haver grande benefício na redução de dados uma vez que lidar com grandes quantidades de dados requer muito tempo e capacidade de processamento. No entanto, é importante que a redução seja feita de maneira a preservar as características originais da amostra.

Redução dimensional (*dimensionality reduction*): Procedimento para promover a redução do número de variáveis, ou atributos, a serem analisadas, o que pode ser feito via diferentes métodos. São as principais técnicas para a redução dimensional:

1 - *Wavelet Transforms,* ou DWT (*Discrete Wavelet Transform*),: é uma técnica de processamento que tem amplas aplicações na Física, Matemática, Ciências Naturais e, é claro, na análise de dados. A técnica, quando aplicada a um vetor X, gera um novo vetor X', composto por coeficientes de *Wavelet*. Embora possuam os mesmos tamanhos, o vetor X' pode ser truncado, comprimido, mantendo-se apenas os coeficientes de *Walvet* suficientemente relevantes. Por isso, a *Walvet transform* também se mostra eficiente como método de limpeza de dados, eliminando aqueles parâmetros cujas relações com dada variável mostram-se fracas. Existem diferentes tipos de DWT que se diferenciam pelo número de *vanishing moments* e interações matemáticas de acordo com o algoritmo. São populares as seguintes aplicações: Haar-2, Daubechies-4 e Daubechies-6.

2 - Análise dos componentes principais, ou *Principal Components Analysis:* segundo (Shlens, 2005), a PCA é uma importante ferramenta para a análise de dados moderna, tendo sua aplicabilidade recorrente em problemas diversos, da neurociência à computação gráfica, em especial por ser um método simples e não paramétrico para revelar importantes informações de conjuntos de dados com alta dimensionalidade. A PCA está intimamente relacionada à técnica matemática "*Singular Value Decomposition*" (SVD). Tendo por *n* o conjunto de atributos pertencentes a uma tupla, a PCA procurará por k *n*-dimensionais vetores ortogonais que possam representar a amostra, onde $k \leq n$. Portanto, a amostra inicial de dados pode ser projetada em um espaço menor, de acordo com o valor de *n* definido, resultando em uma redução dimensional. A técnica permite a seleção dos atributos mais relevantes, criando, assim, um novo conjunto (menor) para futuras análises.

3 - Seleção de Atributos, ou *Atribute Subset Selection*: nem sempre todos os atributos existentes em uma amostra são relevantes para a análise que será feita. Na maioria dos casos, muitos atributos podem ser deixados de lado da análise por não possuírem correlação suficientemente forte com a classe estudada. Esse tipo de decisão pode ser feita por profissionais e experts que, em um dado problema, possuam conhecimento tácito e empírico de quais variáveis podem ou não ser excluídas do problema. No entanto, nem sempre esse tipo de opinião está à disposição e, é nesse momento, que a seleção automática se faz indicada. A eliminação de atributos irrelevantes ao problema se traduz em menores tempos de processamento e achados mais precisos.

Para Han; Kamber; Pei (2011), o objetivo desse tipo de abordagem pode ser resumido em achar o mínimo de atributos necessários que garantam uma distribuição probabilística de resultados o mais próximo possível da encontrada nos dados originais. A ideia por trás dessa abordagem é simples, porém, na prática, a tarefa é extremamente complexa devido a quantidade de possibilidades presentes, isso é, ao alto número de possíveis subconjuntos. Por exemplo, uma amostra de dados que possua n atributos, geraria 2^n-1 subconjuntos diferentes. Testar a eficiência de cada uma dessas combinações seria impraticável, na maioria dos casos. Para lidar com essa complexidade, recorre-se, então, às heurísticas. As heurísticas irão encontrar soluções boas, soluções essas que são ótimas localmente, mas não necessariamente globalmente. Vale ressaltar que essas heurísticas podem ser aplicadas utilizando-se de diferentes abordagens em um conjunto de dados.

Um algoritmo de seleção muito popular é o *Information-Gain Attribute Ranking*, ou "Ganho de Informação". Esse algoritmo é capaz de ranquear os atributos mais importantes por meio do ganho de informação gerado por cada atributo de maneira individual, baseando-se no cálculo da entropia. Entropia diz respeito ao grau de homogeneidade dos valores do atributo em relação à sua classe, ou seja, um alto grau de entropia representa um alto grau de aleatoriedade. O InfoGain, portanto, irá selecionar os atributos que melhor contribuem para a redução de entropia de forma individual.

Outro algoritmo popular é o Cfs, acrônimo para *Correlation-based feature selection*. Neste caso, a análise não é feita para cada atributo individualmente. O Cfs utiliza uma heurística para definir quais atributos, em conjunto, apresentam as melhores correlações com as classes. Por isso, nesse caso, não existe ranqueamento, apenas um subconjunto de atributos. O algoritmo irá primeiro calcular a matriz de correlação entre atributo e classe bem como classe e classe e, então, através de um sistema de pontuação e uma heurística de *best-first-search* selecionar os

melhores subconjuntos de parâmetros, tendo como critério de parada padrão 5 subconjuntos consecutivos que não melhorem a pontuação, ou mérito, da seleção.

4 - Redução Numérica, ou, *Numerosity Reduction* - Técnica que substitui o os dados originais por formatos de dados menores. Essa técnica pode ser paramétrica ou não-paramétrica.

5 - Compressão de dados, ou *Data Compression* - Algoritmos de compressão são aplicados para obter uma representação comprimida ou reduzida da amostra original. Caso os dados originais possam ser reconstruídos sem perdas de informação a partir da amostra comprimida, diz-se que o processo de redução é "sem perdas" ou "*lossless*". Caso contrário, ou seja, caso exista perda de dados a partir da reconstrução, diz-se que o processo é "*lossy*". Existem muitos algoritmos "sem perdas" para compressão, no entanto, de maneira geral, eles permitem uma limitada manipulação dos dados enquanto nesse formato.

Existem, ainda outros métodos para garantir a redução, cada qual possuindo mais ou menos eficiência de acordo com a natureza do problema. Por isso, recomenda-se ao analista que um extenso trabalho exploratório seja conduzido.

2.4 Mineração de dados apoiada por algoritmos de aprendizado de máquina

2.4.1 Problemas típicos

É importante compreender em quais cenários a utilização das ferramentas de mineração de dados podem ser indicadas. Serão abordadas classes de problemas que podem ser tipicamente compreendidos pela utilização de técnicas de mineração. Para Fayyad; Piatetsky; Smyth (1996), em um primeiro momento existem fundamentalmente dois tipos de abordagens possíveis de mineração, as de ordem preditiva (aprendizado supervisionado) e descritiva (aprendizado não supervisionado). Enquanto a primeira diz respeito à utilização de variáveis para prever resultados futuros, a segunda foca em encontrar relações e descrever padrões nos dados analisados. Todas as abordagens diferentes dessas podem ser caracterizadas como derivadas da predição e descrição. Dizem respeito a predição as seguintes aplicações: Classificação e Regressão. Dizem respeito a descrição: Agrupamento e Sumarização. Sobre as aplicações:

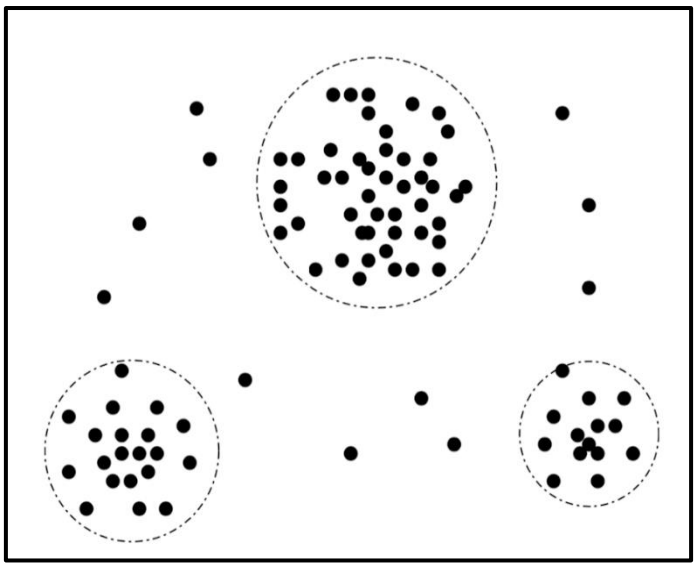
1 – Classificação, quando o objetivo é determinar a classe (categórica) de um objeto com base nos seus respectivos atributos. Convém utilizar-se de dados passados para que o algoritmo seja capaz de aprender as relações entre classe e conjunto de atributos. É uma aplicação muito valiosa, em especial quando ainda não são bem conhecidas as influências dos

atributos na definição de classe. São algoritmos de classificação: *Neural Network, Support Vector Machine,* Árvore de e classificação Bayesiana.

2 – Regressão, esse tipo de aplicação utiliza-se de modelos matemáticos para estimar dados futuros. Por exemplo, tendo acesso aos dados de um hospital veterinário seria possível inferir através de variáveis passadas as relações entre raça, idade, gênero e peso dos animais. Dessa forma, através dos dados passados e suas relações é possível estimar qual o peso de determinado cão levando em consideração as variáveis (independentes) raça, idade e gênero. É muito comum a utilização de modelos de regressão nesse tipo de problema assim como algoritmos do tipo: *Support Vector Machine* e Árvore de decisão.

3 - Agrupamento - Quando utilizado para agrupar dados semelhantes. Um agrupamento (ou cluster) é um subconjunto de dados similares entre si e diferentes dos demais agrupamentos. A semelhança entre os dados é definida por meio da função distância entre os objetos em um plano podendo ser utilizadas concepções de distâncias tradicionais como a Euclidiana. A "qualidade" de um cluster pode ser representada por seu diâmetro, a maior distância entre quaisquer dois objetos no cluster ou pela média de distância entre cada objeto do cluster e o centroide do cluster. Diferente da classificação esse processo não requer que os dados existentes estejam previamente classificados ou agrupados, trata-se de um aprendizado não supervisionado. Algoritmos do tipo kNN, isso é, *k-Nearest Neighbors,* são bastante utilizados em problemas de agrupamento. Veja exemplo simbólico na Figura 12.

Figura 12 - Exemplo visual de agrupamento de dados em *clusters*.



Fonte: Adaptado de Han; Kamber; Pei (2011, p.91)

4- Associação: Quando utilizada para identificar a relação entre dois ou mais atributos. Por exemplo a relação entre o atributo X e Y através da forma: “SE atributo X ENTÃO atributo Y”. É uma análise muito comum nos estudos de comportamento de consumidores para identificar quais produtos são comprados juntos, por exemplo.

2.4.2 Algoritmos de classificação

Tradicionalmente os métodos de mineração podem ser divididos em dois segmentos, o de aprendizado supervisionado (natureza preditiva) e não-supervisionado (natureza descritiva). Como este trabalho emprega apenas algoritmos de aprendizado supervisionado, serão descritos a seguir, e ao longo do restante do texto, algoritmos relacionados apenas a esse segmento.

O aprendizado supervisionado depende das classificações prévias das instâncias que compõem o TS. O objetivo é encontrar relações entre os atributos e a classe de forma que instâncias futuras possam ter seus atributos coletados e as respectivas classes sejam preditas pelo modelo resultante do treinamento. A seguir, são apresentados os algoritmos de classificação mais utilizados:

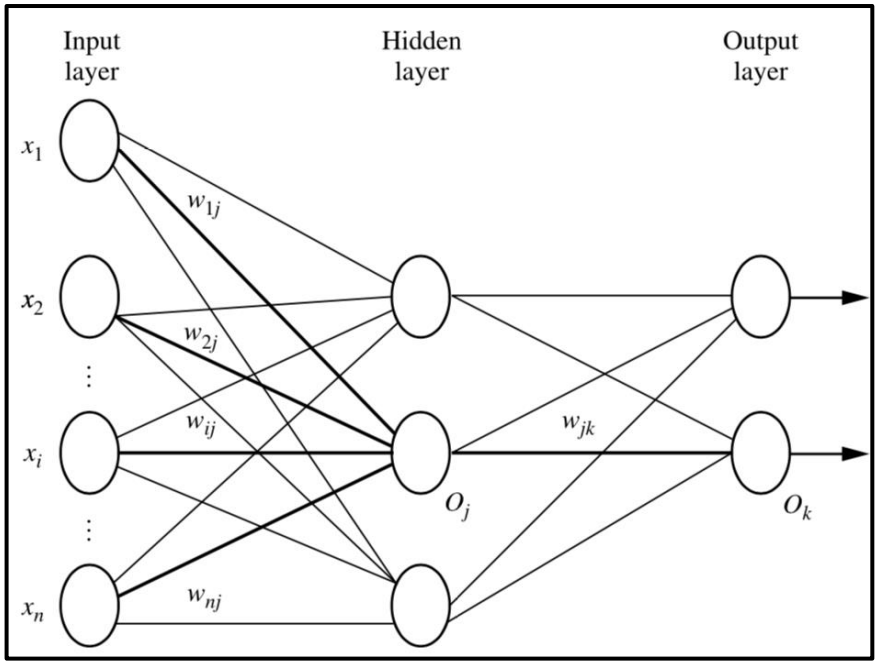
Árvores de Decisão - É um dos métodos mais tradicionais para classificação de dados. O método funciona como um fluxograma, que se assemelha ao formato de uma árvore, por isso o homônimo. Cada nó simbólico indica um teste feito sobre um atributo. As ligações entre os nós representam todos os valores possíveis do teste do nó superior, enquanto as folhas indicam a categoria a qual o registro pertence. Para classificar um novo registro basta seguir o fluxo da árvore, do nó raiz até as folhas. As árvores de classificação podem ser também traduzidas em regras de classificação. Han; Kamber; Pei (2011) fazem uma revisão de literatura abordando as origens e características dessa técnica e suas aplicações em algoritmos. Um dos primeiros algoritmos a utilizar o conceito de árvores de decisão foi o ID3 (*Iterative Dichotomiser),* que precedeu o famoso C4.5, algoritmo que é até hoje utilizado como comparativo e referência no meio. Outro algoritmo muito popular que foi desenvolvido na mesma época por um segundo conjunto de pesquisadores é o CART (*Classification and Regression Trees*). De maneira similar, esses algoritmos possuem uma abordagem muito semelhante que se dá “de cima para baixo”. Idealmente cada ramo proveniente de um nó conteria apenas registros semelhantes de mesma categoria (folha). Para escolher qual atributo será testado em cada nó, frequentemente utiliza-se o conceito de entropia. O atributo de menor

entropia, em um certo nível da árvore sendo construída, será aquele escolhido para formar o nó desse nível.

Classificação Bayesiana (*Bayesian Classification*) – Compõem diversos métodos baseado no teorema de probabilidade condicional de *Bayes:* Probabilidade de (A|B) = Probabilidade de (B|A) X Probabilidade (A) / Probabilidade (B). Um algoritmo bastante popular – e que foi utilizado neste trabalho – é o Naive Bayes. Neste método, para facilitar o cálculo de probabilidade condicionada, os atributos são considerados independentes. Quando, de fato, esta é a realidade, o algoritmo costuma resultar em bons modelos preditivos, com a vantagem adicional de um rápido tempo de treinamento. Quando há dependência entre os atributos, uma alternativa é utilizar o método *Bayesian Belief Networks*. Neste caso, os atributos compõem os nós de um grafo acíclico direcionado. Há um arco partindo-se de um nó i para um nó j se o atributo j é dependente do atributo i.

Redes Neurais (*Neural Networks*) – Esse algoritmo simula o processo das estruturas neurais humanas. Segundo Haykin (2007), o cérebro humano processa informações de uma forma inteiramente diferente do computador digital convencional, de forma altamente complexa, não-linear e paralela. As redes neurais são inspiradas nesse tipo de mecanismo, absorvendo e distribuindo conteúdo em unidades simples, análogas aos neurônios, que possuem a propensão natural para armazenar conhecimento experimental e comunicam- se também entre si. Dessa forma, é possível simular um aprendizado compartilhado. De maneira resumida, os dados de entrada são processados por uma rede neural que pode possuir uma ou mais camadas intermediárias, *hidden layers*. Possuindo mais de uma camada intermediária, a Rede Neural é Classificada como Multicamadas. A cada camada são atribuídos aos atributos de entrada um peso. Do ponto de vista da estatística, o algoritmo realiza a tarefa de regressão, que pode ou não ser linear, a depender de transformações (funções de ativação) das saídas de cada nó. Durante o aprendizado, a rede ajusta os pesos das conexões de forma a melhor classificar o objeto de saída. É uma técnica que pode envolver longos períodos de treinamento e ajuste de parâmetros sendo difícil concluir a relação direta entre dados de entrada e saída. Um dos algoritmos de treinamento mais famosos de redes neurais é o *backpropagation*, em que os pesos são ajustados desde a camada de saída em direção à camada de entrada, sendo corrigidos de acordo com a diferença entre aquilo que se esperava como saída de um nó e o que de fato se obteve, dados os pesos vigentes. A Figura 13 ilustra o funcionamento de uma rede neural.

