**Projeto de *Data Mining***

Manutenção Preditiva de uma Máquina Industrial

Tiago Ribeiro, 2210785

Rute Santos, 2210805

Índice

[1. Introdução 1](#_Toc109218212)

[2. Compreensão do Negócio 5](#_Toc109218213)

[2.1 Descrição dos Objetivos 5](#_Toc109218214)

[2.2 Questões Legais 7](#_Toc109218215)

[2.3 Glossário de Negócio 12](#_Toc109218216)

[2.4 Análise de Custo-Benefício 20](#_Toc109218217)

[2.5 Descrição do Contexto 21](#_Toc109218218)

[2.6 Planeamento do Projeto 24](#_Toc109218219)

[2.6.1 Atividades e Subatividades 24](#_Toc109218220)

[3. Compreensão dos Dados 27](#_Toc109218221)

[3.1 Recolha dos Dados 27](#_Toc109218222)

[3.2 Descrição dos Dados 27](#_Toc109218223)

[3.3 Exploração dos Dados 31](#_Toc109218224)

[3.4 Qualidade dos Dados 39](#_Toc109218225)

[4. Início do Projeto Prático 40](#_Toc109218226)

[4.1 Preparação dos Dados 41](#_Toc109218227)

[4.2 Seleção dos Dados 41](#_Toc109218228)

[4.3 Limpeza de Dados 42](#_Toc109218229)

[4.4 Formatação de Dados 42](#_Toc109218230)

[4.5 Conjunto de Dados Final 44](#_Toc109218231)

[5. Modelação 45](#_Toc109218232)

[5.1 Selecionar as Técnicas de Modelação 46](#_Toc109218233)

[5.2 Conceção do Teste do Modelos 46](#_Toc109218234)

[5.3. Construção dos Modelo 49](#_Toc109218235)

[5.3.1 Árvores de Decisão 49](#_Toc109218236)

[5.3.2 Máquinas de Suporte de Vetores 50](#_Toc109218237)

[5.3.3 Floresta Aleatória 51](#_Toc109218238)

[5.3.3. Rede Neuronal 53](#_Toc109218239)

[5.3 Avaliação do Modelo 54](#_Toc109218240)

[6. Avaliação 55](#_Toc109218241)

[6.1 Avaliação dos Resultados 55](#_Toc109218242)

[6.2 Processo de Revisão 57](#_Toc109218243)

[6.3 Determinar os Próximos Passos 57](#_Toc109218244)

[7. Desenvolvimento 58](#_Toc109218245)

[7.1 Plano de Implementação 58](#_Toc109218246)

[7.2 Plano de Monitorização e Manutenção 61](#_Toc109218247)

[7.3 Elaboração do Relatório Final 61](#_Toc109218248)

[7.4 Revisão do Projeto 61](#_Toc109218249)

[8. Conclusão 61](#_Toc109218250)

[9. Referências 63](#_Toc109218251)

[6. Apêndice **Erro! Marcador não definido.**](#_Toc109218252)

Índice de figuras

[**Figura 1 -** Fases do modelo de referência CRISP-DM 1](#_Toc109218253)

[**Figura 2 -** Representação de um sistema de caixa preta. Neste tipo de sistemas apenas a entradas e saídas são conhecidas, desconhecendo-se o seu funcionamento interno 3](#_Toc109218254)

[**Figura 3 -** Representação de um sistema transparente e interpretável. O funcionamento interno e previsões geradas pelo modelo são explícitas e percetíveis para os humanos 3](#_Toc109218255)

[**Figura 4 -** Matriz de confusão e métricas 4](#_Toc109218256)

[**Figura 5 -** Gráficos de barras relativos às falhas e qualidade dos produtos 32](#_Toc109218257)

[**Figura 6 -** Histogramas e caixas-de-bigodes das variáveis da Temperatura e Velocidade Rotacional 33](#_Toc109218258)

[**Figura 7 -** Histogramas e caixas-de-bigodes das variáveis de Binário e Desgaste da Ferramenta 34](#_Toc109218259)

[**Figura 8 -** Gráfico de dispersão 3D da Velocidade de Rotação e diferença de Temperatura do Ar e do Processo 35](#_Toc109218260)

[**Figura 9 -** Gráfico de dispersão do Desgaste da Ferramenta 36](#_Toc109218261)

[**Figura 10 -** Gráfico de dispersão da Potência Elétrica 36](#_Toc109218262)

[**Figura 11 -** Gráfico de dispersão da Falhas por Fratura por Fadiga para os diferentes níveis de qualidade dos produtos 37](#_Toc109218263)

[**Figura 12 -** Gráfico variável binária **random\_failure** 37](#_Toc109218264)

[**Figura 13 -** Matriz de correlação das variáveis dos sensores 38](#_Toc109218265)

[**Figura 14 –** Excerto do script Python relativo a dados omissos e entradas duplicadas 39](#_Toc109218266)

[**Figura 15 -** Excerto do script Python para renomeação e conversão das variáveis 40](#_Toc109218267)

[**Figura 16 –** Captura da tela do resultado da contabilização do número e categoria de falha 40](#_Toc109218268)

[**Figura 17** - Método de amostragem por Validação Cruzada com k = 3 47](#_Toc109218269)

[**Figura 18** – Esquema da sequência de passos seguidos para testar os modelos. A cor, a diferentes vias criadas para cada um dos algoritmos 48](#_Toc109218270)

[**Figura 19** - Captura de ecrã de excerto do *Jupyter notebook* relativo à busca bayesiana de hiperparâmetros 49](#_Toc109218271)

[**Figura 20** - Treino e desempenho de algoritmo com os dados de treino 50](#_Toc109218272)

[**Figura 21** - Criação e definição do espaço de busca de hiperparâmetros 50](#_Toc109218273)

[**Figura 22** - Detalhe da função de busca bayesiana para o modelo com algoritmo de Máquina de Suporte de Vetores 51](#_Toc109218274)

[**Figura 23** – Modelo encontrada após busca de hiperparâmetros com dados sobreamostrados 51](#_Toc109218275)

[**Figura 24** – Criação do modelo base de através 52](#_Toc109218276)

[**Figura 25** -Código relativo ao treino do modelo de Floresta Aleatória 52](#_Toc109218277)

[**Figura 26** - Detalhe do desenho do teste do modelo com rede Neuronal 53](#_Toc109218278)

[**Figura 27** – Configuração do melhor modelo de Rede Neuronal encontrado 53](#_Toc109218279)

[**Figura 28** – Página de Exploração de Dados desenvolvida com a plataforma Stremlit 59](#_Toc109218280)

[**Figura 29** - Aspeto gráfico do protótipo desenvolvido para *deployment* do modelo preditivo 60](#_Toc109218281)

Índice de tabelas

[**Tabela 1 -** Tipos de Defeitos de Produção, Consequências e Ações para a sua correção 6](#_Toc109218282)

[**Tabela 2 -** Resumo dos Custos e Benefícios da Implementação do Projeto de *Data Mining* 21](#_Toc109218283)

[**Tabela 3 –** Plano de atividades das 6 fases do Projeto de *Data Mining* 25](#_Toc109218284)

[**Tabela 4 -** Fonte, formato e tamanho do ficheiro do conjunto de dados 27](#_Toc109218285)

[**Tabela 5 -** Descrição dos variáveis do conjunto de dados 29](#_Toc109218286)

[**Tabela 6 -** Causas das falhas da máquina 30](#_Toc109218287)

[**Tabela 7 -** Tipos de falha quando variável **machine\_failure** é verdadeira 31](#_Toc109218288)

[**Tabela 8 -** Medidas de tendência central e de distribuição de frequências 31](#_Toc109218289)

[**Tabela 9 -** Captura de tela dos primeiros 10 registos do conjunto de dados final 45](#_Toc109218290)

[**Tabela 10** – Resultados do Desempenho dos Modelos 54](#_Toc109218291)

[**Tabela 11** – Métricas de Desempenho do Modelo de Árvore de Decisão 55](#_Toc109218292)

[**Tabela 12** - Métricas de Desempenho do Modelo de Floresta Aleatória 55](#_Toc109218293)

[**Tabela 13** - Métricas de Desempenho do Máquinas de Suporte de Vetores 56](#_Toc109218294)

[**Tabela 14** - Métricas de Desempenho do Máquinas de Suporte de Vetores 56](#_Toc109218295)

[**Tabela 15** – Tempo de cálculo de predição de todos os dados do conjunto de treino 57](#_Toc109218296)

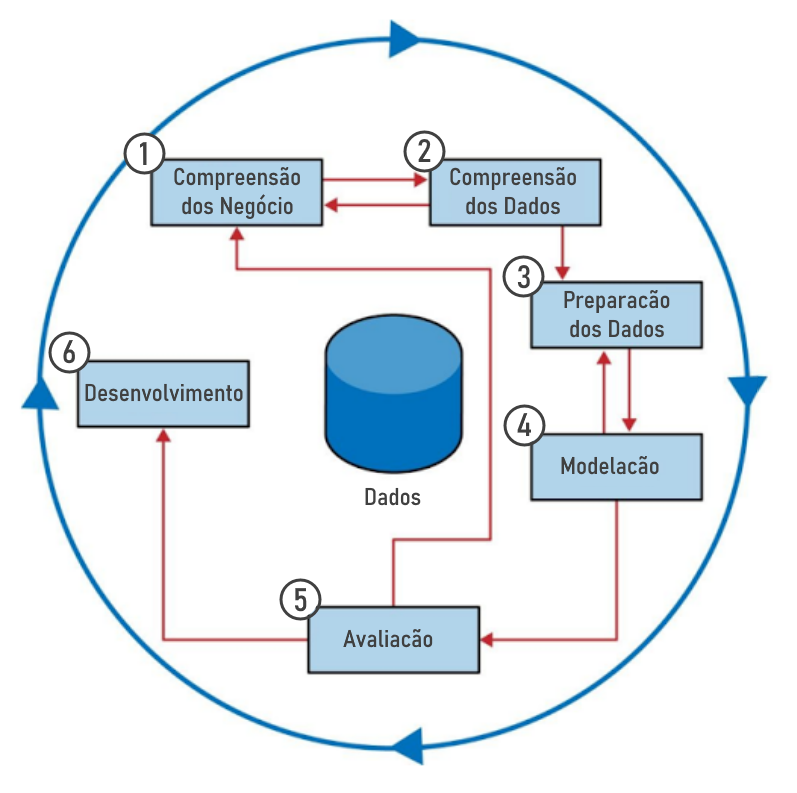
[**Tabela 16** – Síntese das características das plataformas de implementação de modelos 58](#_Toc109218297)

# 1. Introdução

No âmbito da unidade curricular de *Data Mining*, inserida no Mestrado em Ciência de Dados da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, lecionada pelo professor Ricardo Malheiro, foi-nos proposta a elaboração de um projeto de *Data Mining*.

O objetivo deste projeto passa pela resolução de um problema de negócio através da utilização de algoritmos de previsão e técnicas de classificação tendo por base a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*).

A metodologia CRISP-DM consiste num conjunto de boas práticas para executar um projeto de ciência de dados e divide-se em 6 etapas fundamentais, são elas: compreensão do negócio, compreensão dos dados, preparação dos dados, construção do modelo de predição, avaliação e validação do modelo e colocação do modelo em produção.



**Figura 1 -** Fases do modelo de referência CRISP-DM

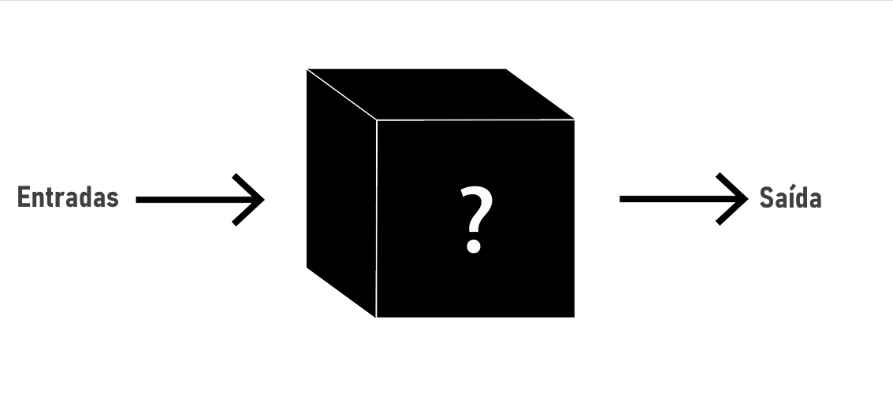
A Região Centro de Portugal Continental, fruto de vários ciclos de industrialização (Caetano e Gama 2004) e das suas características socioeconómicas, contém vários polos industriais do quais se destacam os de Águeda, Leiria, Marinha Grande e Porto de Mós. Particularizando ao distrito de Leiria, sobressaem-se os setores dos moldes metálicos e indústria metalomecânica, do vidro, da cerâmica, dos plásticos, da produção de carne e rações e da transformação de madeira (Fórum Estudante 2019) (Jornal de Leiria 2019).

O futuro destas indústrias depende da eficiência com que emprega os seus recursos e, numa altura em que preconiza a visão da chamada Indústria 4.0[[1]](#footnote-2), entende-se que as decisões tomadas para otimizar a produção devem ser suportadas por dados recolhidos por um crescente número de sensores. É neste contexto que entendemos ser relevante estudar técnicas de apoio à decisão que permitam reduzir o tempo de inatividade dos meios de produção, sem prejuízo para a explicabilidade das sugestões dados pelos sistemas automáticos (Loyola-González 2019) e da compreensão das causas de eventuais falhas das máquinas.

Este trabalho aborda o caso de uma unidade fabril que produz uma gama de produtos de cosmética. Esta unidade de produção fictícia recolhe dados dos sensores de uma máquina (temperatura, binário, velocidade de rotação e desgaste da ferramenta) e associa estes dados a falhas ocorridas na máquina durante o processo produtivo. É do interesse da unidade de produção, evitar que as máquinas parem de forma inesperada devido a avaria. Neste sentido, focar-nos-emos em desenvolver modelos de manutenção preditiva do equipamento, por oposição à manutenção preventiva e corretiva (Marcorin e Lima 2003).

As técnicas de manutenção preditiva são concebidas para ajudar a determinar o estado do equipamento em serviço, a fim de estimar quando a manutenção deve ser executada. Esta abordagem promove uma economia de custos em relação à manutenção preventiva de rotina baseada no tempo, porque as tarefas só são executadas quando se justifica. A manutenção preditiva também tem sido considerada como uma das forças motrizes para melhorar a produtividade reduzindo o tempo de inatividade (Amruthnath e Gupta 2018). Um aspeto que não está refletido no conjunto de dados recolhido, é a possível deterioração da qualidade dos produtos produzidos caso a manutenção seja deficiente.