Figura 13 - Ilustração de Rede Neural.



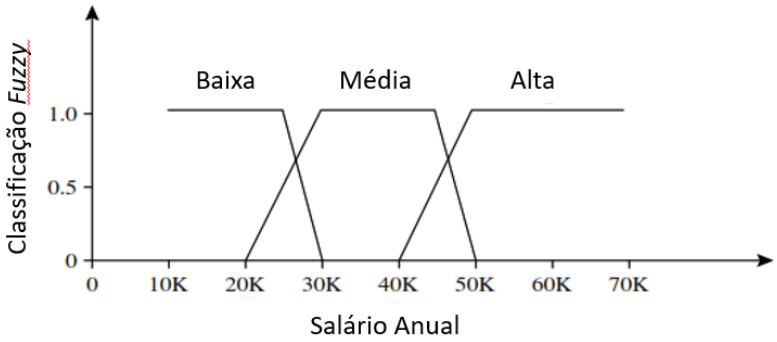
Fonte: Han; Kamber; Pei (2011, p.399)

SVM (*Support Vector Machines*) – Esse método objetiva definir o melhor hiperplano capaz de separa classes com melhor precisão e garantir a maior margem possível. Entende-se por margem a distância entre o hiperplano e o primeiro objeto de cada classe. O hiperplano é definido através de um problema de otimização de ordem quadrática. Quando a separação de objetos em classes não é linear, uma função *kernel* pode ser utilizada a fim de elevar o espaço dos atributos em uma dimensão maior, onde os objetos são mais facilmente separáveis pelo hiperplano sugerido.

Aprendizado Tardio (*Lazy Learners*) – Algoritmos de aprendizado tardio representam a classe de algoritmos em que o aprendizado só é efetuado quando mediante a entrada de novos dados, ou seja, existe um aperfeiçoamento do próprio algoritmo à medida que novos dados são adicionados. No entanto, esses métodos são extremamente exigentes de capital computacional e de confiança nos dados de entrada. Diferente de modelos como as redes neurais, que são pouco afetados por entradas inconsistentes, os algoritmos de aprendizado tardio têm seus resultados bastante dependentes da precisão e acuracidade dos dados de entrada, sendo bastante sensíveis a erros (De Raedt, 1998). Os algoritmos conhecidos como kNN (*k - Nearest Neighbor*) são algoritmos de aprendizado tardio que se popularizaram em 1960, com os avanços de capacidade computacional. Quando uma nova instância tem que ser classificado, o algoritmo procura as *k* instâncias mais próximas da instância sendo classificada. A classe dessa instância é então definida como aquela que for maioritária entre os vizinhos mais próximos encontrados.

Conjuntos Nebulosos (*Fuzzy Set*): Essa abordagem foi desenvolvida para lidar com atributos contínuos, como abordado por Han; Kamber; Pei (2011). Ao contrário da classificação baseada em regras que estabelece pontos fixos de corte para estabelecer a classe dos atributos, os conjuntos nebulosos fazem a classificação das instâncias com base na proximidade das classes definidas, isso é, não há um ponto de corte, ou faixa de valores delimitada, mas a aproximação dessas faixas. Não é uma classificação dicotômica, ou booleana, em que um objeto está dentro ou fora do valor estabelecido. É estabelecido um *score*, um peso, que irá demonstrar o quão próximo de determinado limite está o objeto analisado. Por exemplo, imagine que: "se SALÁRIO > 30K" ENTÃO; "crédito = aprovado" SENÃO; "crédito = negado". Em uma classificação baseada em regras, uma pessoa com salário de 29.999 não teria aprovação de crédito, embora o valor esteja extremamente perto do limite definido. Em uma avaliação nebulosa, no entanto, não seriam considerados apenas os valores em si, mas a proximidade dos valores aos requisitos definidos. Portanto, em uma análise de crédito, todo objeto pertenceria em diferentes proporções às categorias de aprovado e negado. E, no exemplo proposto, o cliente teria muito provavelmente seu crédito aprovado. Observe a Figura 14 que ilustra o funcionamento de uma classificação *Fuzzy* fictícia, que relaciona a chance de aprovação de crédito com salário anual.

Figura 14 - *Fuzzy*.



Fonte: Adaptado de Han; Kamber; Pei (2011, p.428)

Regressão Linear - As técnicas de regressão consistem em relacionar as chamadas variáveis independentes (preditoras) a uma variável dependente (solução). As variáveis preditoras são os atributos dos registros, enquanto a resposta é o que se quer prever. Portanto, trata-se de uma tarefa similar à de classificação, ou seja, também está é considerada uma tarefa de predição, mas, neste caso, estima-se um valor numérico ao invés de uma classe. As

regressões podem ou não ser de ordem linear, a depender da natureza das relações entre as variáveis (ex.: relações lineares, polinomiais ou logarítmicas).

2.4.3 Mineração de dados e desbalanceamento

Dizemos que um conjunto de dados está desbalanceado quando as classes se distribuem de maneira desproporcional em uma população estudada. A performance dos algoritmos de mineração de dados pode, portanto, ficar comprometida pela falta ou excesso de dados de determinadas categorias uma vez que a natureza desses algoritmos é essencialmente preditiva. Para exemplificar, podemos citar a classificação de pixels em imagens de mamografia para detectar possíveis sinais de ocorrência de câncer utilizando algoritmos (Woods, K. et al, 1993). De maneira geral, em casos como esse, cerca de 98% dos dados, isto é, pixels, são "normais" enquanto apenas 2% são "anormais" podendo ser indicativos de câncer, ou seja, a amostra é extremamente desbalanceada. Sem balancear a amostra, dada a natureza preditiva dos algoritmos, os resultados iriam sempre tender para uma predição negativa para câncer, uma vez que existem mais dados dessa natureza (98% contra 2%). No entanto, no presente exemplo, o custo de errar uma predição para ausência de câncer, quando na verdade há câncer é imenso. Para contornar problemas de desbalanceamento, algumas abordagens já foram propostas, das quais duas são aqui citadas. Uma é atribuir diferentes custos para as predições (Pazzani, M. et al, 1994) e (Domingos, 1999). A outra abordagem é rearranjar a amostra inicial, seja hiper-populando a categoria minoritária, seja sub-populando a categoria majoritária (Kubat; Matwin, 1998), (Japkowicz, 2000), (Lewis; Catlett, 1994), o que é feito através da criação de instâncias sintéticas.

Undersampling - Consiste em obter uma subamostra das instâncias da classe majoritária, que pode ser por uma seleção com ou sem reposição. Para Gonzalez (2008), o principal ponto negativo do método é que ele pode causar prejuízo à análise descartando dados que podem ser relevantes. Outro ponto relevante é que ao eliminarem-se dados, pode-se alterar a distribuição probabilística original da amostra uma vez que apenas os dados majoritários estão sendo removidos.

Gryzmala (2005) propôs uma abordagem interessante utilizando o conceito de *Tomek Link* (Tomek, 1976) em que remove dados da classe majoritária da seguinte forma: imagine que se tenha dois elementos Xma e Xmi, ambos pertencentes a classes distintas. Dizemos que

$d(Xma, Xmi)$ é a distância entre os dois objetos, o que é chamado de "par Tomek" ou *Tomek Link*. Caso não exista um exemplo X_1de tal modo que $d(Xma, X_1) < d(Xma, X_1)$ ou $d(Xmi, X_1) < d(Xma, Xmi)$ então duas possibilidades são possíveis, ou os dois exemplos (Xma e Xmi) são ruído ou "*borderlines*". Em ambos os casos a análise será beneficiada com essa exclusão. Esse método pode ser utilizado tanto para redução de ruído, quanto para balanceamento de classes quando apenas exemplos da classe majoritária são removidos.

Kubat e Matwin (1997) propuseram uma dinâmica com elementos das metodologias propostas pôr Hart (1968) e Grazymala (2008). Após a exclusão dos *Tomek Links*, propõem-se que sejam selecionados, aleatoriamente, um elemento da classe majoritária e todos elementos da classe minoritária para compor um conjunto C'. Então, deve-se usar um algoritmo do tipo *1-NN* usando os elementos de C' para classificar os elementos da amostra, conjunto inicial C. Todo exemplo que for classificado de maneira errada pelo algoritmo é então movido para o conjunto C'. A ideia por trás do método é eliminar aqueles dados, da classe majoritária, que são menos relevantes ou redundantes.

Oversampling – Neste caso, o balanceamento da frequência das classes é realizado através da replicação aleatória de objetos da classe minoritária. Um dos pontos negativos, fica evidente logo de início. Com a replicação de objetos, o perfil da distribuição probabilística é alterado o que somado à existência de réplicas favorece casos de *overfitting* dos modelos gerados, isso é, os modelos assumem a existência de regras e similaridades entres objetos artificiais, réplicas que não existem na amostra original. No entanto, existem algoritmos capazes de produzir instâncias sintéticas somadas a métodos heurísticos para evitar esse efeito de *overfitting*, como no caso do SMOTE. O SMOTE (*Synthetic Minority Over-sampling Technique*) é um dos algoritmos capazes de criar instâncias sintéticas das classes minoritárias a fim de balancear os dados. Primeiro, o SMOTE seleciona uma instancia aleatória A da classe minoritária e encontra os k-vizinhos mais próximos, selecionando aleatoriamente um vizinho B. Então, as instancias sintéticas são criadas por meio de uma combinação convexa das duas instâncias A e B.

Decisões sensíveis ao custo - Diferente das técnicas de *oversampling* e *undersampling*, essa técnica incorpora custos à tomada de decisão atribuindo valores fixos e diferentes para o erro de classificação de objetos. O modelo utiliza uma matriz de custo que é definida pelo analista de acordo com o problema estudado. De maneira geral a diagonal da matriz é nula, isso é, classificar um objeto da classe "a" como "a" tem zero custo, enquanto classificar de forma errada determinado objeto apresenta algum custo. Através da manipulação dos custos é possível construir um algoritmo que, por exemplo, entende como custoso definir um objeto da classe

minoritária como majoritária, mas não o contrário. Esse é o caso do algoritmo que analisa pixels de imagens de mamografias para identificar câncer de mama apresentado por (Woods, K. et al, 1999). No cenário em questão, atribuir um falso negativo, isso é, classificar uma imagem como negativa para câncer quando na realidade é positiva é extremamente indesejável, ao passo que um falso positivo não é tão problemático, afinal, uma vez que testes secundários sejam conduzidos o paciente saberá que não tem câncer.

Combinação de Métodos – Essas alternativas oferecem soluções de aplicações conjuntas de diferentes métodos, reconhecendo, assim, que ainda não é claro qual alternativa é melhor para lidar com o tratamento de desbalanceamento de dados. Resultados e trabalhos científicos recentes mostram que que a combinação de métodos diferentes costuma resultar em soluções melhores do que a aplicação de métodos isolados. (Casanova, 2005).

Yan e Liu (2003), por exemplo, testaram diferentes abordagens para a classificação de dados desbalanceados e chegaram à conclusão que, o método SVM aliado a um pré-processamento e agrupamento foi uma ótima técnica para lidar com a presença de casos raros. Weiss e Provost (2003), por sua vez, realizaram uma pesquisa que estuda a abordagem de um algoritmo iterativo que executa uma seleção progressiva que produz a cada etapa conjuntos de dados de treinamento baseados na performance da classificação anterior, alterando, de acordo com a acurácia, a proporção de classes adicionadas na etapa seguinte.

Alguns estudos também abordam os chamados *Boosting Algortihms*, algoritmos iterativos que podem ser combinados à técnica de seleção sensível a custo. Esses algoritmos aumetam os pesos ou custos associados à classificação incorreta e diminuem os custos das classificações corretas a cada iteração, forçando o aprendizado das classificações erradas a cada etapa. Esse é o caso do Adacost. O Adacost é um algoritmo que resultou da combinação do Adaboost e técnicas de seleção sensível ao custo. Em (Fan, W. et al, 1999) o Adacost mostrou uma melhor performance que Adaboost, mostrando que a combinação de técnicas é uma solução mais adequada em muitos casos.

Outra combinação interessante, proposta por Chawla, N. V. et al (2003) é o SMOTE-*Boost.* Algoritmos do tipo "*Boost*" possuem uma tendência de *overfitting* uma vez que a cada iteração os pesos são ajustados para focar na eliminação de falsos negativos, o que tende a gerar um viés para falsos positivos. Para lidar com esse problema, o SMOTE-*Boost* combina as iterações do *Boost* com a mecânica do SMOTE. Ao invés de alterar a distribuição dos dados de treinamento a cada iteração, o SMOTE-*Boost* altera a distribuição adicionando novos dados sintéticos da classe minoritária por meio da técnica do SMOTE.

2.4.3.1 Métricas de performance

Sabendo-se que existem diversos algoritmos e técnicas de mineração, são necessárias métricas para avaliar os resultados, de maneira a compará-los e medir a qualidade individual de cada um, especialmente no que se refere à generalidade dos modelos obtidos. Na questão de se medir a generalidade, em particular, o ideal é aplicar o modelo resultante em um conjunto de dados para validação que seja totalmente independente dos dados utilizados para treino. Normalmente, isto não é possível, pois conseguir dados suficientes para treinamento já é por si só complexo.

Assim, durante a etapa de mineração, para evitar que modelos e suposições sejam inferidos a partir de um único subconjunto, é comum que os dados sejam divididos em alguns subsets, isso é, subconjuntos. A tendência é que quanto menor o número de subconjuntos utilizados para desenvolver o modelo, menor será sua precisão quando aplicado aos demais subconjuntos pois o modelo estará restrito aos subconjuntos selecionados inicialmente. A esse fenômeno dá-se o nome de bias, do inglês "tendencioso" ou "enviesado". Para contornar esse efeito, os dados costumam ser divididos em ao menos três subconjuntos, sendo eles:

1 - Training Set ou Conjunto de Treinamento: Um subconjunto de dados é selecionado e utilizado para desenvolver um modelo através de aprendizado supervisionado.

2- Test Set ou Conjunto de Testes: É um subconjunto independente utilizado para testar o modelo desenvolvido a partir do Training Set.

3- Validation Set ou Conjunto de Validação: Um subconjunto de dados também independente e, da mesma forma, para validação do modelo, mas é utilizado para ajustar os parâmetros do algoritmo de aprendizado empregado. Somente após esgotado os ajustes e construído o modelo final, este é avaliado pelo test set.

Estas análises também são importantes no contexto de desbalanceamento de classes. De maneira geral, sem aplicar técnicas para "balancear" os dados, quando analisa-se uma amostra de dados desbalanceados, os dados de classe minoritária são ignorados, afinal, como descrito anteriormente, o algoritmo tentará ser o mais eficiente possível entendendo os dados raros como ruído, promovendo classificações tendenciosas para casos majoritários. Por isso, a média de acuracidade (taxa de classificações corretas por total de classificações) não costuma ser uma boa métrica. No domínio estudado por Lewis e Catlett (1994), apenas 0,2% exemplos são positivos e um algoritmo padrão, que não considera o desbalanceamento irá atingir 98,8% de acuracidade ignorando a classe minoritária. Ou seja, apesar de alta acurácia o modelo não é capaz de fazer uma classificação minimamente útil. Portanto, é importante considerar a precisão

do algoritmo na classificação de cada classe e não exclusivamente do número total de classificações corretas e incorretas.

A título de melhor entendimento, a literatura costuma abordar os problemas dessa natureza através de um estudo de caso padrão de classes binárias: positiva e negativa. A classe positiva é convencionada como a minoritária e a negativa aquela que é majoritária. A Tabela 1 apresenta a matriz de confusão, popularmente utilizada na interpretação de resultados de algoritmos onde TP (*true positives*) e TN (*true negatives*) quantificam, respectivamente, testes positivos e testes negativos classificados corretamente, e FP (*false positives*) e FN (*false negatives*) os testes positivos e negativos classificados de forma errônea, os falsos positivos e falso negativos.

Tabela 1 - Matriz de Confusão.