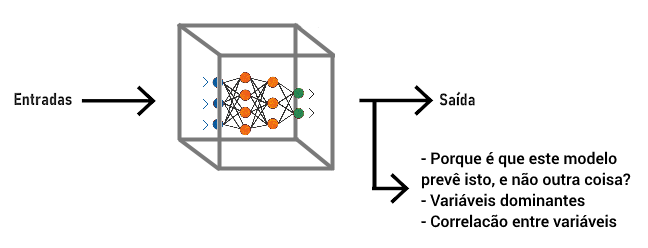
A manutenção preditiva difere da manutenção preventiva porque se baseia no estado real do equipamento, em vez das estatísticas de vida média ou esperada, para prever quando será necessária a manutenção. Tipicamente, são adotadas abordagens de aprendizagem automática para a definição do estado real do sistema e para a previsão dos seus estados futuros. No entanto, modelos de aprendizagem automática com complexidade suficientemente elevada deixam de ser facilmente interpretáveis pelos humanos e passam a estar na categoria dos sistemas de caixa preta (Figura 2). Isto é, sistemas cuja estrutura e funcionamento internos são desconhecidos e onde apenas as suas entradas e saídas são visíveis (Loyola-González 2019).



**Figura 2 -** Representação de um sistema de caixa preta. Neste tipo de sistemas apenas a entradas e saídas são conhecidas, desconhecendo-se o seu funcionamento interno

Nalguns casos não é vital perceber como a decisão foi feita e só o desempenho do algoritmo interessa, seja porque não tem consequências significativas caso haja uma predição errada, seja porque o problema é suficientemente bem estudado e validado em aplicações reais. Noutros, o modelo deve explicar como chegou à decisão, pois apenas ter uma decisão correta não resolve o problema na sua totalidade. Estes casos incluem os cenários onde o objetivo é obter conhecimento científico e compreender aprofundadamente cada um dos passos do processo, aplicações onde a segurança seja vital e pouco ou nada tolerante a falhas e sistemas onde questões de ordem ética possam estar em causa e que necessitam intervenção e responsabilização humanas, por exemplo (Doshi-Velez e Kim 2017).

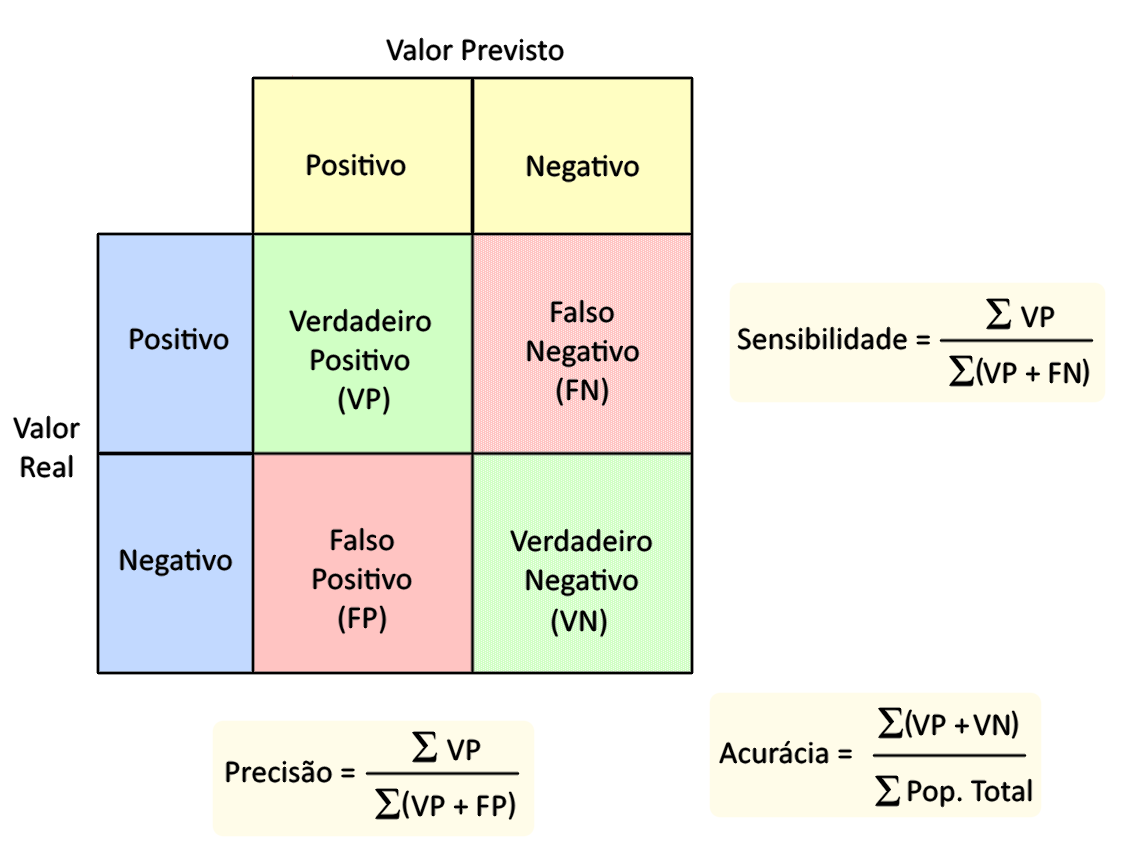
Fazendo uso do nosso caso, não interessa só saber quando uma máquina falhará (previsão), mas importa também perceber a razão da falha para perceber como prevenir o problema e evitar os prejuízos decorrentes da inoperabilidade do equipamento. Quanto mais explicável for o modelo, mais os envolvidos no processo poderão antecipar com eficácia os problemas. Por outro lado, se o modelo for transparente, a identificação de eventuais erros e posterior afinação do algoritmo serão facilitadas.



**Figura 3 -** Representação de um sistema transparente e interpretável. O funcionamento interno e previsões geradas pelo modelo são explícitas e percetíveis para os humanos

Para além do considerado em cima, estando este caso inserido no âmbito de uma produção industrial (ainda que fictícia), parece-nos importante introduzir previsibilidade e segurança, e reduzir os riscos de falha tanto quanto possível, mesmo que isso implique algumas paragens de produção (*downtime*) a mais. Isto é, em ambiente industrial as paragens programadas das máquinas para manutenção, apesar de implicarem um período de inatividade, provocam menor disrupção e permitem melhor planeamento e alocação de meios do que uma situação onde existe uma avaria inesperada que implica manutenção corretivaSupe. Acresce que nos casos onde há avaria, a reparação pode ser mais dispendiosa e prolongada.

Tipicamente, os conjuntos de dados recolhidos de máquinas industriais com o fim de identificar falhas são desbalanceados[[2]](#footnote-3) por conterem uma proporção de amostras em que as máquinas estão a trabalhar sem falhas críticas e uma fração de entradas onde houve falha (Matzka 2020). Se fosse o contrário é que seria de estranhar. O desequilíbrio de classes é comum em muitos outros domínios, incluindo em conjuntos de dados de deteção de fraude e spam, rastreio de doenças ou desastres naturais (Barella 2015). O caso de estudo apresentado não foge à regra. Se as falhas forem eventos relativamente raros, o modelo tem se ser particularmente sensível às falhas e, ao mesmo tempo, as medidas de erro têm de ser interpretadas com especial cuidado; em casos como este, um modelo hipotético que previsse sempre que a máquina não está em risco de falha, o valor da acurácia seria elevado, ainda assim (vide figura 4) estaríamos a ter em consideração uma métrica apropriada para conjunto de dados não-balanceados.



**Figura 4 -** Matriz de confusão e métricas

Deste modo, como métrica de comparação de modelos de previsão teremos como desígnio a minimização do número de Falso Negativos (situações onde a máquina precisaria de manutenção, mas o modelo prevê o contrário). A métrica mais suscetível aos Falsos Negativos é a Sensibilidade.

# 2. Compreensão do Negócio

## 2.1 Descrição dos Objetivos

A unidade fabril em causa é uma empresa criada em 2010, com o objetivo de lançar produtos cosméticos capilares, inovadores e de alta qualidade no mercado profissional nacional e internacional. Sendo o negócio em causa uma unidade fabril, os seus objetivos são essencialmente a rápida produção e a produção ao menor custo possível, de forma a maximizar o lucro gerado e reduzir o desperdício de recursos. Para isso, é importante que as máquinas se encontrem em bom estado e que não existam falhas que facilitem o prejuízo, não só de matérias-primas, como também de tempo despendido durante as falhas no processo produtivo. Assim, é fundamental que seja possível prevenir os problemas e evitar erros e falhas que possam decorrer da inoperabilidade dos equipamentos da unidade fabril.

Esta unidade fabril é composta por 5 departamentos distintos, contando com um total de 26 colaboradores, são eles:

· Administração – composto por 3 sócios-gerentes;

· Faturação – composto por apenas 1 membro;

· Marketing – composto por 3 membros;

· Recursos humanos – composto por apenas 1 membro;

· Comercial – composto por 2 membros;

· Laboratório – composto por 3 membros;

· Produção – composto por 10 membros e 1 gestor de produção;

· Expedição – composto por 2 membros.

Relativamente aos equipamentos e maquinaria, a fábrica é composta por um laboratório, onde são pensadas, estudadas e desenvolvidas todas as fórmulas que são produzidas na empresa; uma zona de enchimento, onde são enchidas todas as embalagens com os produtos criados; uma zona de pesagem, onde é garantido o peso certo a todas as embalagens; uma zona de rotulagem, onde são aplicados os rótulos das embalagens; uma zona que une a produção e expedição, onde são transportados todos os produtos já finalizados, para serem devidamente divididos e preparados para transporte. Todas estas zonas são compostas por máquinas específicas, nomeadamente:

· Máquina de rotulagem;

· Máquina de enchimento;

· Equipamentos técnicos laboratoriais (como é o caso das balanças, misturadoras, etc);

· Máquina de pesagem de produto;

· Batedeiras;

· Máquina de envase;

· Máquina de fermentação;

O mau funcionamento destes equipamentos pode levar a prejuízos bastante elevados para a empresa, como são exemplo:

|  |  |
| --- | --- |
| Tipos de Defeitos de Produção | Descrição |
| Rotulagem errada | Neste caso, os rótulos têm que ser novamente todos produzidos e alterados; |
| Medidas erradas | No laboratório é importante que toda a maquinaria funcione na perfeição, uma vez que uma quantia errada de uma certa substância ou matéria-prima pode danificar toda a produção, o que implica que sejam utilizados novos recursos para desenvolver a fórmula correta; |
| Divergência de volumes | Este caso implica que, no caso de a máquina possuir alguma avaria e o produto não seja completamente cheio, ou seja cheio em demasia, a produção também fique danificada, e caso não se verifique a tempo, pode causar insatisfação na entrega ao cliente; este caso, pode também danificar ou alterar as características específicas do produto; |
| Consistências inapropriadas | O produto é produzido com uma fórmula específica de forma a ter certas características e consistências, caso um produto que seja necessário ficar mais espesso, não fique com a consistência certa, tudo terá que ser produzido novamente; |

**Tabela 1 -** Tipos de Defeitos de Produção, Consequências e Ações para a sua correção

Além dos prejuízos que podem causar relativamente à produção do produto, os erros da maquinaria, podem também ser responsáveis por eventuais falhas ou atrasos na entrega dos produtos e no processo produtivo, que geram insatisfação nos clientes.

Dada a necessidade de uma rápida e elevada quantidade de produção, o projeto em causa, torna-se importante, de forma a garantir o bom funcionamento da empresa em causa, desenvolver modelos de manutenção preditiva dos equipamentos que nos permitam evitar desperdícios, erros e prejuízos. Esta técnica de manutenção preditiva permitirá à empresa determinar o estado dos equipamentos acima referidos, e estimar quando deve ser executada qualquer tipo de manutenção dos mesmos.

Para implementar este processo, a empresa terá que optar por recorrer à contratação externa para acompanhar o Gestor de Produção nas tarefas a desempenhar. O Gestor de Produção será responsável por transmitir todas as informações aos técnicos contratados e acompanhar todo o processo em causa. Pretende-se que o projeto seja implementado o mais rápido possível, mas como qualquer projeto, e dada a vasta quantidade de maquinaria disponível para análise, demora o seu tempo. Prevê-se assim, que todo o processo seja implementado num prazo total de 9 meses (contratação, desenvolvimento do projeto e implementação).

Espera-se que este projeto permita não só garantir a qualidade de produção, como também de segurança e o desperdício de tempo. Caso seja prevista a necessidade de manutenção de maquinaria corretamente, evita-se que sejam postas em causa outras questões no caso de o equipamento avariar, por exemplo a meio de uma produção.

Podem ser verificadas algumas restrições, como por exemplo a disponibilidade total do Gestor de Produção, uma vez que existe apenas um e este também terá que se dedicar à Gestão de todos os processos de produção dos quais já é responsável. Relativamente à recolha e análise dos dados, não se verificam quaisquer tipos de restrições.

A capacidade de produção da fábrica possui um mínimo de 300 unidades por produto, o que implica que, dependendo do cliente, as variedades e quantidades desenvolvidas em cada projeto podem ser bastante dispersas. Isto pode ser uma condicionante no processo preditivo, uma vez que numa produção para o cliente X pode ser necessário desenvolver apenas 300 unidades num dia, mas para o cliente Y, no mesmo dia, podem ser necessárias desenvolver 6000 unidades do produto.