Instância/Classificação	Classificado Como Negativo	Classificado Como Positivo
Negativo	TN	FP
Positivo	FN	TP

Fonte: (Autor, 2020)

Além esses valores absolutos, existem ainda estatísticas que podem ser calculadas a partir da matriz de confusão, que são apresentadas nas Equações 2.1, 2.2, 2. 3, 2.4 e 2.5, onde β (Equação 2.5) diz respeito a importância relativa entre Precisão e Sensibilidade, normalmente definida como igual a 1.

$$Acuracidade = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.1}$$

$$TaxaFP = \frac{FP}{TN + FP} \tag{2.2}$$

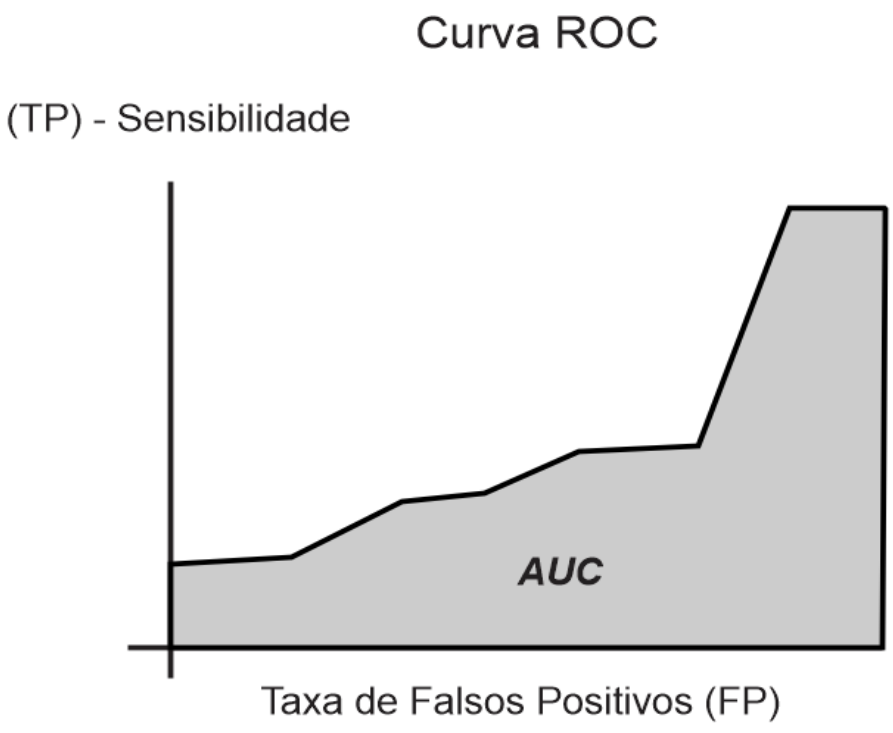
$$TaxaTP, Sensibilidade = \frac{TP}{TP + FN} \tag{2.3}$$

$$Precisão = \frac{TP}{TP + FP} \tag{2.4}$$

$$Fvalor = \frac{(1+\beta^2) * Sensibilidade * Precisão}{\beta^2 * Sensibilidade + Precisão} \quad (2.5)$$

A relações também podem ser exploradas pela curva ROC, *Relative Operating Characteristics*, onde o eixo horizontal diz respeito à taxa de Falsos Positivos (FP) e o eixo vertical diz respeito aos casos positivos classificados de maneira correta (TP), sensibilidade. A ROC irá dar forma às matrizes de confusão (que exploram as métricas TP, TN, FP e FN) de acordo com os limites escolhidos para classificação. A Figura 15 representa uma curva ROC fictícia, através dela é possível observar que existe um *tradeoff*, um dilema fundamental onde a melhora da classificação de determinada classe se dá ao custo de perda de acuracidade na outra classificação.

Figura 15 - Curva ROC.



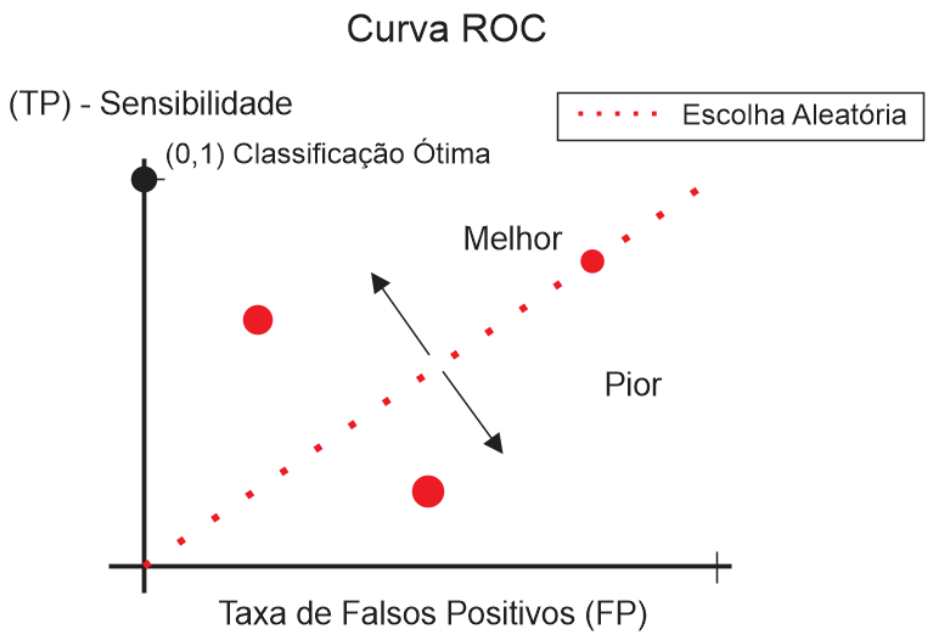
Fonte: (Autor, 2020)

A área sob a curva (AUC, Area *Under* ROC *Curve*) está associada à capacidade de classificação de determinado modelo. Quanto maior a AUC, melhor o poder preditivo, pois

significa que há um ponto de escolha dos limites de classificação em que se consegue obter alto TP e baixo FP.

O melhor modelo de previsão possível seria representado pelo ponto (0,1), no canto superior esquerdo do plano descrito pelo espaço ROC. Nesse ponto a taxa TP, ou sensibilidade, é de 100%, assim como a especificidade *(TN/ TN+FP)*. Portanto não existem falsos positivos ou negativos. Para compreender melhor o que a curva ROC representa, note a diagonal imaginária representada na Figura 16. Qualquer ponto sob a diagonal representa soluções em que *TP = FP*, isso é, a proporção de instancias corretamente classificadas é a mesma que instancias incorretamente classificadas. Soluções aleatórias, em caso de distribuições simétricas, produziriam pontos pertencentes à linha diagonal do espaço (Figura 16). Por isso, considera-se soluções que estejam acima da diagonal soluções melhores que aleatórias enquanto as que se situam abaixo da diagonal são soluções piores que seleções aleatórias.

Figura 16 - Curva ROC, Diagonal de Aleatoriedade.



Fonte: (Autor, 2020)

A fórmula “valor de F” (5) é também muito popular nesse tipo de análise. É uma média harmônica entre TP (3) e Precisão (4). Um alto valor de F indica que tanto precisão (4) quanto TP rate estão altos. A fórmula pode ainda ser ajustada através do parâmetro β, que corresponde ao peso dado à relação precisão (4) e taxa de TP (3).

2.5 Softwares

Hoje estão disponíveis uma grande quantidade de ferramentas desenvolvidas para a Mineração de Dados, sendo grande parte delas amigáveis a profissionais de outras áreas que não a de ciência de dados. Alguns *softwares* e ferramentas populares são: *Clementine* (desenvolvido pela SPSS), IBM *Intelligent Miner* (desenvolvido pela IBM), *Oracle Data Mining* (desenvolvido pela Oracle) e o *Weka*.

Sobre o *Weka*, trata-se de um software de código aberto emitido sob a GNU, *General Public License* que disponibiliza uma série de algoritmos para mineração. Os algoritmos podem ser aplicados utilizando-se a GUI (Graphical User Interface), sem necessidade de programação, ou utilizados importando-se as bibliotecas disponibilizadas para programas Java. Os algoritmos disponíveis no Weka podem ser utilizados em atividade de pré-processamento, classificação, regressão, agrupamento, regras de associação e visualização. Por se tratar de um programa "aberto" é muito comum sua utilização na comunidade acadêmica.

3 ESTUDO DE CASO

Uma vez compreendidos os principais pontos referentes ao funcionamento de motores aeronáuticos a jato, a natureza e impactos gerados por vibração em sistemas mecânicos e as principais metodologias e ferramentas para mineração de dados, será dado início ao estudo de caso do presente trabalho com o objetivo de explorar os temas abordados durante a revisão, trazendo descobertas e contribuindo com a fomentação de mais estudos na área. O estudo será divido em três grandes seções, são elas: Abordagem Inicial, Preparação, Mineração de dados e Conclusão. Cada qual organizada e dividida em subprocessos que serão abordados no início de cada seção.

Dada a sensibilidade dos processos executados, especialmente no que se refere ao sigilo exigido pela instituição fornecedora dos dados, os processos não serão totalmente abertos. Os atributos analisados serão abordados de maneira genérica e terão seus nomes simbolizados por valores inteiros positivos aleatórios.

De acordo com a metodologia escolhida, os dados da empresa foram coletados de duas fontes principais: primeiro, os que Turroni e Mello (2012) chamam de "peso-leve", que são coletados através de entrevistas, observações e discussões entre mecânicos, engenheiros e gestores da produção. Esse tipo de dados é de grande relevância uma vez que consideram a experiência e visão dos processos por meio da percepção dos agentes. segundo os "pesos-pesados", que são dados predominantemente quantitativos e que irão ser traduzidos no formato

matricial de dados. Os atributos relacionados à montagem dos motores, chamados de "*Data Names*" e das assinaturas de vibração de N2 durante teste.

3.1 Abordagem inicial

Durante essa fase, a questão central, definida no início desse trabalho (qual conjunto de atributos são mais influentes no excesso de vibração do eixo de alta rotação do motor (N2) e qual conjunto de algoritmos é capaz de gerar o melhor modelo de classificação/predição de instâncias para o cenário estudado?), foi revisitada. Foi identificado que a rejeição de motores para o excesso de vibração se traduzia em um grande ônus para a empresa. Nesses casos, os motores deveriam ser encaminhados novamente para linha de produção, seguindo para o processo de desmontagem e, então, serem remontados. Como não existem modelos para prever a rejeição dos motores, ou mesmo estudos realizados pela empresa para melhor compreender a relação histórica das vibrações com os registros de montagem, esse trabalho irá promover o estudo das relações entre os atributos de montagem e o excesso de vibração. Assim como, também irá propor o melhor modelo de previsão dentre as alternativas estudadas. Dessa forma, a empresa contará com um método valioso para analisar se o motor deve seguir para teste ou não.

3.2 Preparação

Todo o processo e metodologia, seguidos da condução das análises que se sucedem, tiveram como base os estudos e modelos de conhecimento extraídos de dados apresentados for Han; Kamber; Pei (2011) e Fayyad; Piatetsky; Smyth (1996). Ambos os modelos possuem como pilares da investigação a compreensão clara dos processos ao qual o problema analisa se circunscreve e a definição objetiva dos atributos e classes (conjunto de dados) que serão analisados. A essa etapa convencionou-se chamar "Preparação".

Durante a preparação foram conduzidos estudos sistemáticos das operações de montagem dos motores, descrevendo e compreendendo as principais atividades que são realizadas durante os processos internos da empresa e como seus resultados são registrados.

Identificou-se que os dados referentes a atributos de montagem são registrados em um sistema próprio da empresa, elaborado sob medida para atender aos requisitos da operação e que todos os valores dos processos mais importantes referentes à montagem são registrados

como valores numéricos pelos próprios operadores. Sendo o registro da operação condição para que se possa dar continuidade aos processos.

Uma vez identificado que existem grandes quantidades de dados disponíveis para análise, foi necessário selecionar quais atributos seriam analisados. A análise de todos os dados disponíveis não foi vislumbrada uma vez que a maior parte dos dados de montagem, empiricamente, pouco tem a ver com vibração. Com o intuito de diminuir o tempo de processamento de dados e organizar uma análise mais concisa e objetiva, deu-se início a uma pré-seleção de atributos.

O processo de definição de atributos foi feito sob tutela dos especialistas da área (engenheiros, técnicos e gestores) somadas aos estudos desenvolvidos na revisão de literatura do presente trabalho. Atributos associados a peças rotativas e seus respectivos balanceamentos, por exemplo, foram selecionadas sem exceções uma vez que as relações de causa e efeito com vibração são bem conhecidas. Ficou definido que os processos a atributos estudados seriam restritos aos processos de montagem do compressor de alta pressão, turbina de alta pressão e operações relacionadas a checagem do compressor e peças rotativas, que estão especialmente sujeitas a efeitos de desbalanceamento mecânico. Atributos com baixa probabilidade de relação com vibração foram excluídos dessa análise, assim como valores categóricos relacionados ao tipo do motor, área de teste, horário etc.

Ao fim do processo de preparação, o conjunto de dados a ser analisado começou a tomar forma, sendo composto por 120 atributos e 380 instâncias. Observou-se, nesse momento, que os dados referentes a teste, isso é, atributos relacionados à vibração e à definição de classe de cada instância (motor aprovado ou motor rejeitado) estava registrada em um segundo banco de dados.

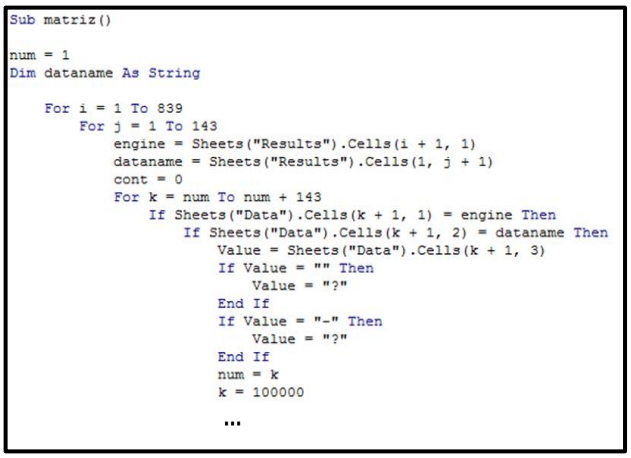
3.2.1 Integração

Para definir o conjunto de dados a ser analisado, foi necessário combinar os atributos pré-selecionados na fase de preparação com os atributos e classe registrados em um segundo banco de dados. Esse processo de combinação de dados é muito comum e chamado, popularmente, como “Integração” segundo Han; Kamber; Pei (2011).

O processo de integração foi executado por meio do software Excel e a elaboração de um código VBA, Visual Basic for Applications, linguagem de programação utilizada para construir algumas das aplicações do Microsoft Office. Dessa forma, o conjunto de dados de

montagem foi combinado ao conjunto de dados de teste. Uma secção do código elaborado em VBA pode ser observada na Figura 17.

Figura 17 - Parte do Código VBA utilizado.



```
Sub matriz()

num = 1
Dim dataname As String

    For i = 1 To 839
        For j = 1 To 143
            engine = Sheets("Results").Cells(i + 1, 1)
            dataname = Sheets("Results").Cells(1, j + 1)
            cont = 0
            For k = num To num + 143
                If Sheets("Data").Cells(k + 1, 1) = engine Then
                    If Sheets("Data").Cells(k + 1, 2) = dataname Then
                        Value = Sheets("Data").Cells(k + 1, 3)
                        If Value = "" Then
                            Value = "?"
                        End If
                        If Value = "-" Then
                            Value = "?"
                        End If
                        num = k
                        k = 100000
                        ...
```

Fonte: (Autor, 2020)

Os dados foram combinados e formatados, inicialmente, em um arquivo do tipo ".csv", formato comum para bancos de dados e aceito em vários softwares voltados à ciência dos dados. Ao término do processo de integração, chegou-se à definição do conjunto de dados para iniciar o processo de mineração de dados , composto por 143 atributos (todos numéricos), o atributo classe (aprovado/rejeitado) e 380 instâncias. Quanto à classe de estudo, trata-se de uma classe binária, isto é, os motores são classificados como rejeitados ou aprovados de acordo com um parâmetro de corte associado à vibração de N2. Esse valor de corte foi fixado pela engenharia (acordo de sigilo), sendo todas as instâncias acima do valor estipulado classificadas como rejeitadas.

3.3 Mineração de dados

Nessa etapa serão abordados os processos utilizados na mineração de dados, lembrando que todas as técnicas aqui utilizadas foram também abordadas na revisão de literatura.