## 2.2 Questões Legais

A nível das questões legais relativas ao negócio em causa, existem algumas normas que esta indústria precisa de seguir rigorosamente (Parlamento Europeu 2009) (Ministério da Saúde 2008) (Ministério do Ambiente 2017). Das quais se destacam:

* **Responsabilidade na produção e colocação no mercado**

|  |
| --- |
| Pessoa Responsável |
| ***Art. 4.º do Regulamento 1223/2009*** |
| “Só podem ser colocados no mercado produtos cosméticos para os quais seja designada uma pessoa singular ou coletiva responsável na comunidade”. A Pessoa Responsável (PR), que pode ser o fabricante, garante o cumprimento das obrigações aplicáveis previstas no Regulamento 1223/2009. |

|  |
| --- |
| Obrigações da Pessoa Responsável |
| ***Art. 5º. do Regulamento 1223/2009*** |
| * A segurança; * Boas práticas de fabrico; * Avaliação de segurança; * Ficheiro de informações sobre o produto; * Amostragem e análises; * Notificação; * Restrições aplicáveis a determinadas substâncias; * Informação ao consumidor; * Fiscalização do mercado; * Adoção de medidas corretivas; * Dever de cooperação com as autoridades. |

|  |
| --- |
| Obrigações da Pessoa Responsável |
| ***Art. 3º. do Regulamento 1223/2009*** |
| Os produtos cosméticos disponibilizados no mercado devem ser seguros para a saúde humana quando aplicados em condições normais ou razoavelmente previsíveis de utilização. |

* **Técnico Responsável**

Os fabricantes e os responsáveis pela colocação no mercado de produtos cosméticos devem ser assistidos por um Técnico Qualificado que com eles assume, solidariamente, a responsabilidade pela observância do disposto nos normativos aplicáveis ao exercício desta atividade.

|  |
| --- |
| Obrigações da Pessoa Responsável |
| ***Art. 25º. nº 1 do Decreto-Lei 189/2008*** |
| Este profissional, designado por Técnico Responsável, deve possuir uma das qualificações seguintes:  a) Licenciatura ou bacharelato em Ciências Farmacêuticas, Química, Biologia, Medicina ou Engenharia Química, obtidas em Universidade Portuguesa ou de outro estado membro;  b) Licenciatura ou bacharelato, reconhecidos em Portugal, em Química Cosmética ou Cosmetologia por Universidades estrangeiras;  c) Licenciatura ou bacharelato em curso especificamente reconhecido por despacho do Ministro da Educação como equivalente a algum dos indicados nas alíneas anteriores. |

* **Avaliador de Segurança**

A fim de demonstrar que os produtos cosméticos são seguros para a saúde humana, nos termos do artigo suprarreferido, a Pessoa Responsável, antes da colocação no mercado, deve certificar-se de que os mesmos foram submetidos a uma avaliação de segurança.

|  |
| --- |
| ***Art. 3º. do Regulamento 1223/2009*** |
| A avaliação de segurança é efetuada por pessoa que reúna os requisitos previstos:  a) Diploma ou outra prova formal de habilitações adquiridas com a conclusão de um curso universitário teórico e prático, em farmácia, toxicologia, medicina ou disciplina semelhante, ou de um curso reconhecido como equivalente por um Estado Membro. |

* **Gestão de embalagens e resíduos de embalagens**

De acordo com o ***Art 3º. do Decreto-Lei n.º 152-D/2017***, define-se como “embalagens” todos e quaisquer produtos feitos de materiais de qualquer natureza utilizados para conter, proteger, movimentar, manusear, entregar e apresentar mercadorias, tanto matérias-primas como produtos transformados, desde o produtor até ao utilizador ou consumidor, incluindo todos os artigos descartáveis utilizados para os mesmos fins.

No mesmo artigo, entende-se por “embalador” aquele que, a título profissional, embale ou faça embalar os seus produtos e que é responsável pela sua colocação no mercado. Os “embaladores”, são responsáveis pela gestão das embalagens e resíduos de embalagens, numa perspetiva de proteção ambiental, sendo igualmente responsáveis pela prestação de contrapartidas financeiras destinadas a suportar os acréscimos de custos com a recolha seletiva e a triagem de resíduos de embalagens, podendo estes recorrer à Sociedade Ponto Verde para executar esta recolha.

* **Boas práticas para o fabrico de cosméticos**

Segundo o **Regulamento (CE) n. 1223/2009**, é um requisito legal que a fim de garantir a segurança dos produtos cosméticos colocados no mercado, estes deverão ser produzidos segundo boas práticas de fabrico”.

Assim, entende-se por boas práticas de fabrico os procedimentos e práticas que asseguram que os produtos são continuamente produzidos e controlados de acordo com os padrões de qualidade adequados ao uso pretendido.

A **Norma ISO 22716** (*International Organization for Standardization* 2007) apresenta uma abordagem abrangente para o sistema de gestão da qualidade da cadeia produtiva envolvida no fabrico, embalagem, controlo da qualidade, armazenamento e transporte dos produtos cosméticos. Esta norma é composta por vários capítulos, nomeadamente:

1. Sistema de Gestão da Qualidade de Cosméticos e a Organização

2. Instalações e Equipamentos

3. Fabrico de Produtos e Gestão de Materiais

4. Desvios, Reclamações e Recolhas

5. Melhoria Contínua

|  |
| --- |
| ***Art. 11º. do Regulamento 1223/2009*** |
| Quando um produto cosmético é colocado no mercado, a Pessoa Responsável deve conservar, por um período de 10 anos a contar da data de colocação do último lote, um ficheiro de informações sobre o produto (FIP), prontamente acessível à autoridade competente do Estado-Membro onde o ficheiro se encontra. O FIP deve conter os seguintes dados e informações:    a) Uma descrição do produto cosmético que permita estabelecer uma associação clara entre o ficheiro de informações sobre o produto e o produto cosmético a que diz respeito  b) O relatório de segurança do produto cosmético, da responsabilidade do Avaliador de Segurança, elaborado nos termos do Anexo II  c) Uma descrição do processo de fabrico e uma declaração de conformidade com as boas práticas de fabrico |

* **Notificação à Comissão Europeia**

|  |
| --- |
| ***Art. 13º. do Regulamento 1223/2009*** |
| Antes da colocação de um produto cosmético no mercado, a Pessoa Responsável deve transmitir à Comissão Europeia, por via eletrónica através do portal CPNP (Cosmetic Products Notification Portal) (Comissão Europeia 2009), as seguintes informações:    a) A categoria a que pertence o produto cosmético e a sua designação ou designações, que permitam a sua identificação específica;  b) O nome e o endereço da Pessoa Responsável onde o ficheiro de informações sobre o produto se encontra disponível;  c) O país de origem em caso de importação;  d) O Estado-Membro em que se prevê a colocação do produto cosmético no mercado;  e) As coordenadas de uma pessoa singular a contactar em caso de necessidade;  f) A presença de substâncias sob a forma de nano materiais, caso se verifique, (…);  g) A denominação e o número CAS ou o número CE das substâncias classificadas como cancerígenas, mutagénicas ou tóxicas para a reprodução (CMR), pertencentes às categorias 1A ou 1B, nos termos da Parte 3 do anexo VI do Regulamento 1272/2008, caso estejam presentes;  h) A formulação-quadro que possibilite a prestação de um tratamento médico rápido e adequado em caso de dificuldades. |

* **Rotulagem**

|  |
| --- |
| ***Art. 19º. do Regulamento 1223/2009*** |
| Os produtos cosméticos só podem ser disponibilizados no mercado se o seu recipiente e a sua embalagem ostentarem em caracteres indeléveis, facilmente legíveis e visíveis, resumidamente, as seguintes informações:  a) O nome ou a firma e o endereço da Pessoa Responsável (…)  b) O conteúdo nominal no momento do acondicionamento, indicado em peso ou em volume (…);  c) A data até à qual o produto cosmético, armazenado em condições adequadas, continua a desempenhar a sua função inicial e, em especial, se mantém conforme com o disposto no art.º 3.º («data de durabilidade mínima»).  d) As precauções especiais de utilização (…);  e) O número de lote de fabrico ou a referência que permita identificar o produto cosmético (…);  f) A função do produto cosmético, salvo se esta decorrer claramente da respetiva apresentação;  g) Uma lista de ingredientes (…), precedida do termo em língua inglesa «ingredients». (…) A lista de ingredientes deve ser estabelecida por ordem decrescente do peso dos ingredientes no momento da incorporação no produto cosmético (…). |

## 2.3 Glossário de Negócio

**A**

* **Ácido Bórico:** adstringente, antisséptico com propriedades bacteriostáticas e fungistáticas.
* **Ácido Cítrico:** obtido de frutas ácidas como limão, laranja, tangerina, maracujá, entre outras. Possui propriedades acidificante nos cosméticos.
* **Ácido Glicólico:** alfa-hidroxiácido, é encontrado naturalmente na cana-de-açúcar. Aplicado sobre a pele provoca vasodilatação, diminui a espessura e a compactação do estrato córneo, acelera a renovação celular da epiderme e estimula a síntese de colágeno.
* **Ácido Ferúlico:** potente antioxidante fenólico encontrado em altas concentrações em plantas, principalmente no farelo de arroz e derivado também da matéria-prima da cevada maltada. Possui alto potencial fotoprotetor, age como uma barreira de membrana celular impedindo a atividade de radicais livres, diminuindo a absorção dos raios UVA e UVB, além de diminuir o eritema causado na pele devido à exposição. Excelente coadjuvante em tratamentos de melasma.
* **Ácido Hialurônico (Hyaxel):** substância altamente hidrofílica (afinidade pela água) combinada com Silício possui atividade biológica como principal glicosaminoglicana da derme, proporcionando viscoelasticidade a esta camada. Colabora no controle da hidratação e no tônus da pele, além da prevenção da integridade das fibras de colágeno.
* **Ácido Kójico:** obtido a partir da fermentação do arroz possui efeito inibidor sobre a tirosinase, por quelação dos iões de Cobre e consequente, diminuição da síntese de melanina.
* **Ácido Lático:** faz parte também do fator NMF (fator natural de hidratação), que retém a umidade na pele. Trata-se de um alfa hidroxiácido.
* **Ácido Salicílico:** possui propriedades queratolíticas. É usado em aplicações tópicas em tratamentos cosméticos de hiperqueratinização e escamação da pele. Beta hidroxiácido.
* **Alantoína:** princípio ativo encontrado no confrei, nas sementes de tabaco, na beterraba e no germe de trigo, favorece a proliferação celular acelerando a regeneração da pele lesada. É excelente cicatrizante, calmante e amaciante sobre tecidos.
* **Alpha bisabolol:** anti-inflamatório, cicatrizante e antisséptico suave.
* **Alfa-Hidroxiácido (AHA):** constituem um grupo de substâncias normalmente encontradas em frutas e alimentos e por isso, são conhecidos como ácidos de frutas. São eles: o ácido glicólico, lático, cítrico, tartárico, málico e mandélico. Os alfas hidroxiácidos aumentam a retenção de água da epiderme em baixas concentrações e aumentam o processo de renovação celular em altas concentrações.
* **Água de coco:** extrato de origem vegetal que além dos minerais presentes, inclui aminoácidos essenciais, polissacarídeos e vitaminas que realizam a verdadeira hidro nutrição da pele mantendo seu equilíbrio hídrico. Todos esses elementos equilibram a barreira natural de defesa da pele (NMF) mantendo a pele hidratada e protegida da ação do vento, frio, ar condicionado, radiação solar, entre outras, ajuda a recuperar o tecido pré e pós-tratamentos estéticos.
* **Aquasense:** extraído da casca do Angico, com coloração avermelhada natural, ativa os canais de água da pele (aquaporinas) oferecendo excelente hidratação.
* **Arnica:** com ação emoliente, estimulante, anti-inflamatória, antisséptica, analgésica, entorses, distensões musculares e estimulantes da circulação sanguínea.
* **Algisium C (Metilsilanol Manuronato):** utilizado nos tratamentos cosméticos contra celulite. Atua na lipólise, prevenindo e atenuando os processos inflamatórios.
* **Argisil C:** estimula os mensageiros endógenos envolvidos na comunicação celular capazes de modular o estoque de gordura nos adipócitos e promover uma lipólise intensa. Protege o colágeno da glicação, ou seja, da reticulação das fibras de sustentação da pele, preservando a tonicidade e firmeza da pele.
* **Algas vermelhas:** contém altíssimos níveis de sais minerais como Iodo, Magnésio, Potássio, Ferro, Sódio e Zinco, além das Vitaminas A, B1, B3, B6, B12, C, D e E e outras substâncias como glicoproteínas, oligoelementos e aminoácidos.

**B**

* **Baby Hair:** são os fios de cabelo curtinhos nas bordas do cabelo, por isso têm esse nome (cabelo de bebé).
* **Bantu Knot Out:** uma técnica para obter um penteado com caracóis que consiste em fazer coques (bantu knots) enquanto o cabelo está húmido e desfazê-los quando já estiver seco. No final, o cabelo fica com um formato diferente de caracóis. Esta técnica pode ser feita tanto em cabelos naturais como frisados.
* **Betaglucan:** ativo calmante e reestruturante.
* **Big Chop (BC):** significa “um grande corte ao cabelo”. Quando alguém com o cabelo desfrisado decide cortar as pontas desfrisadas e só deixar a parte natural do cabelo (raiz) diz-se que fez um Big Chop.
* **Bioex Antilipêmico:** blend de ativos anticelulite, estimulante metabólico, anti-inflamatório, descongestionante, ativador da microcirculação, vasoconstritor e antilipêmico.
* **Braid Out:** uma técnica para obter um penteado com caracóis que consiste em fazer tranças no cabelo húmido/hidratado. Quando o cabelo seca desfaz-se as tranças para criar uns caracóis diferentes e definidos. Esta técnica pode ser feita tanto em cabelos naturais como desfrisados.

**C**

* **Cafeisilane C:** associa a ação lipolítica da cafeína com as propriedades biológicas dos silanóis. A cafeína inibe a fosfodiesterase e os silanóis estimulam o AMP cíclico. Ambos facilitam a lipólise.
* **Cafeína 5%:** estimula a lipólise ao inibir a fosfodiesterase, resultando numa ação lipolítica no metabolismo dos ácidos graxos, diminuindo o tamanho do adipócito. Caulim: abrasivo e protetor da pele.
* **Capsicum:** ação rubefaciente, aumenta a permeabilidade capilar, estimulando a circulação sanguínea, promovendo maior oxigenação e nutrição.
* **Ceramidas:** são lipídios. Os lípidos atuam como uma barreira na cutícula dos fios de cabelo, o que leva as camadas da cutícula a terem uma melhor ligação entre elas, impedindo assim que os fios percam nutrientes. Óleos naturais como o Óleo de cártamo, Óleo de semente de uvas, ou Óleo de Girassol contêm alto teor de ceramidas o que beneficia no selamento da cutícula do fio.
* **Shampoo Purificante/Clarificante/Clarificador/Anti-Resíduos:** é um shampoo capaz de limpar a fundo todos os resíduos, impurezas e produtos deixados no cabelo e couro cabeludo. Estes shampoos contêm sulfatos.
* **Chá Verde:** potente ação antioxidante que combinando aos AHA´s, diminui o potencial de irritação dos mesmos.
* **Cobre:** atua na formação de colágeno, elastina e ácidos nucleicos. Possui ação antisséptica e bacteriostática.
* **Colágeno:** agente de hidratação.
* **Co-Wash:** significa lavar o cabelo sem shampoo, ou seja, só lavar com um condicionador. O propósito é manter o cabelo mais hidratado no final da lavagem.
* **Cronograma Capilar:** é uma rotina de cuidados do cabelo que consiste em intercalar etapas de hidratação, nutrição e reconstrução para repor todos os nutrientes e proteínas perdidos pelo fio. A frequência deve ser adaptada a cada tipo de cabelo e para facilitar a organização pode ser montado um cronograma com os dias da aplicação de cada etapa.
* **Custard**: um produto finalizante que é uma combinação de creme + gel. Serve para definir e ao mesmo tempo hidratar os fios.

**D**

* **Day After:** corresponde ao dia seguinte à lavagem do cabelo.
* **Dedoliss:** uma técnica para obter caracóis com a máxima definição que consiste em passar uma pequena mecha do cabelo ao redor de um dedo como se fosse um modelador de caracóis.
* **Deep Conditioning:** é um tratamento capilar que consiste na aplicação de uma máscara hidratante, nutritiva ou reconstrutora e uso de vaporização/calor durante 15 a 45 min dependendo de cada tipo de cabelo. Opcionalmente, pode ser retirado o uso da vaporização/calor e deixar-se atuar durante 1h ou mais.
* **Densiskin D+:** complexo biológico com ação completa e intensiva sobre o microrrelevo cutâneo e sobre os constituintes da matriz extracelular, estimula a produção de enzimas precursoras da vitamina D ativa.

**E**

* **Extrato de Acerola:** ação antioxidante.
* **Extrato de Algas Marinhas:** composto por vários tipos de algas que associadas, potencializam o efeito de emoliência, umectação e hidratação. Protetora do tecido cutâneo e ativadora do metabolismo.
* **Extrato de Aloe Vera:** descongestionante, hidratante e cicatrizante.
* **Extrato de Açaí:** potente ação oxidante.
* **Extrato de Calêndula:**  agente cicatrizante, antisséptico, bacteriostático, calmante, descongestionante e anti sensibilizante, presente em produtos que previnem a acne.
* **Extrato de Camomila:** possui propriedades calmante, fungicida, cicatrizante, anti-inflamatório e anti séptico.
* **Extrato de Capsicum:** rubefaciente, revulsivo, tônico capilar, antisséptico e estimulante da circulação periférica. Estimulante capilar na restauração do bolbo piloso, restaurador da pele do corpo e rosto.
* **Extrato de Castanha-da-Índia:** estimulador da circulação periférica. Usado em produtos anticelulite como estimulante da circulação local. É adstringente, tonificante, antisséptico e anti-inflamatório.
* **Extrato de Centella Asiática:** usado em produtos que têm a finalidade de reduzir a gordura localizada pelo mecanismo de ativação da circulação periférica, drenando os líquidos.
* **Extrato de Copaíba:** possui ação germicida, cicatrizante e anti-inflamatório.
* **Extrato de Gingko Biloba:** hidratante, estimulante da circulação periférica, protetora contra radicais livres, reestruturante, bioenergizante e ação anti-inflamatória.
* **Extrato de Hamamélis:** propriedade adstringente. Usado em produtos para pele oleosa, acnéica e/ou com poros dilatados.
* **Extrato de Hera:** estimulante metabólico, vasoconstritor, descongestionante, anti-inflamatório, antilipêmico e adelgaçante.
* **Extrato de Hortelã:** antisséptico, tonificante e adstringente.
* **Extrato de Laranja Amarga:** estimulante circulatório, tonificante e lipolítico.
* **Extrato de Maracujá:** hidratante, calmante e antioxidante.
* **Extrato de Própolis:** extraído do mel de abelhas. Possui ação secativa, hidratante, antisséptica, adstringente, cicatrizante, hemostático, bactericida e fungicida.
* **Extrato de Quinoa:** rico em aminoácidos, mais precisamente 20, dos quais 10 são essenciais, possui também ácido linoleico, vitaminas e minerais.

**F**

* **Finger Coils:** é o mesmo que Dedoliss. Uma técnica para obter caracóis com a máxima definição que consiste em passar uma pequena mecha do cabelo ao redor de um dedo como se fosse um modelador de caracóis.
* **Fitagem:** é uma técnica feita em cabelos encaracolados e crespos para obter caracóis super definidos, reduzir o frizz e controlar o volume. Consiste em dividir o cabelo em secções, dividir cada secção em várias mechas menores, aplicar um produto ativador e deslizar com os dedos entre os fios até formar caracóis definidos.
* **Flexi Rod:** é um rolo flexível usado para criar caracóis no cabelo. Existem rolos flexíveis de várias espessuras, cada um cria um formato de caracol diferente.

**G**

* **Gel:** o gel é conhecido por ter uma consistência mais leve e líquida e tem o propósito de definir, mas também dar volume aos caracóis.
* **Gelatina:** é um produto de finalização bastante usado pelas naturalistas, principalmente no Brasil, para dar mais definição e fixação aos caracóis e controlar o frizz. Tem alta fixação, resultados duradouros e a consistência é também mais densa e pesada que o gel normal.

**H**

* **Hair Journey:** é a expressão usada para definir a jornada de melhores cuidados capilares.
* **Hidratação:** é uma etapa de tratamento capilar que tem como propósito repor as moléculas de água perdidas nos fios.

**I**

* **Irgasan:** antisséptico, bactericida de amplo espectro e fungicida.

**J**

* **JBCO:** significa “Jamaican Black Castor Oil” (Óleo de Rícino Jamaicano). É um óleo que estimula o fortalecimento, crescimento e volume do cabelo.

**L**

* **L-carnitina:** aliado à cafeína possui um efeito “bomba de ácidos graxos livres”, deixando-os deste modo, à disposição das cadeias respiratórias mitocondriais para sua combustão.
* **Leave-in:** é um condicionador que não precisa de ser removido do cabelo, ou seja, depois da aplicação do leave-in não se enxagua o cabelo. É quase idêntico a um creme, mas foi inventado para ser usado essencialmente no cabelo molhado/húmido após a lavagem para que o cabelo permaneça o mais hidratado e macio possível.
* **Low Poo:** é uma técnica de cuidados de cabelo em que são utilizados shampoos somente sem sulfatos e produtos sem componentes derivados do petróleo (petrolatos/petrolatum, óleo mineral, parafina líquida, vaselina…).

**M**

* **Manteiga de Cupuaçu:** possui alta capacidade de absorção de água. Proporciona elasticidade e suavidade à pele.
* **Manganês:** é fundamental para a síntese de proteínas de DNA e RNA. A associação Manganês-Cobre é recomendada para o tratamento da acne e celulite.
* **Magnésio:** indicado para as reações biológicas da derme, é necessário para a síntese de vários compostos com ligações de muita energia de qualquer tipo.
* **Mentol:** antisséptico, vaso dilator e refrescante.
* **Microesfera de Polietileno:** são utilizadas em esfoliantes físicos. Devido à sua natureza esférica, deixa a pele limpa e com brilho natural sem agredi-la.
* **Mirtilo:** rico em ácido láctico de origem vegetal.

**N**

* **Niacinamida (Vitamina B3):** no envelhecimento, seus níveis sistêmico e intracelular diminuem. A suplementação em cosméticos reforça a função de barreira reduzindo a perda transepidermal de água, estimula as células basais, regula a biossíntese de involucrina e filagrina, importante para o processo de divisão e diferenciação celular, aumenta a produção de colágeno, seu uso tópico na concentração de 4% promove a melhora clínica de rugas de forma moderada a intensa, ajuda também no controle da acne.
* **Nicotinato de Metila:** vasodilatador, hiperemiante e estimulante, facilita a permeação de outras substâncias incorporadas no tratamento.
* **No Poo:** é uma técnica de cuidados de cabelo em que não são utilizados shampoos nem produtos com componentes derivados do petróleo (petrolatos/petrolatum, óleo mineral, parafina líquida, vaselina…) nem produtos com silicones insolúveis.
* **Nutrição:** é uma etapa de tratamento capilar que tem como propósito repor os nutrientes (lípidos) que o cabelo perde. Esta etapa sela as cutículas dos fios para que as moléculas hidratantes não saiam. Com a nutrição, a hidratação dura mais tempo no cabelo.

**O**

* **Óleo de Abacate:** possui propriedades emolientes, dermoprotetoras, hidratantes, lubrificantes, suavizantes e condicionadoras.
* **Óleo de Algodão:** utilizado para restaurar a barreira lipídica.
* **Óleo de Amêndoas:** utilizado como emoliente e umectante nas preparações cosméticas. Fornece maior hidratação e mantém a oleosidade natural dos cabelos e pele, perdidas pela ação do detergente.
* **Óleo de Andiroba:** promove ação anti-inflamatória e regeneradora, destinadas ao tratamento cosmético contra celulite e regeneração cutânea.
* **Óleo de Canola:** auxilia na recomposição da oleosidade natural em peles ressecadas e com problemas de escamação. Tem capacidade de proteger e regenerar a pele danificada e queimada.
* **Óleo de Gérmen de Trigo:** suavizante, hidratante e emoliente. Recomendado para peles sensíveis e desidratadas.
* **Oligomix:** associação de oligoelementos Zinco, Cobre, Magnésio e Manganês, ligados a um aminoácido.
* **Óxido de Zinco:** filtro solar físico.

**P**

* **PCA Na:** ativo hidratante, principal componente do NMF. Excelente função de reconstituição do manto hidrolipídico ideal para uso em tônicos faciais.
* **Prepoo:** é um tratamento com máscaras ou com óleos naturais que se faz antes de lavar o cabelo. Consiste em aplicar, por exemplo, uma máscara no cabelo seco ou humedecido, um dia antes ou umas horas antes de lavar o cabelo, para deixá-lo mais hidratado, macio e fácil de desembaraçar.
* **Porosidade do cabelo:** corresponde à capacidade que o cabelo tem de reter hidratação. Um cabelo com baixa porosidade sente dificuldade a absorver a hidratação, mas assim que absorve retém-na durante um bom tempo.
* **Pudding:** é um produto finalizante que é uma combinação de creme + gel. Serve para definir e ao mesmo tempo hidratar os fios. Ao contrário dos Custards, os Puddings costumam ter uma consistência mais parecida a um creme (meio que um pudim), enquanto os Custards costumam parecer-se mais com um gel.

**R**

* **Raffermine:** estimula a produção de colágeno.
* **Reconstrução:** é o mesmo que Protein Treatment. É uma etapa de tratamento capilar que tem como propósito repor as proteínas nos fios. As proteínas fortalecem o cabelo, deixando-o mais resistente, e também selam as cutículas.

**S**

* **Salicilato de Metila:** agente vasodilatador.
* **Selagem:** significa “tapar” as falhas na cutícula dos fios de cabelo de forma que as moléculas de água e os nutrientes não se percam, assim o cabelo mantém-se hidratado durante um longo período de tempo e sem frizz.
* **Silanol C:** oligoelemento presente no organismo com ação reestruturante e regeneradora da pele.
* **Slimbuster L:** ativo dermocosmético que trabalha de forma potente sobre a celulite, gordura localizada e flacidez da pele, induzindo a lipólise através de múltiplo mecanismo de ação.
* **Sugar Maple:** extrato rico em ácido málico.

**T**

* **Tensine:** agente tensor capaz de se distribuir na superfície da pele e formar um filme contínuo e coeso. Além do efeito “flash”, reduzindo o número e profundidade das rugas.
* **Trietanolamina:** Comumente encontrado em várias formulações para acertar o pH é utilizado como amolecedor de comedões.
* **Texturizer:** é o mesmo que relaxante. É um desfrisante que não tem como intuito alisar os fios, mas sim abrir os caracóis.
* **Texturização:** consiste em influenciar ou estimular os fios a ganhar forma ou até mesmo uma nova textura.
* **Transição:** é a palavra usada para definir o processo de passar de um cabelo desfrisado/pintado/progressiva para um cabelo totalmente natural.
* **Twist Out:** uma técnica para obter caracóis que consiste em fazer twists no cabelo húmido/hidratado. Quando o cabelo seca desfaz-se os twists para criar uns caracóis diferentes e definidos.

**U**

* **Umectação:** é uma técnica *prepoo* com óleos naturais que nutre, faz reter a hidratação e fortalece os fios que estão secos e/ou danificados.
* **Uva Ursi:** rico em arbutin natural, ativo clareador que atua inibindo a tirosinase (enzima envolvida na melanogenese) inibindo assim a formação da melanina. Atua também na absorção do excesso de melanina já formada.
* **Ureia:** ação queratolítica de regeneração celular. Dependendo da concentração, pode ser usada também como um hidratante.

**V**

* **Vitamina A:** desempenha importante papel na regulação do crescimento das células epiteliais e manutenção de sua integridade.
* **Vitamina E:** age como antioxidante biológico protegendo a pele dos radicais livres.
* **Vitamina C estável (Nanoesferas):**  tetraisopalmitato de ascorbila, conhecido como vitamina C oleosa, em nanopartículas lipofílicas de alta performance. Esta substância é muito mais estável do que o ácido ascórbico devido à presença de grupamentos que estabilizam a molécula deixando-a solúvel em lipídios. Antioxidante, clareador e renovador do colágeno.

**W**

* **Wash-N-Go:** é quando decides lavar o cabelo e deixá-lo secar ao natural sem ajuda de brushing e sem recorrer a penteados como Twist Outs, Braid Outs, etc.

**Z**

* **Zinco:** ação anti seborreica. Quando associados, Zinco e Cobre, é recomendada por sua ação bacteriostática e antisséptica excelente em cosméticos para acne.

## 2.4 Análise de Custo-Benefício

Através da análise preditiva, a empresa terá a possibilidade de antever previamente erros, falhas técnicas ou desgaste de peças que poderão ocorrer com a maquinaria da unidade fabril. Assim, será possível, por exemplo, antecipar a necessidade de troca de uma peça, antes que ela se desgaste e que a necessidade de troca da mesma possa impedir o cumprimento de prazos de entregas, fazendo com que, para a situação ficar resolvida e a máquina poder voltar a trabalhar, seja um processo demorado que prejudique o negócio.

Os principais benefícios da implementação do novo sistema são efetivamente a diminuição e prevenção de erros ou falhas e consequentemente a diminuição de custos de manutenção e reparação. Relativamente aos custos, verifica-se que são custos iniciais: a contratação de técnicos da área e a aquisição inicial, formação de empregados e manutenção/atualização do *software* utilizado.