3.3.1 Análise inicial

Com o conjunto de dados bem definidos, deu-se prosseguimento à fase de análise inicial dos dados e pré-processamento através do software, de uso público, Weka. Uma análise

superficial dos dados indicou que muitos dos atributos possuíam dados faltantes, problema comum em análise de dados. Para evitar falsas interpretações, foi adotado o seguinte critério: todos os atributos com mais do que 70% de atributos faltantes foram eliminados da análise. Sendo assim, foram descartados da análise 64 atributos. Esse processo foi acompanhado pelas partes interessadas que confirmaram que os parâmetros em questão tratavam-se de operações não obrigatórias, ou seja, não executadas para todos os motores. Restaram, portanto, 79 atributos para estudo e compreensão.

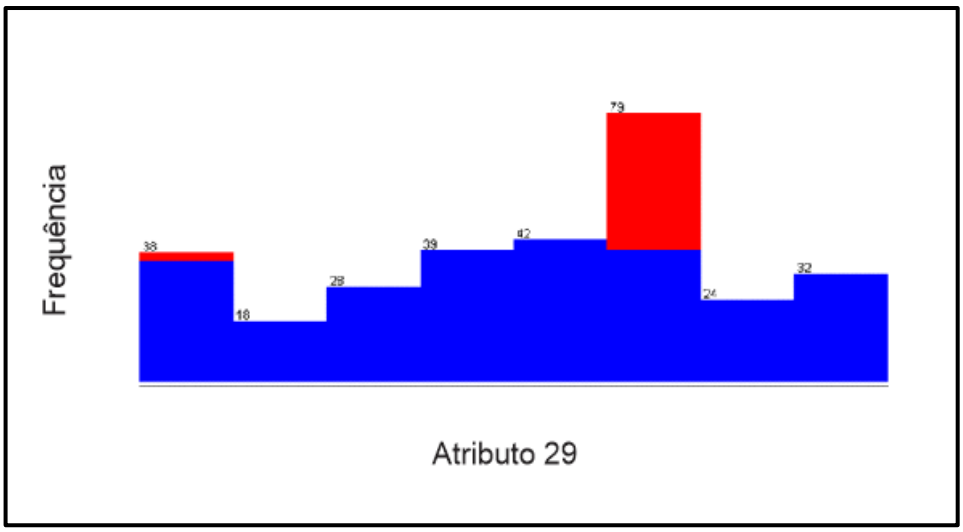
Com a utilização do Weka, foram obtidas estatísticas básicas para todos os atributos: mínimo, máximo, média e desvio padrão e histogramas. Os histogramas, automaticamente produzidos pelo software, ainda diferenciam a classe dos dados através de cores. Para esse caso, em específico, adotou-se o vermelho para designar os motores rejeitados, isso é, as classes positivas, e a cor azul para designar classes negativas, isso é, motores aprovados.

Observe que adotou-se, conforme já descrito, a convenção de chamar as instâncias da classe minoritária (motores rejeitados) de instâncias positivas.

Uma análise visual inicial já permite identificar que certos atributos de montagem se relacionam com a definição de classe. É possível observar que certas instâncias positivas se concentram em certas faixas de parâmetros, não se distribuindo igualmente ao longo do eixo X.

A Figura 18 demonstra um dos casos em que é possível observar a maior frequência de instâncias positivas em determinada faixa do atributo 29, selecionado.

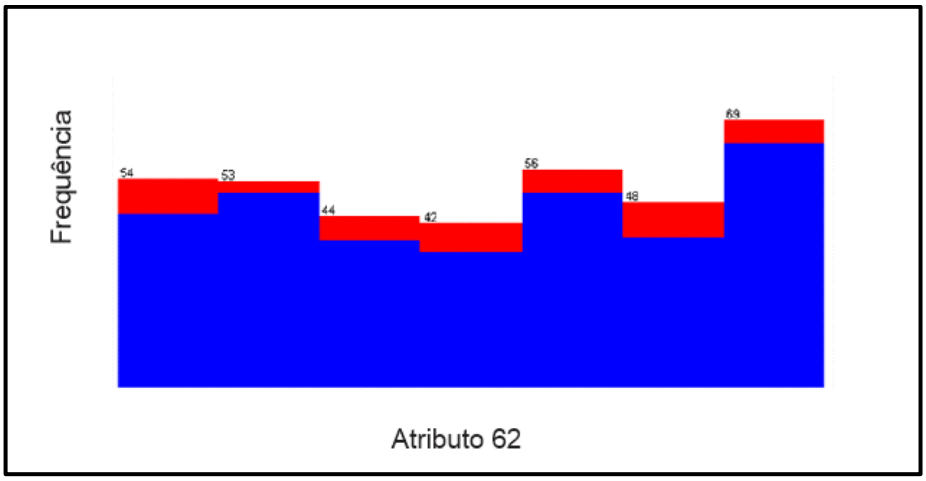
Figura 18 - Histograma - Atributo 29.



Fonte: (Autor, 2020)

Já com relação à maior parte dos dados de montagem, não é possível verificar uma correlação clara visualmente, como mostrado na figura 19. Percebe-se que há motores reprovados distribuídos por toda a faixa de valores do atributo de montagem.

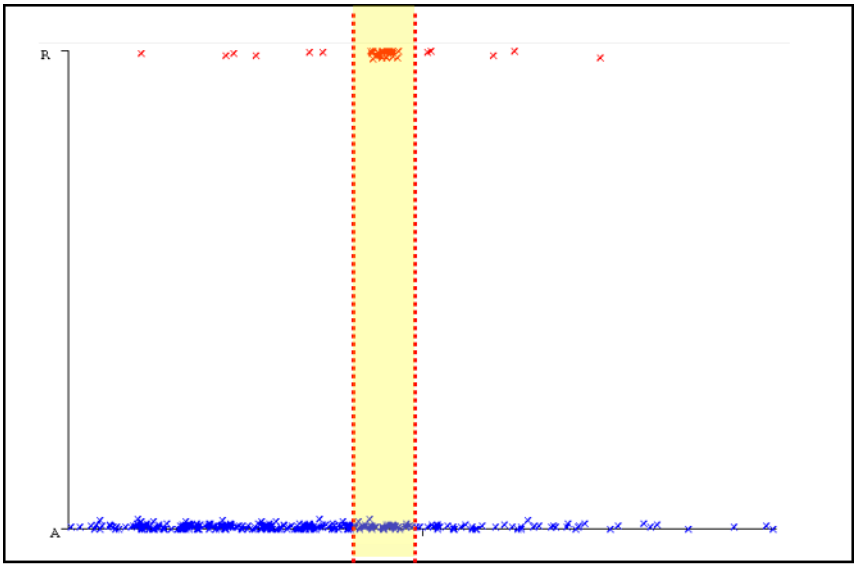
Figura 19 - Histograma – Atributo 62.



Fonte: (Autor, 2020)

Outro recurso interessante, também utilizado durante a análise inicial, foi a visualização da distribuição de instâncias em um plano. Essa é uma excelente ferramenta para a interpretação inicial dos dados. Observe as Figuras 20-23, que dizem respeito à distribuição de motores, em diferentes planos.

Figura 20 - Distribuição de motores, Classe vs. Atributo 6.

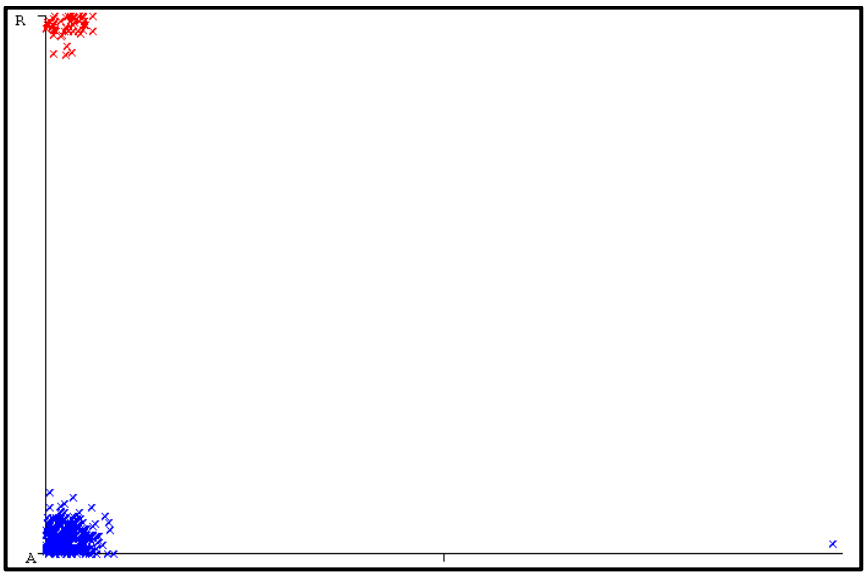


Fonte: (Autor, 2020)

Através desse recurso visual é possível interpretar que segundo a Figura 20 parece haver uma relação entre o atributo 6 e definição de classe. Isso porque os pontos referentes a motores rejeitados concentram-se em determinada faixa (área hachurada) do atributo 6, comportamento que não é visto em motores aprovados (pontos azuis).

Em contra partida, a Figura 21 mostra um caso em que não há relação aparente entre o parâmetro e a classe. Observe que na Figura 21 tanto os motores rejeitados quanto aprovados distribuem-se de maneira similar pelo eixo do atributo 8.

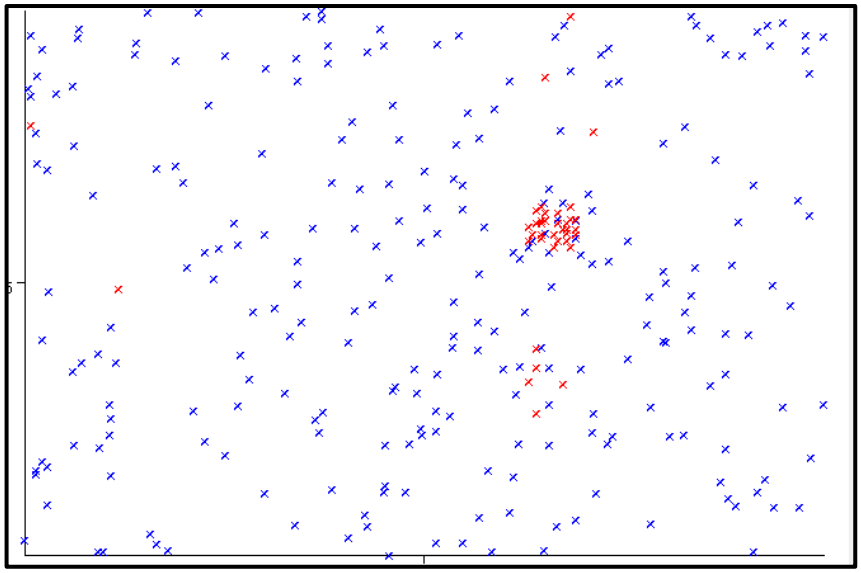
Figura 21 - Distribuição de motores, Classe vs. Atributo 8.



Fonte: (Autor, 2020)

A Figura 22 mostra a relação entre dois atributos, ou seja, desta vez a classe não é considerada. Nesse caso, observa-se a relação entre a distribuição dos motores e dois atributos, o atributo 77 e o atributo 29. É possível observar que motores rejeitados estão concentrados em certa faixa do gráfico, em contraste com os motores aceitos. Essa visualização permite concluir que deve haver uma relação entre os dois parâmetros e a composição de classe.

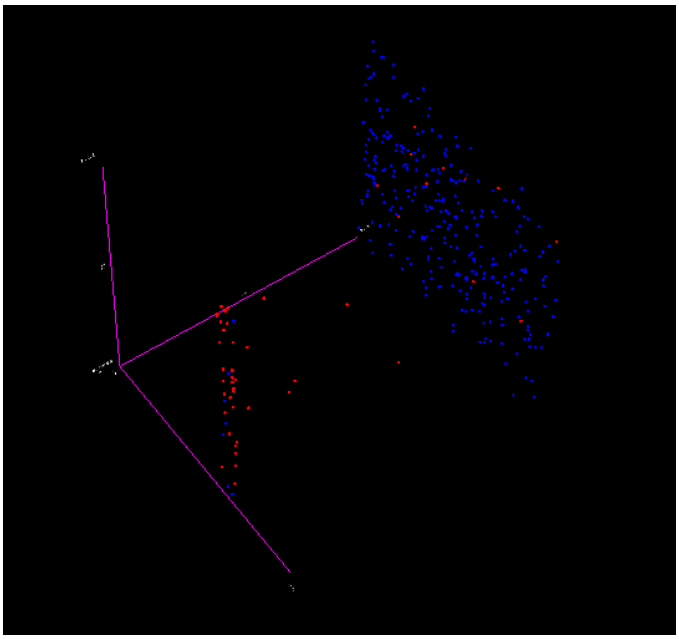
Figura 22 - Distribuição de motores, Atributo 77 vs. Atributo 29.



Fonte: (Autor, 2020)

De maneira análoga, o Weka também permite a visualização da distribuição dos motores em três dimensões, adicionando mais um eixo à análise, como pode ser observado na Figura 23.

Figura 23 - Distribuição de motores, Atributo 19 vs. Atributo 21 vs. Atributo 43.



Fonte: (Autor, 2020)

Como conclusão da etapa de análise inicial foi também observado que o conjunto de dados é bem desbalanceado, isso é, a quantidade de instâncias da classe negativa é muito superior à quantidade de classes positivas. A proporção é de, aproximadamente, 7:1.

3.3.2 Pré-processamento, seleção de atributos

A fase de pré-processamento é vital para garantir uma análise de dados precisa. A maior parte dos algoritmos de mineração já executa estratégias de pré-processamento em suas aplicações, no entanto, como esse trabalho objetiva identificar os atributos que mais se correlacionam com taxas de vibração, essa etapa do trabalho é dedicada a explorar dois algoritmos de seleção de atributos, isso é, de identificação de atributos mais relevantes para a análise. Lembrando que algoritmos de seleção são recursos interessantes para minimizar os tempos de processamento, quando se lida com muitos atributos, e aumentar o grau de precisão das análises removendo certos atributos que contribuem para a composição de ruído.

A amostra de dados analisada possui ao todo 79 atributos. Para avaliar a influência dos atributos na classificação das instâncias, utilizaram-se dois algoritmos de seleção de atributos na amostra de dados sem nenhuma utilização de técnica de balanceamento ou outro pré-processamento.

O primeiro algoritmo testado foi o CfsSubset em conjunto com a heurística *BestFirst*, com o critério de parada configurado para 5 subconjuntos consecutivos de performance inferior. O *CfsSubsetEval* fornece como resultado um subconjunto de atributos relevantes para a definição de classe. Observe os resultados da aplicação desse filtro pelo software Weka na Figura 24. Lembrando que os nomes dos atributos foram convertidos em números inteiros de 1 a 79, sendo 80 o atributo referente à classe.

Figura 24 - Resultados do Algoritmo CfsSubset.

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 818
        Merit of best subset found:    0.529

Attribute Subset Evaluator (supervised, Class (nominal): 80 80):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,19,29,50,51,61,68,71,72,75,79 : 11
                     2
                     19
                     29
                     50
                     51
                     61
                     68
                     71
                     72
                     75
                     79
```

Fonte: (Autor, 2020)

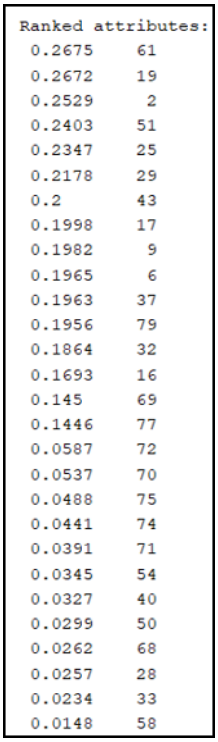
Como foi abordada na revisão de literatura, o Cfs compara a performance de subconjuntos de atributos por meio de heurísticas. Entende-se que os atributos selecionados nesse subconjunto se relacionam de forma a influenciar a classe.

Como resultado do experimento, o subconjunto proposto pelo atributo foi limitado a 11 dos 79 atributos, uma redução de cerca de 86% dos parâmetros analisados. Sendo identificados como os parâmetros mais relevantes os de número: 2, 19, 29, 50, 51, 61, 68, 71, 72, 75 e 79.

O segundo algoritmo de seleção testado foi o InfoGain que, diferente do CfsSubset, identifica a relevância de cada atributo individualmente. Como resultados, tem-se um índice de relevância para cada atributo da amostra, descritos em ordem decrescente de relevância.

Observe os resultados na Figura 25 extraída do software Weka. Foram selecionados 28 atributos (com score > 0) com algum grau de correlação com a classe. Propondo uma redução de cerca de 64,5% dos atributos estudados.

Figura 25 - Resultados do Algoritmo InfoGain.



```
Ranked attributes:
 0.2675    61
 0.2672    19
 0.2529     2
 0.2403    51
 0.2347    25
 0.2178    29
 0.2       43
 0.1998    17
 0.1982     9
 0.1965     6
 0.1963    37
 0.1956    79
 0.1864    32
 0.1693    16
 0.145     69
 0.1446    77
 0.0587    72
 0.0537    70
 0.0488    75
 0.0441    74
 0.0391    71
 0.0345    54
 0.0327    40
 0.0299    50
 0.0262    68
 0.0257    28
 0.0234    33
 0.0148    58
```

Fonte: (Autor, 2020)

3.3.2.1 Comparação de algoritmos de seleção

A fim de compreender qual dos algoritmos de seleção melhor performou no conjunto de dados em questão, realizou-se o seguinte experimento. Seis algoritmos de classificação foram aplicados em três cenários diferentes. No primeiro cenário, os seis algoritmos de mineração foram aplicados no conjunto de dados original, sem nenhum tipo de pré-processamento ou aplicação de algoritmo de seleção de atributos. Os resultados desses testes serviram como base comparativa para a performance dos cenários seguintes, em que foram aplicados antes do uso de algoritmos de classificação o algoritmo *CfsSubset* (cenário 2) e o algoritmo *IngoGain* (cenário 3).

Os resultados podem ser observados na Tabela 2, em que se utilizou a métrica "F-valor" para efeitos de comparação uma vez que essa métrica corresponde à média harmônica dos valores da taxa TP e Precisão. Os algoritmos de classificação testados foram: Naive Bayes, kNN, Adaboost, J48, Random Forest e SMO (*Sequential Minimal Optimization*), que é uma implementação do método SVM. É importante relembrar que quando mais próximo de 1 é o F-valor, melhor o poder preditivo do modelo gerado.

Tabela 2 - Comparação F-valor por Algoritmo de Classificação. Para o kNN, considerou-se os 10 vizinhos mais próximos.

Seletor	Cenário	Classificador	F-Valor
Nenhum	1	Naive Bayes	0,932
Nenhum		kNN (10)	0,953
Nenhum		AdaBoost	0,909
Nenhum		Random Forest	0,953
Nenhum		SMO_Default	0,953
Nenhum		J48	0,937
		Média	0,9395
CfsSubset	2	Naive Bayes	0,946
CfsSubset		kNN (10)	0,953
CfsSubset		AdaBoost	0,96
CfsSubset		J48	0,953
CfsSubset		Random Forest	0,953
CfsSubset		SMO_Default	0,946
		Média	0,951833
InfoGain	3	Naive Bayes	0,953
InfoGain		kNN (10)	0,953
InfoGain		AdaBoost	0,953
InfoGain		Random Forest	0,953
InfoGain		SMO_Default	0,946
InfoGain		J48	0,944
		Média	0,950333

Fonte: (Autor, 2020)

A Tabela 3 resume os melhores resultados (em amarelo na Tabela 2) e mostra, em termos percentuais, quanto a aplicação dos algoritmos de seleção melhoraram o F-valor.

Tabela 3 - Comparação de Melhora do F-valor médio pós Seleção de Atributos.

Seletor	Cenário	F-Valor, Média	Melhora (Percentual) Pós aplicação de Seletor
Nenhum	1	0,9395	
CfsSubset	2	0,951833333	1,3128%
InfoGain	3	0,950333333	1,1531%

Fonte: (Autor, 2020)

É possível observar que em ambas as aplicações, os resultados das classificações melhoram ainda que com um menor número de atributos em consideração. Esse é um exemplo de que métodos de seleção são um conjunto de ferramentas importantes na análise de dados, capazes não só de diminuir tempo de processamento como melhorar a performance de mineração. Para o caso em questão, o algoritmo *CfsSubset* mostrou-se um pouco melhor.

Com relação à seleção de atributos, é curioso observar que todos os atributos selecionados pelo *CfsSubset* foram também selecionados pelo *InfoGain*, sendo que a grande maioria possui influência já conhecida do ponto de vista de engenharia. A maioria diz respeito a valores de balanceamento das peças rotativas, como concentricidade, circularidade, *run-out* axial e *run-out* radial.

Essas relações já são conhecidas como as principais causas de vibração em sistemas rotativos, no entanto, algumas relações são novas e poderão ser melhor exploradas pela empresa e em possíveis trabalhos futuros.

Dada a confidencialidade de dados, não será possível aprofundar nos resultados e relações entre os atributos mais relevantes.

3.3.3 Construção de modelos preditivos para diversos cenários

Nessa etapa, foram testadas diferentes combinações de algoritmos para definir qual melhor performa no cenário estudado. Uma vez identificada a melhor combinação, a empresa terá a seu dispor um modelo para previsão de rejeição de motores pós teste. Essa será a maior contribuição desse trabalho.

3.3.3.1 Definição das métricas

O conjunto de dados estudado, como visto na análise inicial, é um conjunto desbalanceado. Esse tipo de situação não é nova na literatura e já foi abordada por muitos autores e explorado nesse trabalho. Para efeitos de métricas, portanto, a acuracidade (taxa de acertos totais, ou seja, para ambas as classes) é pouco relevante uma vez que um algoritmo que possuísse um viés de classificação de instâncias como negativas teria ainda uma alta taxa de acuracidade. Por exemplo, caso o algoritmo opte por ignorar as classes positivas, sempre classificando qualquer classe como negativa, ainda teria uma acuracidade relativamente alta, próxima a 88%.

Por isso, as análises dos resultados das combinações exploradas não basearam-se nos valores de acuracidade. Como o objetivo principal da análise é desenvolver um modelo que melhor identifique instâncias positivas, isso é, de rejeições, as métricas referentes à classificação de instâncias positivas serão as mais relevantes para o estudo. As principais métricas escolhidas foram:

- Taxa TP ou sensibilidade: Taxa de verdadeiros positivos, i.e., instâncias corretamente classificadas de uma classe. Em especial, houve maior foco na TP para a classe de rejeição, isso é, a taxa de instâncias de rejeição corretamente classificadas.

- Taxa FP: Taxa de falsos positivos, i.e., instâncias incorretamente classificadas de uma classe. Houve atenção especial aos casos de FP para a classe de aprovados, isso é, os casos em que instâncias de rejeição foram classificadas como aprovadas. Deseja-se, obviamente, um valor de FP próximo a zero.

- Precisão: Dentre todas as instâncias classificadas como positivas, consiste na proporção daquelas que, de fato, são positivas. Nesse caso, quanto mais próxima de 1 for o valor da precisão, melhor.

- F-valor: É a média harmônica das métricas precisão e sensibilidade. Um F-valor igual a 1 indica que as instâncias são perfeitamente classificadas, isso, é apresentam tanto sensibilidade quanto precisão iguais a 1. É uma medida muito relevante em casos como o estudado, em que há desbalanceamento.

- ROC Área, ou AUC – Diz respeito a área da curva ROC. É uma métrica muito importante nessa análise, oferecendo uma visão global sobre a performance dos algoritmos em todos os *thresholds* de classificação possíveis. Quanto maior o valor da área ROC, melhor a performance do algoritmo.

3.3.3.2 Definição dos algoritmos e cenários de estudo

Nessa etapa, serão definidos os algoritmos que serão testados e suas combinações. Serão alvo desse estudo comparativo 6 algoritmos de mineração (NaiveBayes, kNN, SMO, AdaBoost, J48 e Random Forest) combinados a 2 algoritmos de seleção de atributos (*CfsSubset, InfoGain*) e um algoritmo de balanceamento (SMOTE). O que irá resultar em 36 combinações possíveis. A título de melhor compreensão e entendimento, as etapas de teste serão divididas em 6 cenários. São eles:

- Cenário 1: Conjunto de dados sem tratamento. Isso é, os algoritmos foram aplicados nos dados sem nenhum algoritmo de pré-processamento, como um algoritmo de seleção de atributos ou um algoritmo para balanceamento de dados, como o SMOTE.

- Cenário 2: Conjunto de dados previamente submetidos ao algoritmo de seleção de atributos *CfsSubsetEval* utilizando a heurística *BestFirst.*

- Cenário 3: Conjunto de dados previamente submetidos ao algoritmo de seleção de atributos InfoGain.

- Cenário 4: Conjunto de dados previamente balanceados a partir da aplicação do algoritmo SMOTE, regulado para criação de 600% de instâncias sintéticas na classe de rejeição.

- Cenário 5: Conjunto de dados previamente submetidos ao algoritmo de seleção de atributos CfsSubsetEval+Best-first e então submetidos ao algoritmo SMOTE para balanceamento.

- Cenário 6: Conjunto de dados previamente submetidos ao algoritmo de seleção de atributos InfoGain e então submetidos ao algoritmo SMOTE para balanceamento.

A Tabela 4 resume os cenários:

Tabela 4 - Composição dos Cenários de Estudo.

Cenário	Algoritmo de Seleção	Algoritmo de Balanceamento
1	Nenhum	Nenhum
2	CfsSubset+BestFirst	Nenhum
3	InfoGain	Nenhum
4	SMOTE	Nenhum
5	CfsSubset+BestFirst	SMOTE
6	InfoGain	SMOTE

Fonte: (Autor, 2020)

3.3.3.3 Fase de testes

A aplicação dos algoritmos se deu através da divisão, aleatória, do conjunto de dados em dois subconjuntos. Um conjunto de treinamento, possuindo 2/3 das instâncias totais, utilizado para a construção dos modelos de classificação, e um conjunto de teste, referente ao 1/3 restante dos dados totais, utilizado para avaliar o modelo resultante através das métricas já descritas.

Os 6 algoritmos selecionados foram, então, testados nos 6 cenários propostos. Os dados foram compilados em uma tabela única para fins de comparação e visualização, conforme Figura 26. Lembrando que cada métrica possui três valores, referentes à classe positiva, à classe negativa e à média destes dois casos.

Figura 26 - Matriz de Resultados dos Algoritmos.

	A	B	C	D	E	F	G	H	I	J
1	Classe	Amostra	Test Options	Classificador	Acuracidade	TP RATE	FP RATE	Precisão	F-Measure	ROC Area
2	A	Default	Test_Set (1/3)	Naive Bayes	92,9134	0,946	0,2	0,972	0,959	0,912
3	R	Default	Test_Set (1/3)	Naive Bayes	92,9134	0,8	0,054	0,667	0,727	0,911
4	Avg.	Default	Test_Set (1/3)	Naive Bayes	92,9134	0,929	0,183	0,936	0,932	0,912
5	A	Default	Test_Set (1/3)	IBk (10)	95,2756	0,973	0,2	0,973	0,973	0,892
6	R	Default	Test_Set (1/3)	IBk (10)	95,2756	0,8	0,027	0,8	0,8	0,892
7	Avg.	Default	Test_Set (1/3)	IBk (10)	95,2756	0,953	0,18	0,953	0,953	0,892
8	A	Default	Test_Set (1/3)	AdaBoost	91,3386	0,964	0,467	0,939	0,952	0,882
9	R	Default	Test_Set (1/3)	AdaBoost	91,3386	0,533	0,036	0,667	0,593	0,882
10	Avg.	Default	Test_Set (1/3)	AdaBoost	91,3386	0,913	0,416	0,907	0,909	0,882
11	A	Default	Test_Set (1/3)	J48	93,7008	0,964	0,267	0,964	0,964	0,836
12	R	Default	Test_Set (1/3)	J48	93,7008	0,733	0,036	0,733	0,733	0,836
13	Avg.	Default	Test_Set (1/3)	J48	93,7008	0,937	0,239	0,937	0,937	0,836
14	A	Default	Test_Set (1/3)	Random Forest	95,2756	0,973	0,2	0,973	0,973	0,882
15	R	Default	Test_Set (1/3)	Random Forest	95,2756	0,8	0,027	0,8	0,8	0,882
16	Avg.	Default	Test_Set (1/3)	Random Forest	95,2756	0,953	0,18	0,953	0,953	0,882
17	A	Default	Test_Set (1/3)	SMO_Default	95,2756	0,973	0,2	0,973	0,973	0,887
18	R	Default	Test_Set (1/3)	SMO_Default	95,2756	0,8	0,027	0,8	0,8	0,887
19	Avg.	Default	Test_Set (1/3)	SMO_Default	95,2756	0,953	0,18	0,953	0,953	0,887
20	A	cfs	Test_Set (1/3)	Naive Bayes	94,4882	0,964	0,2	0,973	0,969	0,914
21	R	cfs	Test_Set (1/3)	Naive Bayes	94,4882	0,8	0,036	0,75	0,774	0,914
22	Avg.	cfs	Test_Set (1/3)	Naive Bayes	94,4882	0,945	0,181	0,947	0,946	0,914
23	A	cfs	Test_Set (1/3)	IBk (10)	95,2756	0,973	0,2	0,973	0,973	0,889
24	R	cfs	Test_Set (1/3)	IBk (10)	95,2756	0,8	0,027	0,8	0,8	0,889
25	Avg.	cfs	Test_Set (1/3)	IBk (10)	95,2756	0,953	0,18	0,953	0,953	0,889
26	A	cfs	Test_Set (1/3)	AdaBoost	94,063	0,982	0,2	0,973	0,978	0,861
27	R	cfs	Test_Set (1/3)	AdaBoost	94,063	0,8	0,018	0,857	0,828	0,861
28	Avg.	cfs	Test_Set (1/3)	AdaBoost	94,063	0,961	0,178	0,96	0,96	0,861
29	A	cfs	Test_Set (1/3)	J48	95,2756	0,973	0,2	0,973	0,973	0,872

Fonte: (Autor, 2020)

3.4 Análise comparativa dos resultados

Inicialmente, os seis cenários foram comparados de maneira geral, sem atentar-se aos algoritmos de maneira independente. Para realizar essa análise foi utilizada a ferramenta de formatação condicional do Excel, programada para definir através do gradiente de cores os melhores resultados para cada métrica. Convencionou-se com a cor vermelha os melhores resultados e amarelo para os piores resultados. A ferramenta colore, gradualmente, as células a fim de facilitar a interpretação visual dos dados que foram divididos em 6 cenários e três classes de métricas, as métricas que dizem respeito à classificação de instâncias positivas, negativas e a média de ambas. Observe que os cenários 1, 2 e 3 carecem de balanceamento, enquanto que os cenários 4, 5 e 6 foram balanceados através de adições de instâncias sintéticas via algoritmo SMOTE.

Observe a Tabela 5.

Tabela 5 - Cenários e suas métricas formatados condicionalmente.

Cenário / Métricas		TP RATE	FP RATE	Precisão	F-Measure	ROC Area	TP RATE	FP RATE	Precisão	F-Measure	ROC Area	TP RATE	FP RATE	Precisão	F-Measure	ROC Area
DESBALANCEADO	1	0,8	0,054	0,667	0,727	0,911	0,946	0,2	0,972	0,959	0,912	0,929	0,183	0,936	0,932	0,912
		0,8	0,027	0,8	0,8	0,892	0,973	0,2	0,973	0,973	0,892	0,953	0,18	0,953	0,953	0,892
		0,533	0,036	0,667	0,593	0,882	0,964	0,467	0,939	0,952	0,882	0,913	0,416	0,907	0,909	0,882
		0,733	0,036	0,733	0,733	0,836	0,964	0,267	0,964	0,964	0,836	0,937	0,239	0,937	0,937	0,836
		0,8	0,027	0,8	0,8	0,882	0,973	0,2	0,973	0,973	0,882	0,953	0,18	0,953	0,953	0,882
		0,8	0,027	0,8	0,8	0,887	0,973	0,2	0,973	0,973	0,887	0,953	0,18	0,953	0,953	0,887
	2	0,8	0,036	0,75	0,774	0,914	0,964	0,2	0,973	0,969	0,914	0,945	0,181	0,947	0,946	0,914
		0,8	0,027	0,8	0,8	0,889	0,973	0,2	0,973	0,973	0,889	0,953	0,18	0,953	0,953	0,889
		0,8	0,018	0,857	0,828	0,861	0,982	0,2	0,973	0,978	0,861	0,961	0,178	0,96	0,96	0,861
		0,8	0,027	0,8	0,8	0,872	0,973	0,2	0,973	0,973	0,872	0,953	0,18	0,953	0,953	0,872
		0,8	0,027	0,8	0,8	0,92	0,973	0,2	0,973	0,976	0,92	0,953	0,18	0,953	0,953	0,92
		0,8	0,036	0,75	0,774	0,882	0,964	0,2	0,973	0,969	0,882	0,945	0,181	0,947	0,946	0,882
	3	0,8	0,027	0,8	0,8	0,928	0,973	0,2	0,973	0,973	0,928	0,953	0,18	0,953	0,953	0,928
		0,8	0,027	0,8	0,8	0,892	0,973	0,2	0,973	0,973	0,892	0,953	0,18	0,953	0,953	0,892
		0,8	0,027	0,8	0,8	0,867	0,973	0,2	0,973	0,973	0,867	0,953	0,18	0,953	0,953	0,867
		0,733	0,027	0,786	0,759	0,808	0,973	0,267	0,965	0,969	0,808	0,945	0,238	0,943	0,944	0,808