Na tabela seguinte são apresentados todos os custos *versus* benefícios associados à implementação do projeto em causa:

|  |  |  |
| --- | --- | --- |
| Custo | Benefício | Descrição |
| Alocação de recursos humanos e possível perda de produtividade durante processo de implementação | Diminuição de tempo despendido a resolver problemas relativos à maquinaria | Para implementação deste projeto terá que ser despendido bastante tempo de produção para dedicar à sua implementação. |
| *Stock* de materiais | Diminuição de custos de manutenção | Conseguindo prever o desgaste de algum equipamento é possível antecipar a sua necessidade para que funcione e se encontre em ótimo estado de utilização, reduzindo o tempo de inatividade. |
| Contratação de técnicos da área e/ou serviços externos | A prestação deste serviço através de contratação externa | Se a contratação de um consultor externo delega parte do trabalho deste projeto tecnicamente desafiante, também comporta custos e dependência de terceiros. |
| Formação de profissionais | Profissionais com os conhecimentos certos são menos propícios ao erro na utilização dos *softwares* | Todos os profissionais envolvidos nos processos que se caracterizam pela utilização da maquinaria devem ter conhecimentos sobre o *software* em utilização. |
| Manutenção e atualização do sistema | Tecnologias atualizadas permitem o aparecimento de novas ferramentas. | Os *softwares* são tecnologias que precisam de estar sempre atualizadas para que não existam erros ou falhas. |
| Implementação | Diminuição de erros ou falhas | O processo de implementação poderá ser demorado. |

**Tabela 2 -** Resumo dos Custos e Benefícios da Implementação do Projeto de *Data Mining*

## 2.5 Descrição do Contexto

Esta unidade fabril tem como principal serviço não só a comercialização dos seus produtos, mas também a terceirização, através da criação e desenvolvimento de marcas e produtos para clientes.

Os seus objetivos são maioritariamente focados no lucro de receitas e na produção ao custo mais reduzido possível, para definir com maior clareza os objetivos da empresa em questão será utilizada a análise dos objetivos SMART (*Specific, Measurable, Achievable, Realistic and Time Based*), que são uma ferramenta que auxilia as pessoas ou empresas a definirem os seus objetivos e metas. Doran, G. T. (1981) afirma, no artigo “*There's a S.M.A.R.T. way to write management's goals and objectives*", que os objetivos que seguem os critérios SMART têm maior possibilidade de serem bem-sucedidos, pois estes ajudam uma empresa ou pessoa a definir um futuro claro, orientando as ações para o que realmente importa fazer.

Assim, no contexto empresarial, para alavancar um negócio a atingir os melhores resultados, é importante definir objetivos claros e estratégicos que permitam alcançar o sucesso dos negócios. As principais vantagens de ter objetivos delineados e bem definidos são:

* Definir com clareza e preparar a empresa para situações futuras, de forma a projetar os esforços necessários;
* Direcionar toda a equipa para os mesmos objetivos;
* Motivar a equipa para alcançar os resultados pretendidos;
* Medir a eficiência das ações implementadas.

Em seguida são apresentados os objetivos SMART que a empresa pretende alcançar:

* **Específicos**

Os objetivos específicos referem-se ao que se pretende alcançar e o que se irá fazer para se conseguir alcançar, estes devem ser o mais específico e precisos possíveis.

O principal objetivo que a empresa pretende alcançar é o aumento de pelo menos 20% da produção no espaço de 3 anos, e consequentemente aumentar os lucros da empresa. Para fazer face a esse crescimento esperado, a empresa terá que optar pela contratação de novos colaboradores, a fim de conseguir dar resposta ao crescimento da produção.

* **Mensuráveis**

Os objetivos mensuráveis baseiam-se na parte numérica dos objetivos, traduzindo-se assim nas qualidades necessárias, nas quantidades desejadas e no seu custo associado.

Com a implementação do projeto em causa, espera-se que exista uma redução de tempo despendido na manutenção da maquinaria, assim como do tempo despendido em caso de avarias ou problemas técnicos, sendo este um dos grandes objetivos da empresa, reduzir o tempo de inatividade por falhas ou erros técnicos. Assim, o desejável é que os erros e falhas ocorrentes possam ser antecipados e evitados ou rapidamente solucionados.

* **Alcançáveis**

Ao definir objetivos SMART, é importante ter em conta se os objetivos que se pretendem alcançar são de facto possíveis de atingir ou não. É fundamental perceber e investigar a aceitabilidade e tangibilidade dos mesmos, considerando o tempo, esforço e o custo associado que o objetivo terá, e por fim perceber se haverá lucro sobre o investimento em causa.

Tendo em conta a viabilidade do projeto e a possibilidade da sua implementação, espera-se que a implementação deste novo sistema contribua para atingir todos os objetivos pretendidos.

* **Relevantes**

A relevância dos objetivos define com certeza se estes são ou não importantes para a empresa, assim como, se existe um alinhamento com outros objetivos relevantes no mercado.

Os objetivos que a empresa pretende alcançar são bastante relevantes para o sucesso e evolução do negócio.

* **Temporais**

Os *deadlines* são também fundamentais para alcançar com sucesso os objetivos definidos. Prevê-se que os objetivos anteriormente mencionados possam ser alcançados no prazo de 3 anos conforme pretende a empresa.

Tendo em conta a rápida atualização do mercado e com o avanço das tecnologias, as empresas, cada vez mais, precisam de se preocupar com a inovação e com a concorrência. Numa empresa com a unidade fabril em estudo, é importante que existam processos delineados que permitam facilitar o dia-a-dia, assim como reduzir os custos operacionais e aumentar a eficiência.

Num mercado tão competitivo como o setor dos cosméticos, em que os clientes se tornam cada vez mais exigentes com a qualidade e rapidez de entrega do produto, é importante ter em consideração todos os passos dos concorrentes em relação à inovação da tecnologia utilizada e das estratégias implementadas.

O projeto em causa posiciona-se como uma resposta e solução a uma das maiores fraquezas da empresa, os erros e falhas inesperados que podem provocar atrasos e prejuízos na produção.

O sucesso deste projeto será medido pela diminuição de falhas e erros na maquinaria, assim como na diminuição de atrasos na entrega de produtos aos clientes.

A avaliação dos modelos em causa será da responsabilidade do Gestor de Produção, assim como do responsável pela expedição do produto através das métricas acima referidas.

## 2.6 Planeamento do Projeto

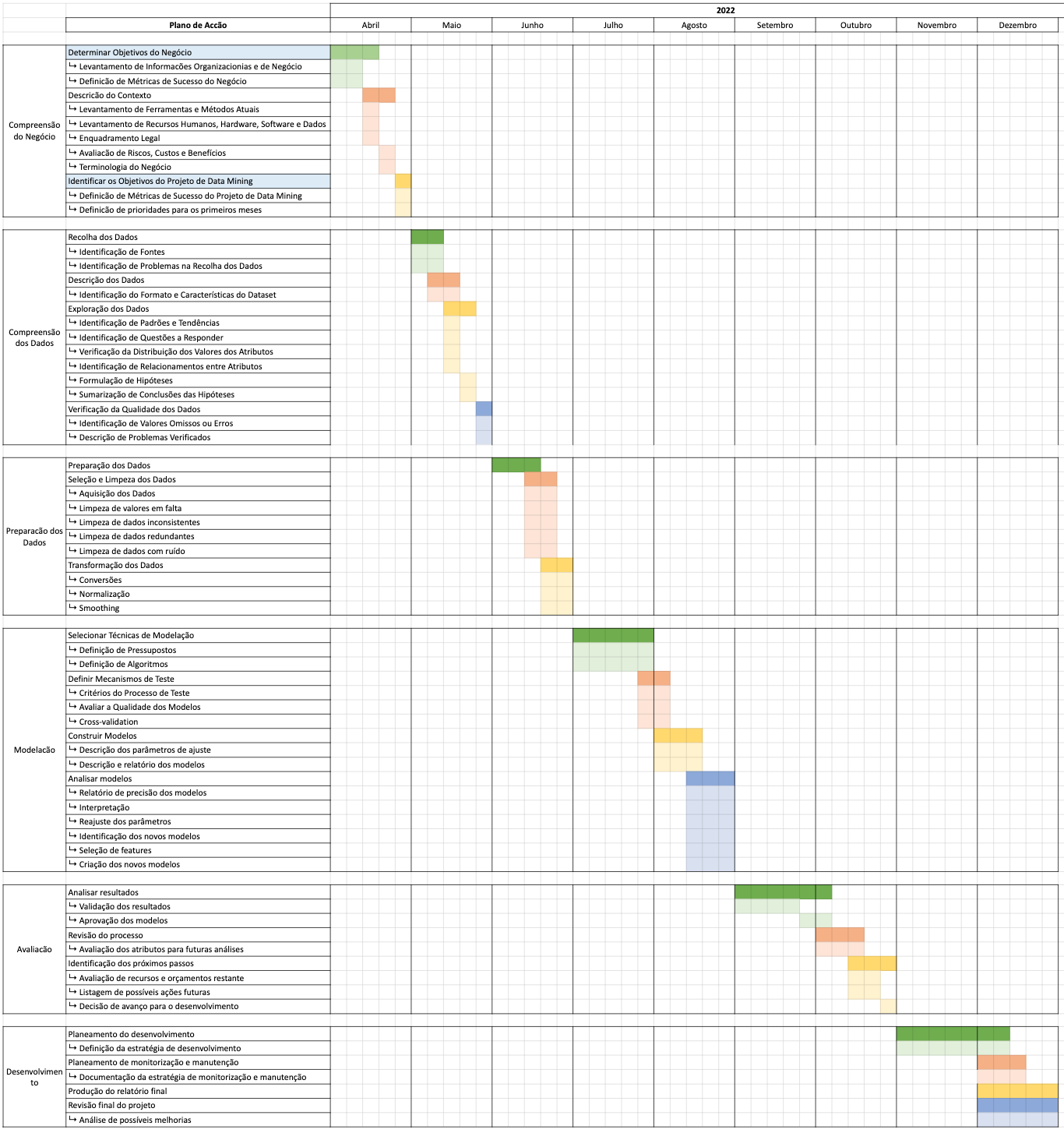
### 2.6.1 Atividades e Subatividades

O planeamento do projeto seguirá a Metodologia CRISP-DM, um modelo de referência que define as fases a seguir, tarefas a implementar e resultados esperados em cada uma das fases do projeto de *data mining*, como referido anteriormente.

Foi executado um plano a 9 meses (abril a dezembro de 2022), onde são identificadas todas as tarefas a desenvolver ao longo do tempo, e os respetivos tempos despendidos para desenvolver cada uma dessas tarefas. Para este planeamento, recorremos a um *RoadMap*, dividido entre as 6 fases do projeto, nomeadamente: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e desenvolvimento.

A fase inicial de compreensão do negócio e dos dados correspondem aos conteúdos abordados respetivamente nos pontos 2 e 3 deste trabalho. Os restantes tópicos serão desenvolvidos na 2ª parte deste projeto, e prevê-se que a preparação e modelação dos dados termine apenas no fim do mês de agosto, e que as restantes atividades se prolonguem até ao fim do ano em causa.

Na tabela seguinte são apresentadas todas as atividades e subatividades correspondentes ao projeto, através do *RoadMap* desenvolvido, sendo que este pode sofrer alterações ao longo do projeto, uma vez que se encontra em constante atualização.



**Tabela 3 –** Plano de atividades das 6 fases do Projeto de *Data Mining*

Tendo em conta a área de atuação onde se pretende implementar o projeto, as ferramentas e tecnologias a utilizar ainda se encontram em fase de estudo, estando algumas delas já definidas. A análise exploratória, a transformação, limpeza e normalização dos dados serão feitas com recurso à linguagem de programação Python, utilizando *Jupyter notebooks* e, quando necessário, dado que estamos a trabalhar com dados tabulares, o *software* de folha de cálculo Excel ou Libre Office Calc. Contamos desenvolver e testar modelos de aprendizagem automática com recurso, por exemplo, às bibliotecas Python *scikit-learn* ou XGBoost mas, estando ainda numa fase precoce do projeto, outras opções podem ser consideradas, havendo ainda abertura para utilização de outras ferramentas, como é o caso da plataforma *RapidMiner*. Consideraremos com particular atenção modelos de aprendizagem automática explicáveis, pelas razões explicitadas na primeira secção deste trabalho.

Findas as fases de Modelação e Avaliação dos modelos, é nosso objetivo implementar uma solução cujo armazenamento e execução seja feita na infraestrutura de computadores local. Numa fase mais adiantada do projeto, em conjunto com a administração da empresa, serão consideradas as opções mais em detalhe.

# 

# 3. Compreensão dos Dados

Nesta segunda seção do trabalho, seguindo o esquema descrito no guia CRISP-DM, começaremos por descrever como foi feita a recolha dos dados, as suas fontes e assim como eventuais problemas daí resultantes. De seguida, faremos a descrição detalhada das características do conjunto de dados recolhido e, numa fase posterior, por meio de técnicas de visualização e consulta de dados, estudaremos as propriedades dos atributos-chave, as relações entre eles e elaboraremos algumas descobertas e hipóteses iniciais por forma a indicar possíveis orientações para as fases posteriores do projeto e prossecução dos objetivos a que nos propusemos. Finalmente, faremos a análise da qualidade dos dados, incluindo se existem dados omissos, se contém erros, assim como se estão corretamente documentados.

## 3.1 Recolha dos Dados

Uma vez que os verdadeiros conjuntos de dados de manutenção preditiva são geralmente difíceis de obter, o autor propôs-se a criar um conjunto de dados sintéticos que reflete, tanto quanto possível, a manutenção preditiva real encontrada na indústria, que apelidou de AI4I 2020 Predictive Maintenance Dataset (Matzka 2020).

Este conjunto de dados está disponível para ser descarregado em vários sítios web, mas a sua fonte original é o repositório de conjuntos de dados de aprendizagem automática da Universidade da Califórnia em Irvine. O ficheiro original destes dados é de texto delimitado por vírgulas, comummente denominado *Comma-separated values* ou CSV, de fácil leitura num software de folha de cálculo como o Microsoft Excel ou Libre Office Calc. Este ficheiro tem pouco mais de 500 Kilobytes, o que facilita o seu estudo, manipulação e armazenamento (tabela 4).

| Formato do ficheiro | \*.csv |
| --- | --- |
| Tamanho do ficheiro | ≅ 510 Kilobytes |
| Fonte dos Dados | [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset) |
| Data de *download* dos dados | 28/03/2022 |

**Tabela 4 -** Fonte, formato e tamanho do ficheiro do conjunto de dados

## 3.2 Descrição dos Dados

Antes de avançarmos para a exploração dos dados, é necessário ter um entendimento perfeito de cada uma das variáveis que o constituem. Nomeadamente, perceber o seu significado, tipos de dados, unidades de medida entre outras informações relevantes. Nesse aspeto, cremos que o conjunto de dados está relativamente bem documentado, o que facilitou a recolha das informações da tabela 5, em baixo.