		0,8	0,027	0,8	0,8	0,919	0,973	0,2	0,973	0,973	0,919	0,953	0,18	0,953	0,953	0,919
		0,8	0,036	0,75	0,774	0,882	0,964	0,2	0,973	0,969	0,882	0,945	0,181	0,947	0,946	0,882
SMOTE	4	0,972	0,063	0,937	0,954	0,956	0,938	0,028	0,972	0,955	0,978	0,954	0,045	0,955	0,954	0,967
		0,972	0,348	0,727	0,832	0,957	0,952	0,028	0,961	0,777	0,957	0,808	0,184	0,847	0,804	0,957
		0,944	0,045	0,953	0,948	0,975	0,955	0,056	0,947	0,951	0,975	0,95	0,05	0,95	0,95	0,975
		0,907	0,089	0,907	0,907	0,92	0,911	0,093	0,911	0,911	0,92	0,909	0,091	0,909	0,909	0,92
		0,935	0,027	0,971	0,952	0,991	0,973	0,065	0,94	0,956	0,991	0,954	0,047	0,955	0,954	0,991
		0,935	0,054	0,943	0,939	0,941	0,946	0,065	0,938	0,942	0,941	0,941	0,06	0,941	0,941	0,941
	5	0,85	0,036	0,958	0,901	0,951	0,964	0,15	0,871	0,915	0,951	0,909	0,094	0,913	0,908	0,951
		0,85	0,107	0,883	0,867	0,905	0,893	0,15	0,862	0,877	0,905	0,872	0,129	0,873	0,872	0,905
		0,963	0,098	0,904	0,932	0,968	0,902	0,037	0,962	0,931	0,968	0,932	0,067	0,933	0,931	0,968
		0,916	0,089	0,907	0,912	0,891	0,911	0,084	0,919	0,911	0,891	0,913	0,087	0,913	0,913	0,891
		0,944	0,027	0,971	0,957	0,987	0,973	0,056	0,948	0,96	0,987	0,959	0,042	0,959	0,959	0,987
	6	0,822	0,036	0,957	0,884	0,893	0,964	0,178	0,85	0,904	0,893	0,895	0,108	0,902	0,894	0,893
		0,85	0,027	0,968	0,905	0,952	0,973	0,15	0,872	0,92	0,958	0,913	0,09	0,919	0,913	0,955
		0,897	0,089	0,906	0,901	0,946	0,911	0,103	0,903	0,907	0,946	0,904	0,096	0,904	0,904	0,946
		0,935	0,036	0,962	0,948	0,983	0,964	0,065	0,939	0,952	0,983	0,95	0,051	0,95	0,95	0,983
		0,944	0,071	0,927	0,935	0,944	0,929	0,056	0,945	0,937	0,944	0,936	0,064	0,936	0,936	0,944
		0,935	0,027	0,971	0,952	0,985	0,973	0,065	0,94	0,956	0,985	0,954	0,047	0,955	0,954	0,985
		0,841	0,063	0,928	0,882	0,889	0,938	0,159	0,861	0,897	0,889	0,89	0,112	0,893	0,89	0,889
	CENÁRIOS	Classificações Positivas					Classificações Negativas					Classificação Média				

Fonte: (Autor, 2020)

Pode-se observar através da tabela 5 que os cenários que apresentaram melhores resultados, independente dos algoritmos utilizados, foram os cenários que foram balanceados via SMOTE. É possível observar que a cor vermelha se concentra mais nas células da parte inferior da tabela, sendo majoritária nos cenários 4, 5 e 6. Como o objetivo é desenvolver um modelo capaz de classificar instâncias positivas da melhor forma possível, as análises se concentram nos resultados das métricas para a classificação de instâncias positivas (rejeição). E, como critério secundário, foram avaliadas as métricas referentes à classificação das instâncias negativas. Vale ressaltar, que além dos artifícios visuais, os dados foram comparados através da seleção dos 10 melhores valores das métricas, considerando apenas instâncias positivas, aquelas que mais interessam. Observe a Tabela 6, em que os 10 melhores valores para cada métrica são hachurados em vermelho. Mais uma vez fica evidente que algoritmos que foram tratados por meio da criação de instâncias sintéticas via SMOTE (cenários 4,5 e 6) são os que têm melhor performance.

Tabela 6 - 10 Melhores Valores Para Cada Métrica.

Classe	Amostra	Cenário	Classificador	TP RATE	FP RATE	Precisão	F-Measure	ROC Area
Positiva	SMOTE	4	Naive Bayes	0,972	0,063	0,937	0,954	0,956
Positiva	SMOTE	4	kNN (10)	0,972	0,348	0,727	0,832	0,957
Positiva	cfs+smote	5	AdaBoost	0,963	0,098	0,904	0,932	0,968
Positiva	cfs+smote	5	Random Forest	0,944	0,027	0,971	0,957	0,987
Positiva	SMOTE	4	AdaBoost	0,944	0,045	0,953	0,948	0,975
Positiva	InfoGain+Smote	6	J48	0,944	0,071	0,927	0,935	0,944
Positiva	SMOTE	4	Random Forest	0,935	0,027	0,971	0,952	0,991
Positiva	InfoGain+Smote	6	Random Forest	0,935	0,027	0,971	0,952	0,985
Positiva	InfoGain+Smote	6	AdaBoost	0,935	0,036	0,962	0,948	0,983
Positiva	SMOTE	4	SMO_Default	0,935	0,054	0,943	0,939	0,941
Positiva	cfs+smote	5	J48	0,916	0,089	0,907	0,912	0,891
Positiva	SMOTE	4	J48	0,907	0,089	0,907	0,907	0,92
Positiva	InfoGain+Smote	6	kNN (10)	0,897	0,089	0,906	0,901	0,946
Positiva	InfoGain+Smote	6	Naive Bayes	0,85	0,027	0,968	0,905	0,952
Positiva	cfs+smote	5	Naive Bayes	0,85	0,036	0,958	0,901	0,951
Positiva	cfs+smote	5	kNN (10)	0,85	0,107	0,883	0,867	0,905
Positiva	InfoGain+Smote	6	SMO_Default	0,841	0,063	0,928	0,882	0,889
Positiva	cfs+smote	5	SMO_Default	0,822	0,036	0,957	0,884	0,893
Positiva	Default	1	Naive Bayes	0,8	0,054	0,667	0,727	0,911
Positiva	Default	1	kNN (10)	0,8	0,027	0,8	0,8	0,892
Positiva	Default	1	Random Forest	0,8	0,027	0,8	0,8	0,882
Positiva	Default	1	SMO_Default	0,8	0,027	0,8	0,8	0,887
Positiva	cfs	2	kNN (10)	0,8	0,027	0,8	0,8	0,889
Positiva	cfs	2	J48	0,8	0,027	0,8	0,8	0,872
Positiva	cfs	2	Random Forest	0,8	0,027	0,8	0,8	0,92
Positiva	InfoGain	3	Naive Bayes	0,8	0,027	0,8	0,8	0,928
Positiva	InfoGain	3	kNN (10)	0,8	0,027	0,8	0,8	0,892
Positiva	InfoGain	3	AdaBoost	0,8	0,027	0,8	0,8	0,867
Positiva	InfoGain	3	Random Forest	0,8	0,027	0,8	0,8	0,919
Positiva	cfs	2	Naive Bayes	0,8	0,036	0,75	0,774	0,914
Positiva	cfs	2	SMO_Default	0,8	0,036	0,75	0,774	0,882
Positiva	InfoGain	3	SMO_Default	0,8	0,036	0,75	0,774	0,882
Positiva	cfs	2	AdaBoost	0,8	0,018	0,857	0,828	0,861
Positiva	InfoGain	3	J48	0,733	0,027	0,786	0,759	0,808
Positiva	Default	1	J48	0,733	0,036	0,733	0,733	0,836
Positiva	Default	1	AdaBoost	0,533	0,036	0,667	0,593	0,882

Fonte: (Autor, 2020)

A Tabela 7 resume os resultados das métricas de cada algoritmo para instâncias positivas, para os cenários 4, 5 e 6.

Tabela 7 - Resultados Para Classificação de Instancias Positivas.

Métricas	Algorítimo	TP RATE (+)	FP RATE (+)	Precisão (+)	F-Measure (+)	ROC Area (+)
4	Naive Bayes	0,972	0,063	0,937	0,954	0,956
	kNN (10)	0,972	0,348	0,727	0,832	0,957
	AdaBoost	0,944	0,045	0,953	0,948	0,975
	J48	0,907	0,089	0,907	0,907	0,92
	Random Forest	0,935	0,027	0,971	0,952	0,991
	SMO_Default	0,935	0,054	0,943	0,939	0,941
5	Naive Bayes	0,85	0,036	0,958	0,901	0,951
	kNN (10)	0,85	0,107	0,883	0,867	0,905
	AdaBoost	0,963	0,098	0,904	0,932	0,968
	J48	0,916	0,089	0,907	0,912	0,891
	Random Forest	0,944	0,027	0,971	0,957	0,987
	SMO_Default	0,822	0,036	0,957	0,884	0,893
6	Naive Bayes	0,85	0,027	0,968	0,905	0,952
	kNN (10)	0,897	0,089	0,906	0,901	0,946
	AdaBoost	0,935	0,036	0,962	0,948	0,983
	J48	0,944	0,071	0,927	0,935	0,944
	Random Forest	0,935	0,027	0,971	0,952	0,985
	SMO_Default	0,841	0,063	0,928	0,882	0,889
CENÁRIOS	Instancias POSITIVAS					

Fonte: (Autor, 2020)

Para determinar qual a melhor combinação de algoritmos e cenário para utilização no problema, foi definida a seguinte metodologia. Primeiro, a performance de cada algoritmo será comparada nos três cenários (4, 5 e 6). Uma vez selecionado a melhor combinação para cada algoritmo, são então postas em comparação as seis combinações entre si.

As métricas de cada algoritmo foram comparadas nos três cenários selecionados. Apenas uma aplicação dentro dos três cenários foi escolhida para ser confrontada com a melhor combinação dos demais algoritmos. Para determinar quais algoritmos performam melhor, recorreu-se à utilização de tabelas comparativas e gráficos estilo “radar” para comparar as 4 métricas principais. Também foram definidos graus de relevância para cada métrica relacionada à classificação de instâncias positivas, sendo da mais importante para a menos importante, no contexto analisado: Taxa TP, área da curva ROC, Precisão e Taxa FP.

3.4.1 Análise comparativa da performance para cada algoritmo de classificação

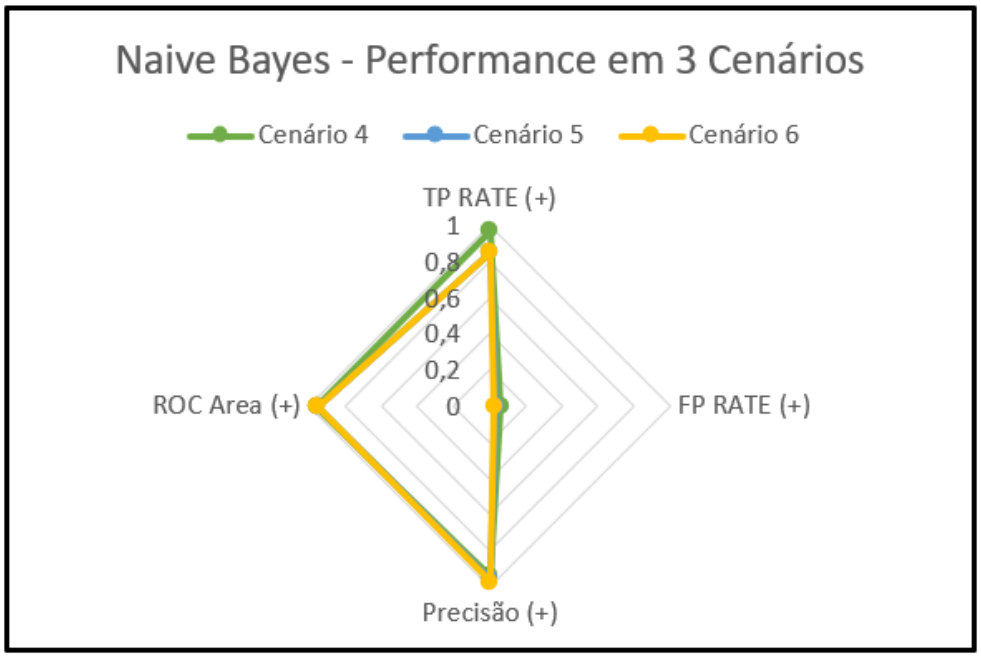
São comparados, nesse momento, as métricas de classificações de instâncias positivas para cada um dos 6 algoritmos em 3 cenários, de maneira individual.

3.4.1.1 Naïve Bayes

Primeiro são analisadas as aplicações do algoritmo Naive Bayes. Observe a Figura 27 que explora a performance do algoritmo nos três cenários selecionados através da tabela e gráfico apresentados. É possível observar que o algoritmo performa melhor no cenário 4 uma vez que essa combinação só tem as métricas FP Rate (+) e Precisão (+) dominadas por outros cenários, possuindo as duas métricas mais relevantes como dominantes. Portanto, a combinação Naive Bayes no cenário 4 é confrontada na segunda etapa dessa análise.

Figura 27 - Comparativo de Métricas para o Algoritmo Naive Bayes em 3 cenários.

Naive Bayes				
Cenário	TP RATE (+)	FP RATE (+)	Precisão (+)	ROC Area (+)
Cenário 4	0,972	0,063	0,937	0,956
Cenário 5	0,85	0,036	0,958	0,951
Cenário 6	0,85	0,027	0,968	0,952



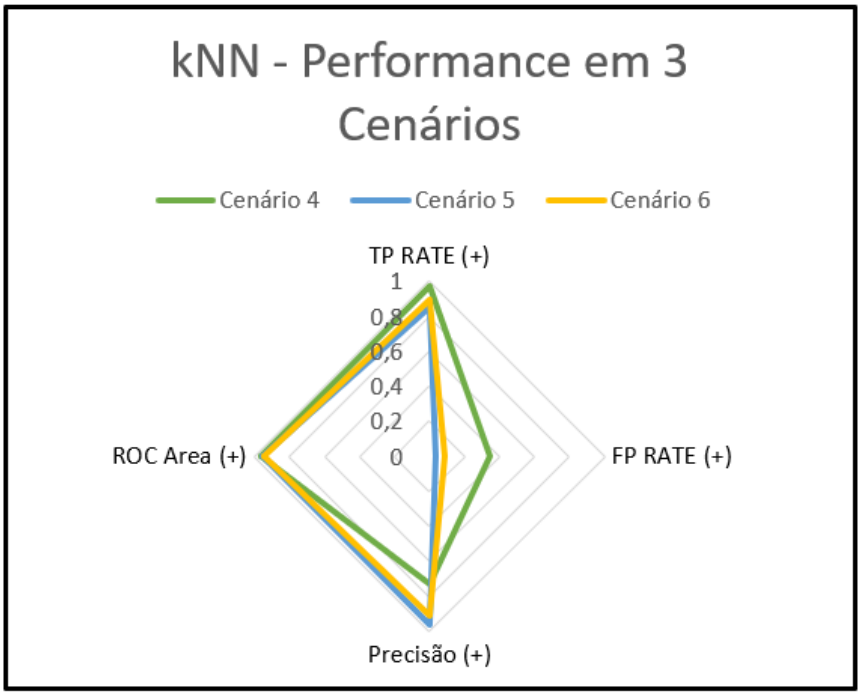
Fonte: (Autor, 2020)

3.4.1.2 kNN

Considerando que dentre as métricas analisadas, a TP para a classe de rejeição é a mais relevante, conclui-se que o cenário em que o algoritmo kNN, k=10, melhor performa é o cenário de número 4, como pode ser observado na Figura 28.

Figura 28 - Comparativo de Métricas para o Algoritmo kNN em 3 cenários.

kNN (10)				
Cenário	TP RATE (+)	FP RATE (+)	Precisão (+)	ROC Area (+)
Cenário 4	0,972	0,348	0,727	0,957
Cenário 5	0,85	0,036	0,958	0,951
Cenário 6	0,897	0,089	0,906	0,946



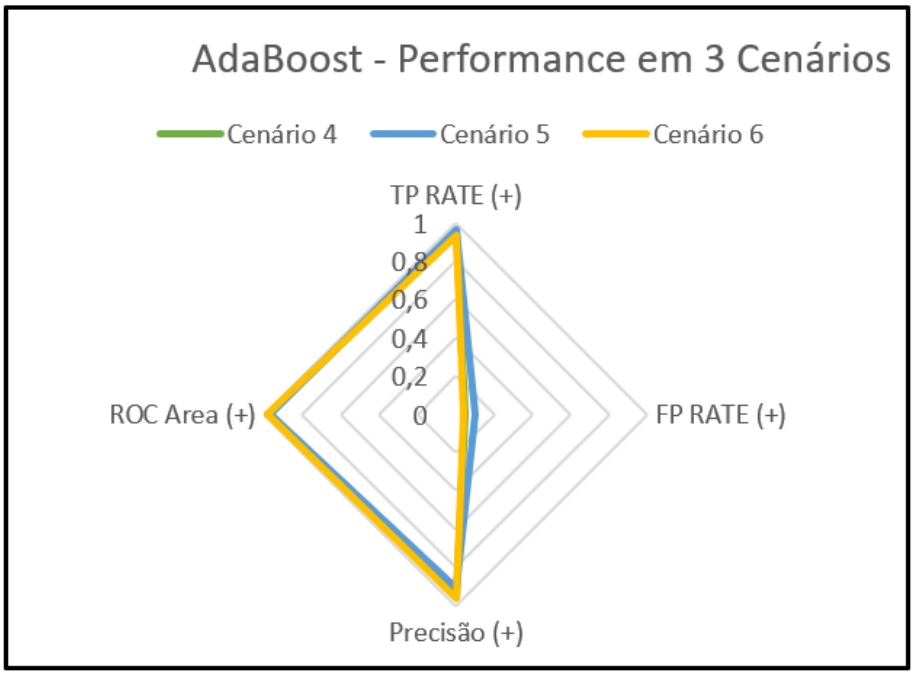
Fonte: (Autor, 2020)

3.4.1.3 Adaboost

A análise do Algoritmo Adaboost pode ser resumida na Figura 29. O cenário 6 mostrou dominar os demais cenários em todas as métricas, exceto na TP Rate para rejeição, em que se mostrou um pouco menos eficiente que os demais cenários. Analisando as demais métricas e, em especial os valores da curva ROC, chegou-se à conclusão de que o melhor cenário foi o de número 6. Observe a Figura 29.

Figura 29 - Comparativo de Métricas para o Algoritmo Adaboost em 3 cenários.

AdaBoost				
Cenário	TP RATE (+)	FP RATE (+)	Precisão (+)	ROC Area (+)
Cenário 4	0,944	0,045	0,953	0,975
Cenário 5	0,963	0,098	0,904	0,968
Cenário 6	0,935	0,036	0,962	0,983



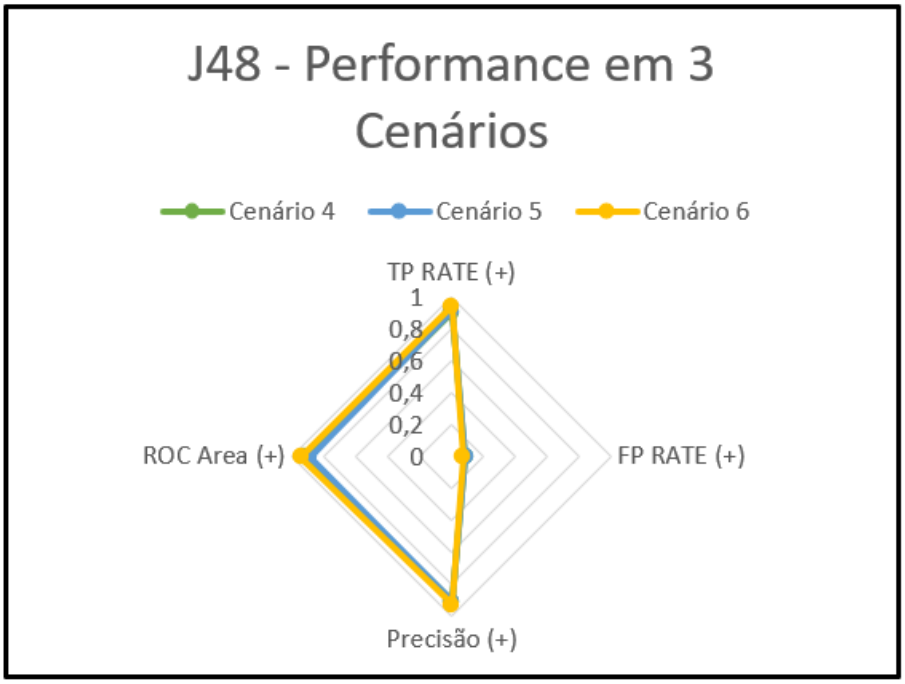
Fonte: (Autor, 2020)

3.4.1.4 J48

A próxima análise é do algoritmo J48, em que a melhor performance ocorre também no cenário 6, como é possível observar na Figura 30. Essa combinação será abordada, de forma mais profunda nas conclusões desse trabalho uma vez que por ser um algoritmo de árvore de decisão ele fornece um "mapa" para classificação de instâncias, sendo as relações encontradas por esse algoritmo muito valiosas para a compreensão de quais atributos são relevantes para a rejeição de motores, e como analisá-los.

Figura 30 - Comparativo de Métricas para o Algoritmo J48 em 3 cenários.

J48				
Cenário	TP RATE (+)	FP RATE (+)	Precisão (+)	ROC Area (+)
Cenário 4	0,907	0,089	0,907	0,92
Cenário 5	0,916	0,089	0,907	0,891
Cenário 6	0,944	0,071	0,927	0,944



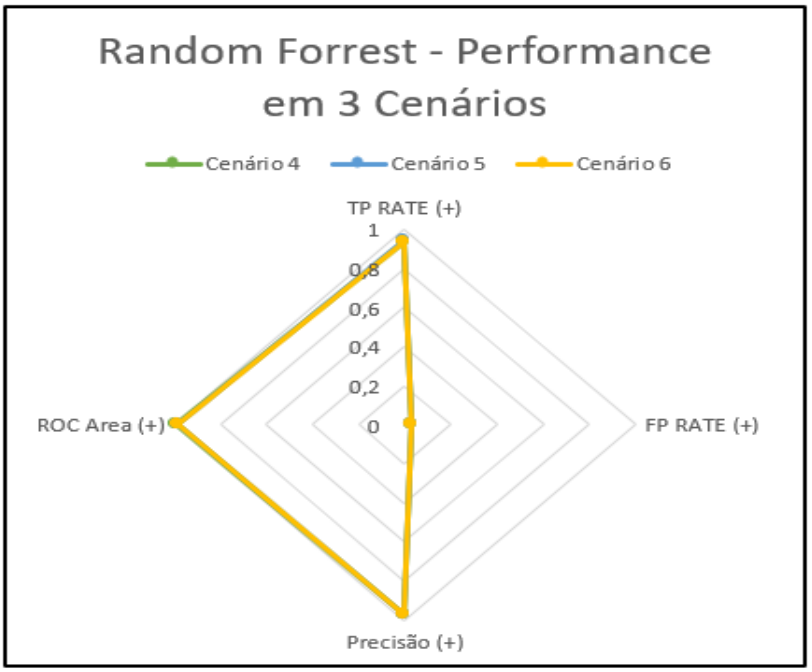
Fonte: (Autor, 2020)

3.4.1.5 Random Forest

Os resultados do algoritmo Randon Forest podem ser observados na Figura 31, em que as melhores métricas ocorrem no cenário 5, que só é superado em uma métrica, a ROC área, do cenário 4.

Figura 31 - Comparativo de Métricas para o Algoritmo Random Forest em 3 cenários.

Random Forrest				
Cenário	TP RATE (+)	FP RATE (+)	Precisão (+)	ROC Area (+)
Cenário 4	0,935	0,027	0,971	0,991
Cenário 5	0,944	0,027	0,971	0,987
Cenário 6	0,935	0,027	0,971	0,985



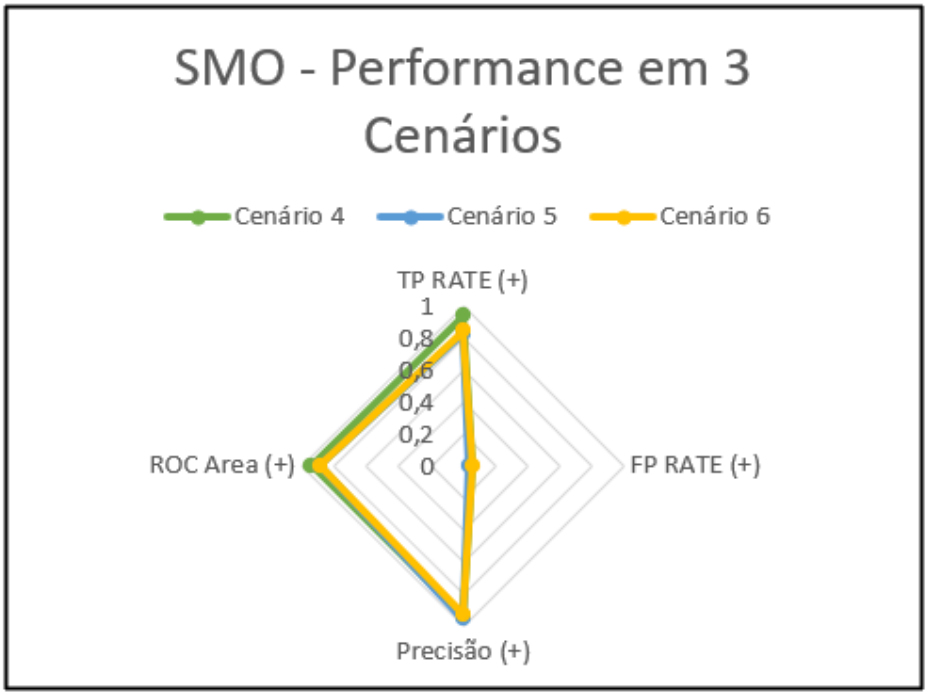
Fonte: (Autor, 2020)

3.4.1.6 SMO

Finalmente, os resultados do último algoritmo, SMO, podem ser encontrados na Figura 32, em que as melhores métricas ocorrem no cenário 4.

Figura 32 - Comparativo de Métricas para o Algoritmo Random Forest em 3 cenários.

SMO				
Cenário	TP RATE (+)	FP RATE (+)	Precisão (+)	ROC Area (+)
Cenário 4	0,935	0,054	0,943	0,941
Cenário 5	0,822	0,036	0,957	0,893
Cenário 6	0,841	0,063	0,928	0,889



Fonte: (Autor, 2020)

3.4.2 Etapa final de comparação

Uma vez selecionados para cada algoritmo qual combinação de cenário gera os melhores resultados, todos os dados foram compilados na Tabela 8 para que se pudesse explorar as comparações entre os mesmos e definir qual será o mais indicado para ser aplicado na situação problema.

Tabela 8 - Comparação Final Entre Combinação do Melhor Cenário Para Cada Algoritmo.

Algoritimo_Cenário	TP RATE (+)	FP RATE (+)	Precisão (+)	ROC Area (+)
Naive Bayes_C4	0,972	0,063	0,937	0,956
Knn_C4	0,972	0,348	0,727	0,957
AdaBoost_C6	0,935	0,036	0,962	0,983
J48_C6	0,944	0,071	0,927	0,944
Random Forrest _C5	0,944	0,027	0,971	0,987
SMO_C4	0,935	0,054	0,943	0,941

Fonte: (Autor, 2020)

Na Tabela 8, os dados hachurados em vermelho correspondem aos melhores valores para cada métrica analisada. Pode-se, então, descartar os algoritmos que são completamente dominados pelos demais, isso é, que nenhuma de suas métricas é a melhor em termos comparativos com as demais. Excluem-se então 4 combinações. O algoritmo kNN associado ao cenário 4, o algoritmo Adaboost associado ao cenário 6, o algoritmo J48 associado ao cenário 6 e o algoritmo SMO aplicado ao cenário de número 4.

As duas combinações mais relevantes são, portanto, as apresentadas na Tabela 9, em que agora são confrontadas as métricas referentes à média de performances para ambas as classes, isso é, levando-se em conta como a combinação performa na média, para classificar classes positivas e negativas.

Tabela 9 - Tabela Comparativa NaiveBayes em cenário 4 e Random Forest em Cenário 5.

Algoritimo_Cenário	TP RATE	FP RATE	Precisão	ROC Area
Naive Bayes_C4	0,954	0,045	0,955	0,967
Random Forrest _C5	0,959	0,042	0,959	0,987

Fonte: (Autor, 2020)

A aplicação do algoritmo Random Forest após a utilização do seletor de atributos CfsSubsetEval+BestFirst e balanceamento via criação de instancias sintéticas com SMOTE (cenário 5) se mostrou superior às demais combinações, apresentando a melhor performance a partir da metodologia utilizada.

Observe a Figura 33, retirada do software Weka referente aos resultados dessa combinação. Nela, é possível observar as métricas comparadas e a matriz de confusão da aplicação. É possível ver que 109 dos 112 casos de motores Aceitos são classificados como aceitos (3 Falsos Negativos) e 101 dos 107 motores Rejeitados são classificados como rejeitados (6 Falsos Positivos).

Figura 33 - Resultados do Modelo de Predição Random Forest + CfsSubset + SMOTE.