O conjunto de dados é composto por 10.000 ocorrências e estão armazenados numa tabela com 14 variáveis (“[AI4I 2020 Predictive Maintenance Dataset](https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset)” 2020):

| # | Designação da variável | Descrição | Tipo de Dados | Unidade de medida | Exemplo |
| --- | --- | --- | --- | --- | --- |
| 1 | **udi** | Identificador único de ocorrência (de 1 a 10000) | Número Inteiro | N/A | *1320* |
| 2 | **product\_ID** | Número de identificação ou número de série do produto | Alfanumérico | N/A | *M16180* |
| 3 | **type** | Letra **L** (*Low)* que indica qualidade de produto baixa, **M** (*Medium)* que indica qualidade de produto média e **H** (*High)* para produto de alta qualidade | Caractere | N/A | *M* |
| 4 | **air\_temperature** | Valores gerados usando um processo de passeio aleatório[[3]](#footnote-4) mais tarde normalizado para um desvio padrão de 2 Kelvin em torno de 300 Kelvin | Número com 1 casa decimal | Graus Kelvin | *299.0* |
| 5 | **process\_temperature** | Valores gerados usando um processo de passeio aleatório normalizado a um desvio padrão de 1 Kelvin, adicionado ao ar temperatura mais 10 Kelvin | Número com 1 casa decimal | Graus Kelvin | *310.3* |
| 6 | **rotational\_speed** | Valores calculado a partir de uma potência de 2860 Watt, sobreposto de ruído normalmente distribuído | Número Inteiro | Rotações por Minuto [rpm] | *1540* |
| 7 | **torque** | Os valores de binário são normalmente distribuídos em torno de 40 Nm com um desvio padrão de 10 Nm, sem valores negativos | Número com 1 casa decimal | Newton-metro [Nm] | *38.1* |
| 8 | **tool\_wear** | As variantes de qualidade H/M/L acrescentam 5/3/2 minutos de desgaste na ferramenta utilizada no processo | Número Inteiro | minutos [min] | *182* |
| 9 | **machine\_failure** | Indica se houve falha na máquina. Se pelo menos um dos modos de falha [#10 a #14] for verdadeiro (= 1) , o processo falha e a variável **machine\_failure** muda de estado (0 → 1) | Binário | N/A | *0* |
| 10 | **tool\_wear\_failure (TWF)** | Indica que ferramenta falha de forma aleatória quando tem desgaste (**tool\_wear**) entre 200 e 240 minutos | Binário | N/A | *0* |
| 11 | **heat\_dissipation\_failure (HDF)** | A dissipação de calor provoca falha no processo, se a diferença entre a temperatura do ar (**air\_temperature**) e a temperatura do processo (**process\_temperature**) for abaixo de 8,6 Kelvin e a velocidade de rotação da ferramenta inferior a 1380 rpm | Binário | N/A | *0* |
| 12 | **power\_failure (PWF)** | Se a potência, ou seja, o produto do binário (**torque**) com a velocidade rotacional (**rotational\_speed**), for inferior a 3500 Watt ou superior a 9000 Watt, o processo falha | Binário | N/A | *0* |
| 13 | **overstrain\_failure (OSF)** | Se o produto do desgaste da ferramenta (**tool\_wear**) e o binário (**torque**) da ferramenta exceder 11000 min∙Nm para a variante de produto L, 12000 para M e 13000 para H, o processo falha | Binário | N/A | *0* |
| 14 | **random\_failure (RNF)** | Para além dos modos de falha #10 a #13, a máquina tem uma chance de 0,1 % de apresentar falha aleatória, independentemente dos parâmetros que causam as falhas descritas em cima | Binário | N/A | *0* |

**Tabela 5 -** Descrição dos variáveis do conjunto de dados

Tal como descrito na tabela 5, o autor do conjunto de dados sintético estabeleceu uma relação de causalidade entre as grandezas medidas pelos vários sensores e as falhas da máquina.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sensores** | | | → | **Falha na Máquina** |
|  |  |  |  |
| Temperatura Ar | Temperatura Processo | Velocidade Rotação ***ω*** | Falha de Dissipação de Calor |
| **𝚫Temp** = Temp. Processo - Temp. Ar **< 8,6 ℃** ∩ ***ω* < 1380 rpm** | | |
|  | | | | |
|  | | → | |  |
| Desgaste da Ferramenta | | Falha de Desgaste da Ferramenta |
| Desg. Fer. entre 200 e 240 minutos | |
|  | | | | |
|  |  | → | |  |
| Velocidade Rotação ***ω*** | Binário *𝛕* | Falha de Potência Elétrica |
| ***ω*** x *𝛕* **<** 3500 W ∪ ***ω*** x *𝛕*  **>** 9000 W | |
|  | | | | |
|  |  | → | |  |
| Desgaste Ferramenta | Binário *𝛕* | Falha de Fratura por Fadiga |
| Desgaste x *𝛕* > 11000 min∙Nm (produtos tipo **L**)  Desgaste x *𝛕* > 12000 min∙Nm (produtos tipo **M**)  Desgaste x *𝛕* > 13000 min∙Nm (produtos tipo **H**) | |
|  | | | | |
| Máquina tem **0,1 %** de chance de falhar independentemente dos parâmetros dos sensores | | Não existe causalidade | |  |
| Falha Aleatória |

**Tabela 6 -** Causas das falhas da máquina

Para que fique claro e para facilitar a análise e exploração dos dados, na tabela 5 desenhou-se um esquema que representa as causas das falhas da máquina, tal como definido pelo criador do conjunto de dados. Se pelo menos uma destas condições de falha for verdadeira, a máquina sofre uma falha.

O facto de nalgumas ocorrências haver mais de um tipo de falha, introduz ambiguidade e nem sempre será transparente para o modelo de *Data Mining* qual é a verdadeira causa da falha. Simultaneamente, após análise dos dados relativos às causas de falha, verifica-se que em 9 ocorrências, apesar da variável *machine\_failure* (falha da máquina) ser verdadeira, não é indicado qualquer tipo de falha, sendo desconhecido a origem da falha por omissão (tabela 7). Importa referir que estas incongruências foram introduzidas propositadamente no conjunto de dados (Matzka 2020).

| **Número de tipos de falha** | **Número de ocorrências** | **Exatidão da causa** |
| --- | --- | --- |
| 1 Falha | 309 | Causa definida e unívoca |
| 2 ou 3 Falhas | 23 | Causa da falha ambígua |
| Nenhuma falha | 9 | Falhas sem causa ou causa aleatória |

**Tabela 7 -** Tipos de falha quando variável **machine\_failure** é verdadeira

## 3.3 Exploração dos Dados

Após a descrição dos dados, como primeira abordagem calculou-se uma tabela de sumário com medidas de tendência central e de distribuição de frequências (percentis) de todas a variáveis numéricas. Como as variáveis relativas às falhas são binárias e, em certa medida, indicam categorias, optou-se por omiti-las nesta tabela.

|  |
| --- |
|  |

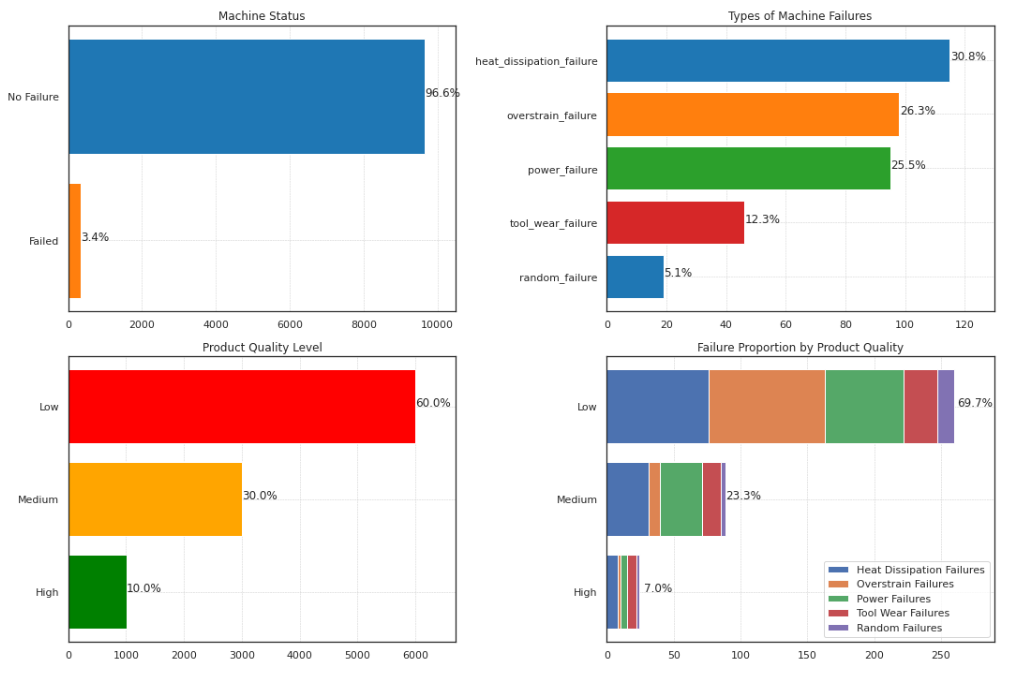
**Tabela 8 -** Medidas de tendência central e de distribuição de frequências

Destacadas a amarelo, estão as variáveis relativas aos sensores. Particularizando para as variáveis de temperatura - que foram convertidas de graus Kelvin para graus Celsius - em média, a temperatura do processo é 10 graus superior à temperatura média do ar, sendo que a variabilidade da temperatura do ar é superior. Quanto à velocidade de rotação, tem uma amplitude interquartil (P75% - P25%) de 444 rpm e média de ≅ 1539 rpm.

Quanto ao binário ou torque, apresenta um valor de média e mediana muito próximo de 40 Nm. A simetria da função de distribuição é evidente quando se traça o histograma.

Finalmente, o desgaste de ferramenta é a única variável que pela sua natureza tem mínimo de 0 minutos, atingindo um valor máximo de 253 minutos.

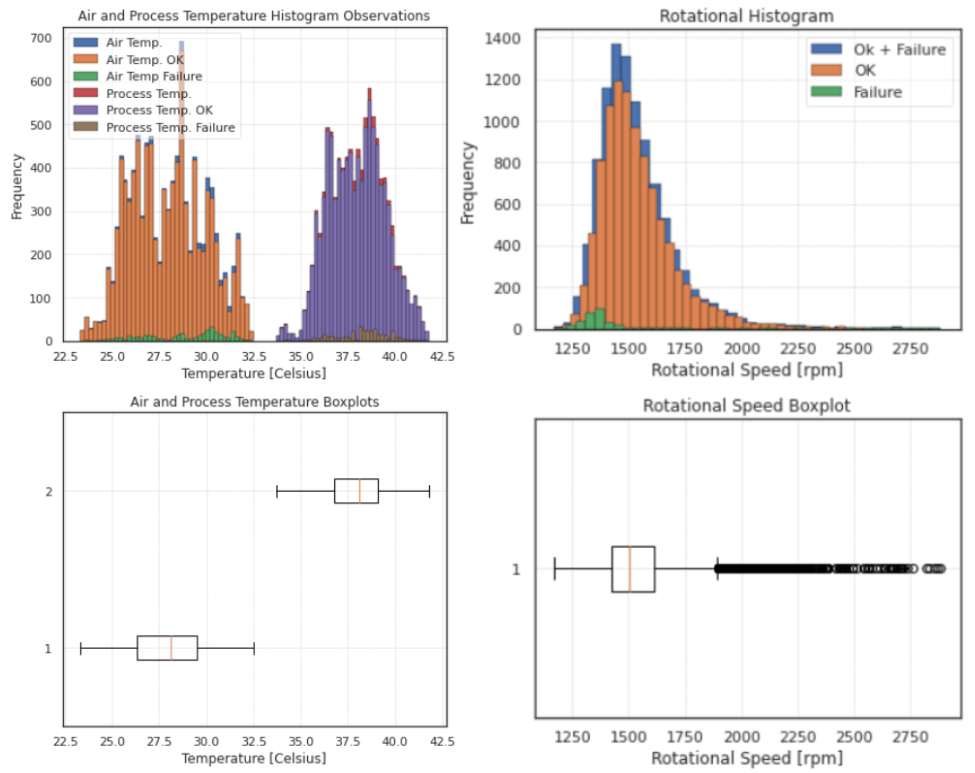
De seguida, analisamos algumas relações que se consideraram relevantes nas variáveis que indicam as falhas e a qualidade dos produtos. Começando pelo gráfico de barras do canto superior esquerdo da figura 5 relativo ao estado de operação da máquina, conclui-se que os dados relativos às situações de falha são de apenas 3,4% de um total de 10000 ocorrências, o que corresponde a um total de 339 entradas com falha na máquina. Examinando o gráfico imediatamente à direita, conclui-se que a maioria das falhas ocorreu devido a falha de dissipação de calor, com 32,4% do total de falhas, o que corresponde a 115 ocorrências. Com valores próximos, as falhas de fratura por fadiga e falha de potência elétrica, com 98 e 95 ocorrências respetivamente, formam um segundo grupo de causas de falha.



**Figura 5 -** Gráficos de barras relativos às falhas e qualidade dos produtos

No gráfico do canto inferior esquerdo podemos observar que a proporção de Produtos de baixa qualidade é de 60%, de média qualidade 30%, e as de alta qualidade representam apenas 10% dos produtos produzidos. Esta observação levou-nos a traçar um quarto gráfico que nos indicasse se a qualidade dos produtos está relacionada com a quantidade e tipo de falhas. O gráfico do canto inferior direito, indica-nos que as peças de qualidade inferior apresentam uma taxa de falha ligeiramente superior às de qualidade média e superior, talvez devido ao facto das falhas de fratura por fadiga terem o limite menor para produtos de qualidade inferior (tabela 5).

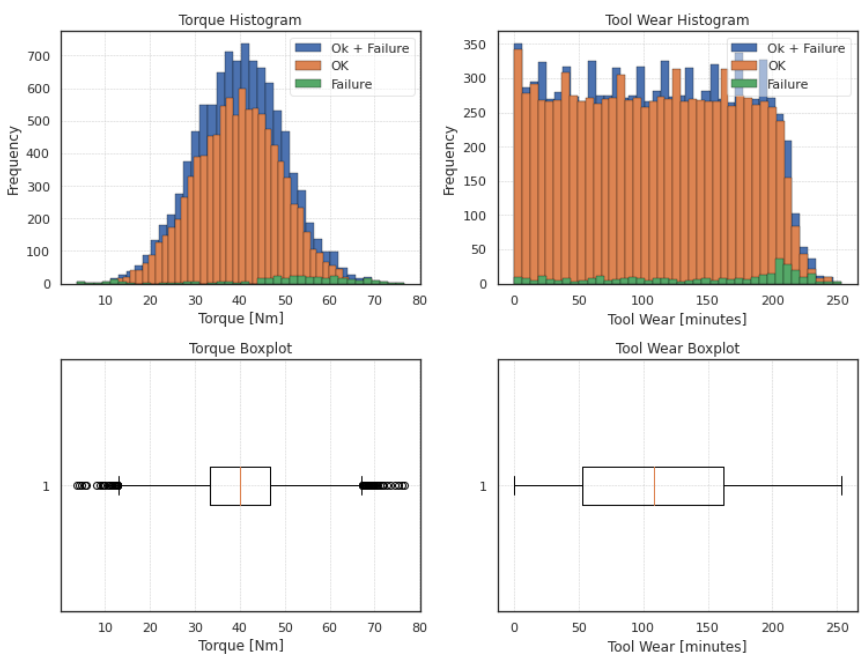
Para termos perceção das distribuições de frequência de cada uma variáveis numéricas dos sensores, traçaram-se os histogramas e os gráficos caixa-de-bigodes. Para o efeito de deteção de anomalias, partindo do pressuposto que não houve erro de medição e/ou inserção e dos dados, os *outliers* não foram descartados ou ignorados pois são os eventos atípicos podem ser os que explicam as falhas.



**Figura 6 -** Histogramas e caixas-de-bigodes das variáveis da Temperatura e Velocidade Rotacional

Os histogramas foram desenhados usando 3 camadas: uma primeira onde são incluídos todos os pontos (com e sem falha), uma segunda só com os pontos em que o estado de funcionamento da máquina é normal, e uma terceira representando as frequências para os pontos de falha. Deste modo, podemos ter alguma perceção de que zonas ocorrem a maioria das falhas.

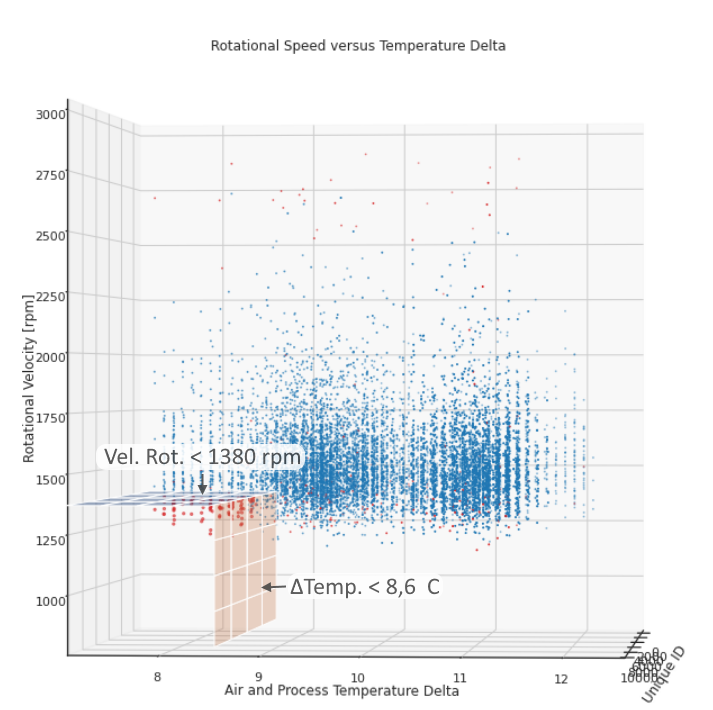
Analisando os histogramas e caixas-de-bigodes das variáveis de temperatura do ar e do processo da figura 6, ambas parecem ter uma distribuição aproximada à distribuição normal com aparente simetria. Quanto ao histograma da variável da velocidade rotacional, este apresenta assimetria positiva. Esta variável apresenta *outliers* superiores moderados e severos, sendo o ponto máximo 2886 rpm. Observa-se, no entanto, que as falhas se concentram mais nos pontos inferiores à mediana.



**Figura 7 -** Histogramas e caixas-de-bigodes das variáveis de Binário e Desgaste da Ferramenta

Quanto às variáveis de Torque ou Binário e Desgaste da Ferramenta - à esquerda e direita da figura 7, respetivamente -, verificamos que a distribuição de frequências do Binário é aproximadamente normal e simétrica, com média e mediana aproximadamente em 40 Nm. As falhas parecem ser mais frequentes para valores superiores à média. Relativamente à variável de desgaste, apresenta uma distribuição que se pode caracterizar como uniforme, com um limite superior aproximadamente em 220 minutos. As falhas estão concentradas sobretudo nos pontos com valor de desgaste superior a 200 minutos.

Na seção que se segue, tomando como base as informações da tabela 6, procurou-se traçar gráficos de dispersão representativos de cada um dos tipos de falha. Os pontos com falha (*machine\_failure* = 1) foram destacados a cor vermelha e as regiões ou áreas cujas condições provocam avaria da máquina, são destacadas num tom avermelhado. Para o caso em que 3 variáveis estão envolvidas na causa de falha da máquina, delimitaram determinadas áreas através de planos.



**Figura 8 -** Gráfico de dispersão 3D da Velocidade de Rotação e diferença de Temperatura do Ar e do Processo

Começando esta análise pela falha de dissipação de calor, a figura 8 representa um gráfico de dispersão tridimensional, com o eixo x representando o ID da entrada, o eixo y a diferença entre a temperatura do processo e a temperatura do ar e o eixo z a velocidade de rotação. Tal como vimos em cima na seção 3.2 Descrição do Dados, quando a velocidade de rotação é menor que 1380 rpm e a diferença entre a temperatura do processo e a temperatura do ar menor que 8,6 ℃, produz-se uma falha de dissipação de calor. Observando o gráfico da figura 8, torna-se evidente que existe de facto uma concentração de pontos de falha (a vermelho) nesta região.

Fazendo o mesmo exercício para o gráfico de dispersão relativo ao desgaste da ferramenta da figura 9, podemos observar que na área destacada a vermelho, que compreende o intervalo de 200 a 240 minutos, parece haver uma concentração maior de pontos a vermelho (falhas) quando comparado com as zonas em que o desgaste da ferramenta é menor que 200, tal como vimos no histograma da figura 7, aliás.



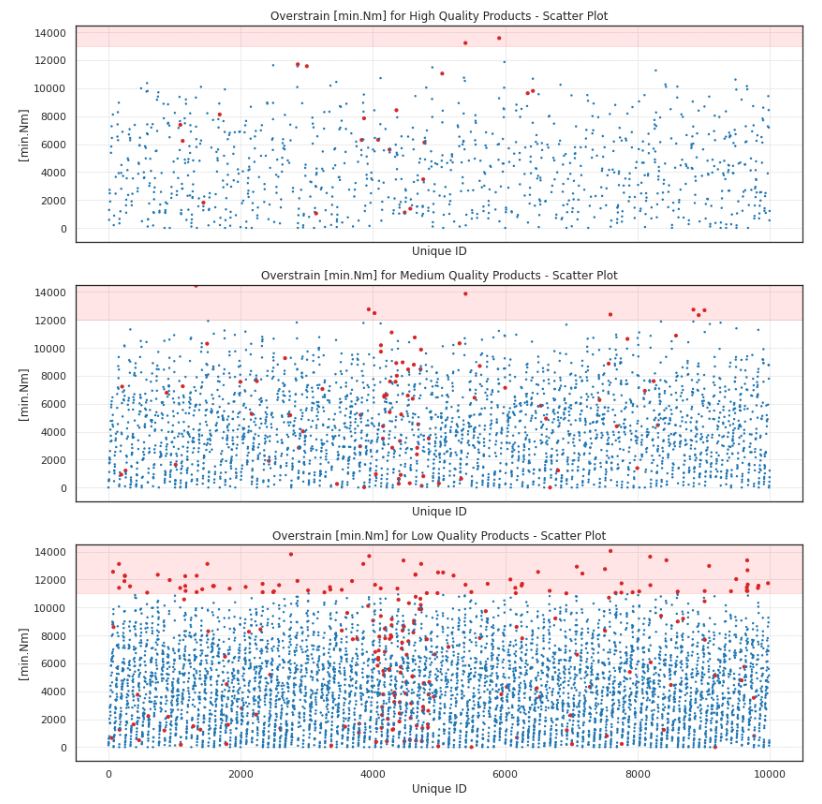
**Figura 9 -** Gráfico de dispersão do Desgaste da Ferramenta

As falhas de Potência Elétrica acontecem quando se ultrapassa os 9600 Watt e quando o valor é menor que 3600 Watt. Estas duas regiões estão representadas a vermelho no gráfico de dispersão da figura 10. Pela observação do gráfico apenas, parece-nos que a quase totalidade de ocorrências nesta região estão a vermelho, isto é, representam falhas da máquina.



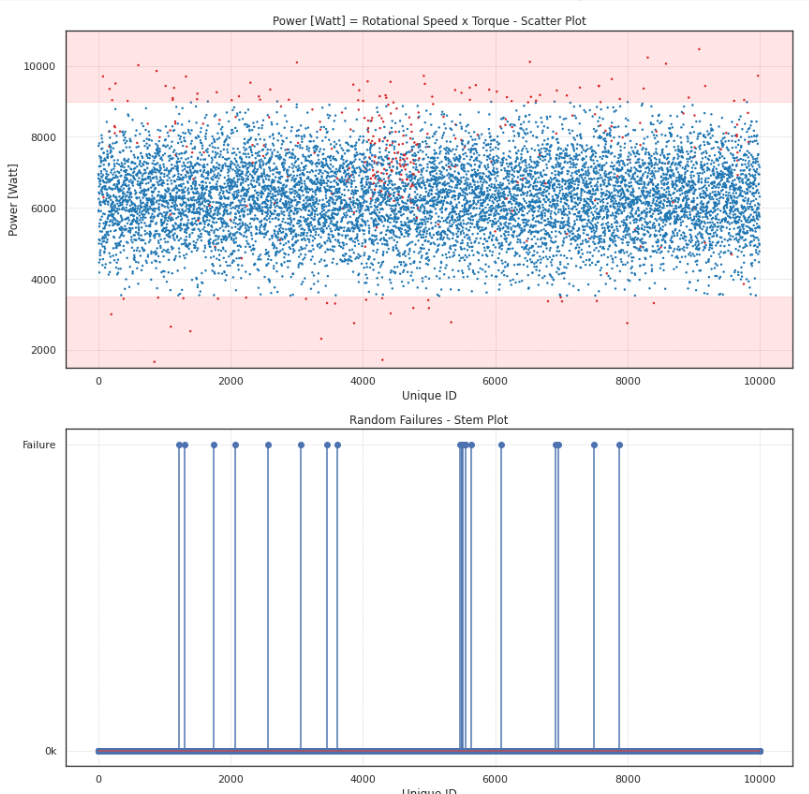
**Figura 10 -** Gráfico de dispersão da Potência Elétrica

Quanto às falhas de Fratura por Fadiga (figura 11), dependendo da qualidade do produto, têm limites diferentes de falha. Para os produtos com qualidade superior (gráfico mais acima), o nível a partir do qual existem falhas é 13000 min∙Nm. Se o produto for de qualidade média, o limite é de 12000 min∙Nm, e para produtos de qualidade inferior, apenas 11000 min∙Nm.



**Figura 11 -** Gráfico de dispersão da Falhas por Fratura por Fadiga para os diferentes níveis de qualidade dos produtos

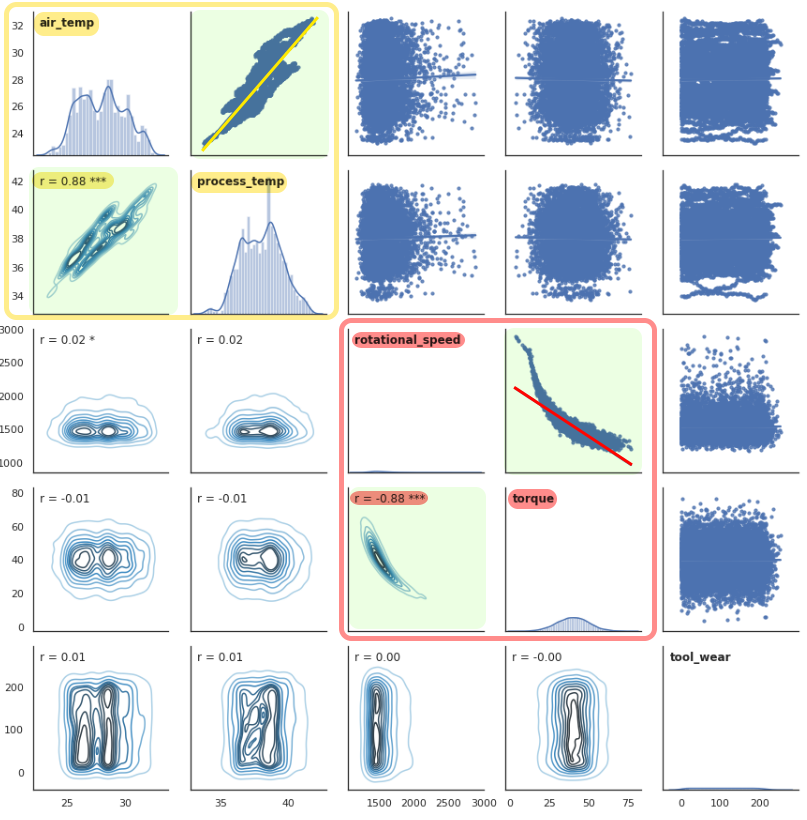
A última causa de falha abordada na tabela 5 foi a Falha Aleatória. Esta falha não é explicada por nenhuma das variáveis dos sensores e, por essa razão, não faz sentido traçar um gráfico que os relacione. Optou-se por traçar um gráfico da variável binária *random\_failure* (figura 12). No total, a Falha Aleatória ocorre por 19 ocasiões em 10000.