```
Correctly Classified Instances         210               95.8904 %
Incorrectly Classified Instances         9                4.1096 %
Kappa statistic                          0.9177
Mean absolute error                      0.1102
Root mean squared error                  0.2041
Relative absolute error                 22.0469 %
Root relative squared error             40.8227 %
Total Number of Instances              219

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,973    0,056    0,948      0,973    0,960      0,918    0,987     0,987     A
                 0,944    0,027    0,971      0,944    0,957      0,918    0,987     0,987     R
Weighted Avg.    0,959    0,042    0,959      0,959    0,959      0,918    0,987     0,987

=== Confusion Matrix ===

   a   b   <-- classified as
 109   3 |   a = A
   6 101 |   b = R
```

Fonte: (Autor, 2020)

Vale ressaltar que apesar de não ter sido o algoritmo que melhor performou, o J48 combinado ao cenário 6 deve também ser melhor investigado. Isso porque, por se tratar de um algoritmo do tipo “arvore de decisão”, é possível extrair relações claras e relevantes da árvore utilizada pelo algoritmo durante o processo de classificação. Observe a Figura 34 com os resultados da aplicação do algoritmo J48, extraídas do Weka, quando aplicado à amostra pré-processada pelo seletor de atributos InfoGain e balanceado pelo algoritmo SMOTE.

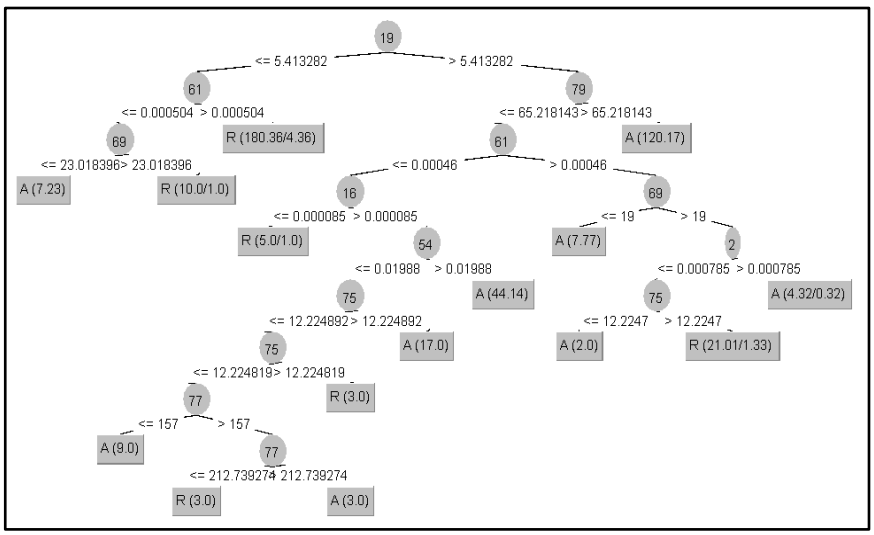
Figura 34 - Resultados do Modelo de Predição InfoGain + SMOTE + J48.

```
Correctly Classified Instances         205               93.6073 %
Incorrectly Classified Instances        14                6.3927 %
Kappa statistic                          0.8721
Mean absolute error                      0.0806
Root mean squared error                  0.2489
Relative absolute error                 16.1167 %
Root relative squared error             49.7949 %
Total Number of Instances              219

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,929    0,056    0,945      0,929    0,937      0,872    0,944     0,925     A
                 0,944    0,071    0,927      0,944    0,935      0,872    0,944     0,923     R
Weighted Avg.    0,936    0,064    0,936      0,936    0,936      0,872    0,944     0,924

=== Confusion Matrix ===

   a   b   <-- classified as
 104   8 |   a = A
   6 101 |   b = R
```

Fonte: (Autor, 2020)

A Figura 33 mostra a performance global do algoritmo, isso é, tanto para a classificação de instâncias positivas, quanto negativas. Sendo possível observar que o algoritmo classifica de maneira mais precisa as instâncias positivas do que as negativas, possuindo uma taxa TP para instâncias positivas de 0,944. Ou seja, 94,4% das instâncias classificadas como positivas são, de fato, positivas. Portanto, pode-se confiar com alto grau de precisão nas relações geradas pela árvore de decisão, em especial as que levam à classificação de instâncias positivas. Observe a Figura 35, extraída do Weka, que diz respeito à árvore de decisão utilizada pelo modelo elaborado.

Figura 35 - Árvore de Decisão do Modelo de Predição InfoGain + SMOTE + J48.



Fonte: (Autor, 2020)

A árvore apresentada é relativamente simples dada a complexidade do problema, sendo as relações sugeridas de extrema relevância. Foram 9 nós, isso é, atributos sugeridos para compreender a natureza de uma instância analisada. Os atributos selecionados foram os seguintes: 2, 16, 19, 54, 61, 69, 75, 77 e 79. As regras propostas pelo algoritmo também foram analisadas e a grande maioria das relações já são de conhecimento da engenharia. No entanto, as relações referentes aos atributos 2,19 e 61 ainda são pouco conhecidas e poderão ser melhor exploradas pela empresa. Dada as questões de confidencialidade, não será possível um aprofundamento quanto a essas relações.

4 CONCLUSÃO

O presente trabalho se sustentou por uma revisão de literatura que garantisse conhecimentos gerais e específicos para aplicação de análise de dados no cenário estudado. Foram abordados temas de ordem mecânica, no que tange ao funcionamento de motores aeronáuticos a jato e a natureza de fenômenos vibratórios, bem como temas referentes à mineração de dados. Uma vez concluídos os estudos necessários para compreensão e aplicação das técnicas de análise de dados e, em especial, de mineração de dados, sucedeu-se o estudo de um caso prático associado a montagem de motores.

No cenário de estudo, foram aplicadas diversas técnicas de mineração de dados, em especial, a utilização de algoritmos de aprendizado supervisionado para elaborar um modelo suficientemente preciso para classificar instâncias no cenário estudado. O objetivo geral foi explorar a aplicação de mineração em um banco de dados complexo, série histórica de dados de montagem e vibração em uma linha de *turbofans*, a fim de encontrar novas relações entre montagem e vibração e teste, assim como desenvolver um modelo de predição para classificação de motores. Dessa forma, a empresa terá a seu dispor um modelo capaz de identificar se um motor pós montagem deve ou não ser enviado para célula de testes.

Os resultados obtidos das aplicações são considerados bastante positivos, tanto do ponto de vista metodológico quanto prático. Do ponto de vista metodológico, é possível observar que a revisão de literatura estudada se mostrou parte essencial para o desenvolvimento das análises. Sem um modelo metodológico como o KDD, a condução desse estudo seria impraticável, uma vez que foram encontrados uma série de dificuldades e obstáculo que só foram superados pelas ferramentas propostas pelo modelo KDD.

Do ponto de vista prático, os resultados foram bastante animadores. Foram encontradas novas relações, até então desconhecidas, pela engenharia acerca de influências

vibratórias. Certos atributos mostraram-se fortemente correlacionados à vibração dos *turbofans* em teste. Essas relações, a dispor da empresa estudada, irão se traduzir em alterações de processos internos e melhor controle dos atributos identificados.

O desenvolvimento de um modelo de classificação/predição dos resultados de vibração é, talvez, a maior contribuição prática desse trabalho. Com o modelo proposto a empresa terá uma nova ferramenta disponível.

O modelo é capaz de prever com suficiente precisão (93,6%) se um motor será aprovado ou rejeitado antes do processo de teste. Caso o motor em questão seja classificado como rejeitado pelo algoritmo, o que ocorre com uma taxa de precisão de 97,1%, a produção poderá optar por não enviar o motor para teste até que parâmetros de montagem sejam alterados. Essa nova possibilidade irá se traduzir em uma grande economia para empresa. Apesar da alta taxa de precisão, há ainda casos de falsos positivos e falsos negativos. Para mitigar essas situações, em especial falsos positivos, isso é, motores classificados pelo modelo como rejeitados, mas que seriam aprovados, recomenda-se que a empresa desenvolva um segundo processo intermediário de testes que possa confirmar, sem muitos gastos, se as assinaturas de vibração realmente estão elevadas.

Quanto a falsos negativos, isso é, motores classificados pelo modelo proposto como aprovados que, no entanto, serão rejeitados em fase de teste, recomenda-se que o modelo seja revisado e aprimorado à medida que mais dados estejam disponíveis. Por se tratar de um modelo de aprendizado supervisionado, quanto mais dados disponíveis, melhor será o aprendizado de máquina. Como o modelo proposto é uma ferramenta que se baseia em dados passados, haverá ganhos incrementais de precisão à medida que mais motores forem testados. Por isso, é fundamental que os novos dados sejam registrados e incorporados ao modelo.

5 LIMITAÇÕES DA PESQUISA E TRABALHOS FUTUROS

Nessa secção serão abordadas as principais ressalvas quanto ao caso prático estudados. Em especial as limitações impostas a esse estudo e possíveis trabalhos futuros.

5.1 Limitações da pesquisa

Devem ser consideradas as restrições e limitações que foram adotas e submetidas à pesquisa de campo desenvolvida por esse trabalho. Essas limitações e restrições foram ambas de ordem compulsória, legal, e de fim metodológicos.

As limitações de ordem legal se impuseram desde o início do trabalho uma vez que a pesquisa está sendo realizada em uma empresa multinacional, cuja natureza de dados é extremamente sensível, incorrendo sobre questões de propriedade industrial e, por isso, as relações encontradas e estudadas não puderam ser reveladas.

Quanto às limitações de ordem metodológica, algumas decisões e restrições tiveram de ser tomadas a fim de garantir que um modelo suficientemente bom para o estudo fosse desenvolvido dada as limitações de tempo e disponibilidade de processamento. E, por isso, apenas uma linha de motores foi estudada podendo haver novas descobertas caso se considere um conjunto de análise maior. Ainda é importante ressaltar a escassez de instâncias disponíveis, pois para a obtenção de modelos mais robustos necessita-se de mais dados.

5.2 Trabalhos futuros

Como trabalhos futuros, algumas possibilidades e temas já podem ser vislumbrados, entre eles destaca-se a possibilidade de aumentar o conjunto de dados analisados, isso é, levar em consideração mais linhas de motores montados assim como outras células de montagem. Tendo mais dados à disposição, novas tendências e descobertas poderão, possivelmente, ser discutidas e um novo modelo ser sugerido.

Há também, como possibilidade de trabalho futuro, a exploração das relações encontradas entre atributos e vibrações, por meio de uma perspectiva de mecânica de sistemas. Isso é, compreender se as relações encontradas pela análise de dados se sustentam pelo conhecimento já existente na literatura e, caso não se sustentem, promover uma investigação a fundo para melhor compreender suas causas.

REFERÊNCIAS

ASLIN, Matthew J.; PATTON, Gary J. ***Central maintenance computer system and fault data handling method: Patent n. 4,943,919***. California, 1990.

CASANOVA, A. A.; LABIDI, S. **Algoritmo da Confiança Inversa para Mineração de Dados Baseado em Técnicas de Regras de Associação e Lógica Nebulosa**. XXV Congresso da Sociedade Brasileira de Computação. Cuiabá, 2005.

CHAPMAN, P; CLINTON, J; KERBER, R; KHABAZA, T; REINARTZ, T; SHEARER, C; WIRTH, R. ***CRISP-DM 1.0: CRISP-DM consortium***. Delaware, 2000.

CHAWLA, N. V.; LAZAREVIC, A.; HALL, L. O.; BOWYER, K. W. ***Smoteboost: Improving prediction of the minority class in boosting. In Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases.*** Croatia, 2003.

CLARENCE, W. de Silva. ***Vibration Fundamentals and Practice*****.** Boca Raton, 2000.

DE AMO, Sandra. **Técnicas de mineração de dados: Jornada de Atualizaçao em Informatica**. Minas Gerais, 2004.

DE RAEDT, L. ***Attribute-value learning versus inductive logic programming - The missing links: Proceedings of the Eighth International Conference on Inductive Logic Programming.*** Berlin, 1998.

DOMINGOS, Pedro. ***Metacost: A general method for making classifiers cost-sensitive. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining.*** Lisboa, 1999.

FAN, W.; STOLFO, S. J.; ZHANG, J.; CHAN, P. K. ***Misclassification cost-sensitive boosting. In Proceedings of the Sixteenth International Conference on Machine Learning***. New Jersey, 1999.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. ***From Data Mining to Knowledge Discovery in Databases.*** *American Association for Artificial Intelligence*. California, 1996.

FERNANDES, Salomão Marques da Silva Duarte. Procedimentos de manutenção na indústria aeronáutica: Tese de Doutorado, Instituto Superior de Engenharia de Lisboa. Lisboa, 2019.

FERREIRA, Marcos José Barbieri et al. **Dinâmica da inovação e mudanças estruturais: um estudo de caso da indústria aeronáutica mundial e a inserção brasileira**. 2009.

*FUNDAMENTALS OF AIRCRAFT TURBINE ENGINE CONTROL. **National Aeronautics and Space Administration***, 2020. Disponível em: <https://www.grc.nasa.gov/WWW/cdtb/aboutus/Fundamentals_of_Engine_Control.pdf>. Acesso em: 1 de Agosto de 2020.

*GE AVIATION, **"CF34-8E Fan Vibration Podcast"**, 2018.*

GE AVIATION'S GENX ENGINES POWER QANTAS 787-9 RECORD-BREAKING NON-TOP FLIGHT FROM NEW YORK TO SYDNEY. ***General Electric***, 2019. Disponível em: <https://www.geaviation.com/press-release/genx-engine-family/ge-aviations-genx-engines-power-qantas-787-9-record-breaking-non>. Acesso em: 2 de Jul de 2020.

GONZÁLEZ, P.; MENASALVAS, E.; RUIZ, S. M. C.; SEGOVIA, J. ***Towards a methodology for data mining project development: The importance of abstraction***. Berlin, 2008.
GRZYMALA-BUSSE, J. W. ***MLEM2 rule induction algorithms: With and with-out merging intervals***. Berlin, 2008.

GRZYMALA-BUSSE JW, STEFANOWSKI J, WILK S. ***A comparison of two approaches to data mining from imbalanced data***. (2005)

HAN, J.; KAMBER, M.; PEI, J. ***Data Mining: Concepts and Techniques***. Massachusetts, 2011.

HAN, J.; KAMBER, M.; PEI, J. ***Data Mining: Concepts and Techniques: The Morgan Kaufmann Series in Data management Systems***. Massachusetts, 2012.

HART, P.E. ***The Condesed Nearest Neighbor Rule: IEE Transactions on Information Theory***. New Jersey, 1968.

HAYKIN, Simon. **Redes neurais: princípios e prática**: Bookman Editora. Porto Alegre, 2007.

HÜNECKE, Klaus. ***Jet Engines: Fundamentals of Theory. Design and Operation***. Berlin, 2003.

IATA. ***WATS: World Air Transport Statistics 2019***. 2019

JAPKOWICZ, Nathalie et al. ***Learning from imbalanced data sets: a comparison of various strategies. In: AAAI workshop on learning from imbalanced data sets***. Nova Escotia, 2000.

KUBAT, M.; HOLTE, R.; MATWIN, S. ***Machine Learning for the Detection of Oil Spills in Satellite Radar Images: Machine Learning***. Boston, 1998.

KUBAT, M.; MATWIN, S. ***Addressing the curse of imbalanced training sets: One sided selection. In Proceedings of the Fourteenth International Conference on Machine Learning.*** Tennessee, 1997.

LAROSE, Daniel. ***Discovering Knowledge in Data: An Introduction to Data Mining***. New Jersey, 2014.

LEWIS, D.; CATLETT, J. **Heterogeneous Uncertainity Sampling for Supervised Learning. In *Proceedings of the Eleventh International Conference of Machine Learning***. California, 1994.

MASON, Andrew; RÖNNQVIST, Mikael. ***Solution methods for the balancing of jet turbines: Computers & operations research***. v. 24. Auckland, 1997.

NUNES, Nuno António Neves; E SILVA, Júlio Montalvão. **Contribuição para a concepção de sistemas inteligentes de diagnóstico em controlo de condição por análise de vibrações de motores de aeronave:** Universidade Técnica de Lisboa. Lisboa, 2005.

PAZZANI, M.; MERZ, C.; MURPHY, P.; ALI, K.; HUME, T.; BRUNK, C. ***Reducing Misclassification Costs: In Proceedings of the Eleventh International Conference on Machine Learning***. San Francisco, 1994.

PEREIRA, José Cristiano. **Modelo causal para análise probabilística de risco de falhas de motores a jato em situação operacional de fabricação**. Rio de Janeiro, 2017.

REZENDE, S. O. **Sistemas Inteligentes –Fundamentos e Aplicações**. Barueri, 2005.

SHLENS, Jonathon. ***A tutorial on Principal Component Analysis*.** California, 2005.

SILVA, Samuel da. **Vibrações Mecânicas**. Foz do Iguaçu: UNIOSTE, 2009.

THE GENX COMMERCIAL AIRCRAFT ENGINE. ***General Electric***, 2020. Disponível em: <https://www.geaviation.com/commercial/engines/genx-engine>. Acesso em 2 de Jul de 2020.

TURBOFAN. ***National Aeronautics and Space Administration***, 2015. Disponível em: < https://www.grc.nasa.gov/WWW/k-12/airplane/turbfan.html>. Acesso em: 2 de Jul de 2020.

TURBOFAN. **Wikipedia**, 2020. Disponível em: < https://pt.wikipedia.org/wiki/Turbofan#/media/Ficheiro:Turbofan_operation.svg>. Acesso em: 2 de Jul de 2020.

TURBOSHAFT AND TURBOPROP. ***Aerospaceweb***, 2005. Disponível em: <http://www.aerospaceweb.org/question/propulsion/q0209.shtml>. Acesso em: 28 de Jul de 2020.

TOMEK, Ivan. ***A generalization of the k-NN rule: IEEE Transactions on Systems, Man, and Cybernetics***. New Jersey, 1976.

TURRIONI, João Batista; MELLO, Carlos Henrique Pereira. **Metodologia de Pesquisa em Engenharia de Produção.** Minas Gerais, 2012.

WEISS, G. M.; PROVOST, F. ***Learning when training data are costly: the effect of class distribution on tree induction. Journal of Artificial Intelligence Research***. New Jersey, 2003.

WITTEN, I. H.; FRANK, E. ***Data Mining - Practical Machine Learning Tools and Techniques***. Elsevier, 2016.

WOODS, K.; DOSS, C.; BOWYER, K.; SOLKA, J.; PRIEBE, C.; KEGELMEYER, P. ***Comparative Evaluation of Pattern Recognition Techniques for Detection of***

Microcalcifications in Mammography: International Journal of Pattern Recognition and Artificial Intelligence. Cidade de Singapura, 1993.

YAN, R.; LIU, Y.; JIN, R.; HAUPYMANN, A. ***On predicting rare classes with SVM ensembles in scene classification: In IEEE International Conference on Acoustics, Speech and Signal Processing***. New Jersey, 2003.