**Figura 12 -** Gráfico variável binária **random\_failure**

Por último, construiu-se uma matriz de correlação com o fim de sumarizar as correlações entre todas as variáveis que resultam dos sensores, observar relações que à partida não consideramos e também como elemento de partida para eventual utilização de técnicas de redução de dimensionalidade como a análise de componentes principais (PCA).

A figura 13 para além de representar uma matriz de correlação (coeficiente de Pearson), traça os gráficos de dispersão, histogramas e gráficos de densidade bivariada pelo método do núcleo, oferecendo no nosso entender, muitos elementos interessantes para analisar. Das 5 variáveis representadas, apenas a temperatura do ar e temperatura do processo e o binário e a velocidade rotacional apresentam valores de coeficiente de correlação de Pearson superiores a |0,02|.



**Figura 13 -** Matriz de correlação das variáveis dos sensores

|  |
| --- |
|  |

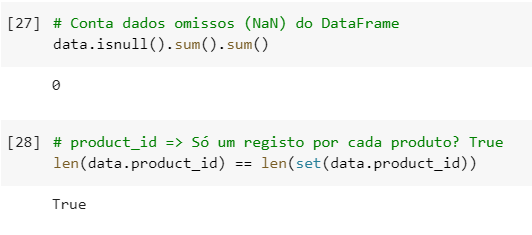
Sem surpresa, as temperaturas do ar e do processo estão fortemente correlacionadas de forma positiva e linear, com valor de coeficiente de Pearson r = 0,88. Do mesmo modo, o binário a velocidade rotação também estão fortemente correlacionadas de forma negativa (r = -0,88). Pode-se dizer, que tendencialmente, quando a velocidade de rotação aumenta o valor do binário diminui, o que mais uma vez faz todo o sentido, se considerarmos uma potência constante. O gráfico sugere que a relação entre estas duas variáveis não é linear.

## 3.4 Qualidade dos Dados

A última tarefa da fase de Compressão de Dados, será verificar a qualidade e descrever os problemas encontrados e as soluções para esses problemas. À medida que a quantidade de dados vai crescendo, maior será a necessidade de criar regras e padronizações que permitam facilitar a gestão da qualidade de dados. As questões da qualidade dos dados prendem-se por exemplo com a omissão de dados, erros de inserção, duplicação de entradas, (in)coerência entre várias fontes de dados, falta de documentação dos dados (metadados), origem desconhecida (linhagem) ou ainda questões relativas à legalidade da utilização dos dados.

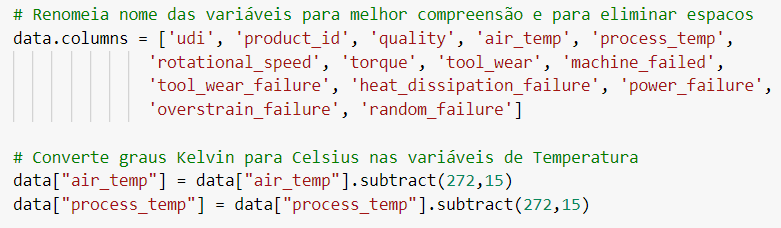
Para o nosso estudo, a origem e os metadados estão bem documentados e a dimensão e diversidade de fontes de dados é limitada pelo que não encontrámos problemas de difícil resolução. Ainda assim, fizemos alguns testes e encontrámos algumas inconsistências que passamos a descrever.

O conjunto de dados não apresenta dados omissos em nenhuma das variáveis e não tem entradas duplicadas para o mesmo produto.



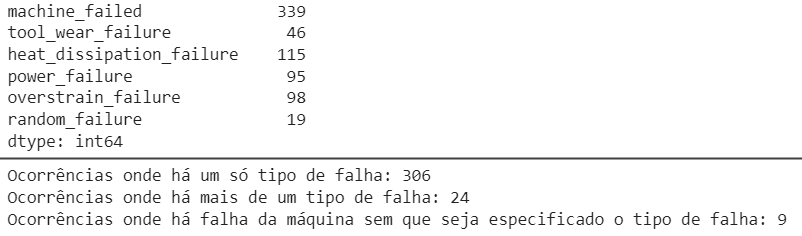
**Figura 14 –** Excerto do script Python relativo a dados omissos e entradas duplicadas

Fizemos uma renomeação das variáveis pois apresentavam espaços e tinham nomes algo inconsistentes. Para além disso, para facilitar a interpretação, convertemos as variáveis de temperatura de graus Kelvin para graus Celsius.



**Figura 15 -** Excerto do script Python para renomeação e conversão das variáveis

No entanto, este conjunto de dados apresenta uma inconsistência que está documentada pelo autor (Matzka 2020). Tal como exposto na tabela 4 da secção 2.2 Descrição de Dados, em 9 das ocorrências de falha (*machine\_failure* = 1) não é especificado o tipo de falha; em 24 ocorrências de falha, mais de um tipo de falha tem valor 1, o que pode ser origem de ambiguidade.



**Figura 16 –** Captura da tela do resultado da contabilização do número e categoria de falha

Ainda assim, talvez por se tratar de um conjunto de dados sintético, aberto e com origem académica, conclui-se que a qualidade dos dados é bastante aceitável.

# 4. Início do Projeto Prático

Esta segunda parte do projeto é focada essencialmente na parte mais prática, nomeadamente, na Preparação de Dados, Modelação, Avaliação dos modelos criados, e Colocação do Modelo em Produção, através da Metodologia CRISP-DM.

Na fase de Preparação dos Dados, começaremos por efetuar a sua seleção e limpeza e documentar os motivos de seleção ou exclusão dos mesmos. Caso necessário criaremos novos registos e formataremos esses novos dados adicionais de forma a ficarem compatíveis com os restantes, a fim de desenvolvermos o nosso Conjunto de Dados final.

Na Modelação dos dados, após escolhermos os modelos a testar ao longo do projeto, planificaremos a divisão do Conjunto de Dados nos respetivos conjuntos de treino, validação e teste e construiremos o método de busca de hiperparâmetros, descrevendo-o. Por fim, iremos avaliar esses mesmos modelos e fazer as afinações finais.

Na fase da Avaliação, utilizaremos os critérios de sucesso anteriormente definidos para avaliar os resultados do modelo e determinaremos ações futuras ou redefinições de qualquer problema ou questão que possa surgir.

A última fase será para fazer os últimos ajustes, colocar o plano em produção e planear a monitorização e manutenção do mesmo.

## 4.1 Preparação dos Dados

Neste ponto do projeto iremos dar início à preparação dos dados, que serão utilizados para a modelação e para o trabalho de análise principal do projeto.

Nesta fase serão feitos processos de seleção, limpeza, construção, formatação e integração dos dados, para chegarmos a um conjunto de dados final, para utilizarmos ao longo deste estudo. Todo este processo é feito recorrendo a *Jupyter Notebooks* e programas de folhas de cálculo.

## 4.2 Seleção dos Dados

As variáveis utilizadas no projeto foram selecionadas de acordo com a sua relevância para o estudo em causa. O conjunto de dados é composto por 10.000 ocorrências e estão armazenados numa tabela com 14 variáveis (*AI4I 2020 Predictive Maintenance Dataset*, 2020).

Para o nosso caso, por serem variáveis identificadoras e não terem influência na ocorrência de falhas, não se consideraram os atributos *udi* (identificador único) e *product\_ID* (código alfanumérica). Deste modo, foram incluídas 5 variáveis de entrada, destacadas a verde-claro na Tabela 1 e 6 variáveis objetivo ou de saída, a azul-claro, perfazendo um total de 11 de variáveis.

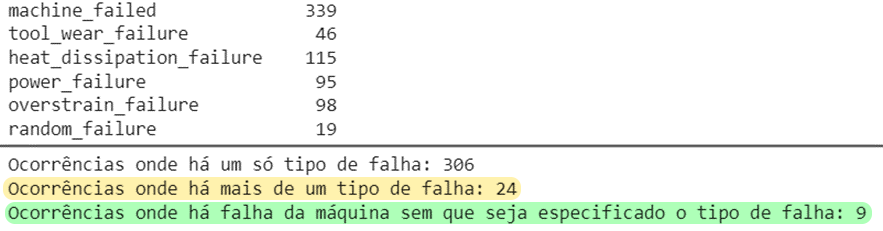
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| # | Designação | Seleção |  | # | Designação | Seleção |
| 1 | **udi** | Eliminada | 8 | **tool\_wear** | Selecionada |
| 2 | **product\_ID** | Eliminada | 9 | **machine\_failure** | Selecionada |
| 3 | **type** | Selecionada | 10 | **tool\_wear\_failure (TWF)** | Selecionada |
| 4 | **air\_temperature** | Selecionada | 11 | **heat\_dissipation\_failure (HDF)** | Selecionada |
| 5 | **process\_temperature** | Selecionada | 12 | **power\_failure (PWF)** | Selecionada |
| 6 | **rotational\_speed** | Selecionada | 13 | **overstrain\_failure (OSF)** | Selecionada |
| 7 | **torque** | Selecionada | 14 | **random\_failure (RNF)** | Selecionada |

**Tabela 9 -** Variáveis selecionadas para as fases seguintes do projeto

## 4.3 Limpeza de Dados

Em relação à limpeza dos dados, como mencionado na primeira parte deste projeto, verificámos que não existem valores omissos, duplicados ou mal inseridos em nenhuma das variáveis, talvez por ser um conjunto de dados sintético e, por esse motivo, não foi necessário fazer alterações a nível de limpeza de dados.

No entanto, este conjunto de dados apresenta uma característica que foi documentada também na primeira parte do projeto, e relatada pelo autor (Matzka, 2020): em 9 das ocorrências de falha (*machine\_failure* = 1) não é especificado o tipo de falha e em 24 ocorrências de falha, existe a indicação de mais de uma falha simultânea, o que pode ser origem de ambiguidade.



**Figura 17 -** Tipo e número de ocorrências de falha de máquina presentes no Conjunto de Dados original

Após ponderar se seria mais adequado eliminar estas entradas potencialmente ambíguas, optou-se por incluí-las no conjunto de dados, por se considerar que foram intenção do autor e por não serem necessariamente incorretas.

## 4.4 Formatação de Dados

Quanto à formatação dos dados, todas as variáveis foram renomeadas para que houvesse uma melhor compreensão dos seus nomes e para eliminação de espaços entre os mesmos. As variáveis de temperatura foram ainda convertidas de graus Kelvin para graus Celsius, de forma a facilitar a nossa compreensão dos valores associados às variáveis de temperatura (*air\_temperature* e *process\_temperature*)

A variável categórica ordinal *quality*, relativa à qualidade de peças em produção, sofreu ainda um processo de transformação por forma a ser convertida em variável numérica, que nos permitiu associar os números “0”, ”1” e “2” a cada nível de qualidade “Baixa”, “Média” e “Alta”, respetivamente.

De seguida, de modo a organizar os dados, realizou-se a divisão do conjunto de dados em Entradas e Saídas, obtendo-se a seguinte divisão:

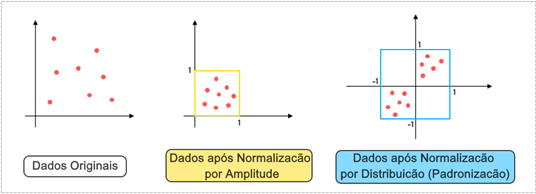
Uma imagem com texto

Descrição gerada automaticamente

**Figura 18 -** Variáveis de entrada e saída

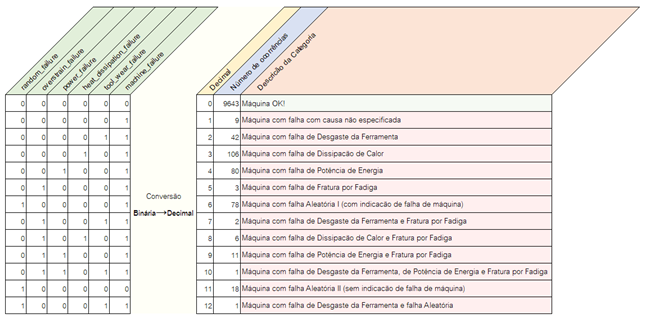
Por fim, foi feita a normalização por padronização das variáveis de entrada, para que todas as amostras transformadas ficassem com uma distribuição aproximadamente normal, com média igual a 0 e desvio padrão unitário, 𝑵~(𝝁=𝟎, 𝝈=𝟏).

A normalização por padronização é uma técnica de transformação de dados especialmente útil para algoritmos onde é conveniente ou necessário ter dados de entrada com distribuição normal, como a Regressão Linear, Regressão Logística ou Análise Descriminante Linear. Outros algoritmos que também beneficiam da normalização em amplitude, são as Redes Neuronais, o algoritmo kNN ou as Máquinas de Vetores de Suporte. Por outro lado, se o objetivo for interpretar os valores das variáveis por serem grandezas que só fazem sentido se o seu valor não sofrer transformações, como é o caso com as Árvores de Decisão, não faz sentido, nem é benéfico para o desempenho.



**Figura 19 -** Esquema simplificado da transformação de Dados por Normalização

Ainda relativamente à fase da formatação de dados, recorremos ao processo de Reamostragem dos dados, uma vez que os dados do conjunto de dados original é desbalanceado. Ao todo identificámos 13 categorias de saída diferentes que optámos por conservar, sendo que as ocorrências em cada uma destas categorias tem frequências muito distintas, tal como podemos observar na Tabela 2.



**Tabela 10 -** Categoria de saída, número de ocorrências em cada categoria e conversão binária-decimal

Finalmente, parece-nos relevante fazer referência à conversão do vetor binário que constitui a variável de saída, num número inteiro de base decimal. Com este efeito, ficamos com 12 categorias de falha (de 1 a 12) e uma em que se estima que a máquina esteja em funcionamento sem problemas (010 ou 0000002).

Para fazer face ao desbalanceamento do conjunto de dados, consideraram várias técnicas de Reamostragem (SMOTE ou ADASYN), mas como algumas categorias tem um número de ocorrências muito baixo, optou-se pela técnica de sobreamostragem ROSE (Lunardon, Menardi, & Torelli, 2014), implementado pela função *RandomOverSampler* do *sklearn*. Este algoritmo possibilita a sobreamostragem das classes minoritárias através da recolha aleatória de amostras com substituição.

## 4.5 Conjunto de Dados Final

Por fim, após todo o processo de seleção, limpeza, tratamento e formatação dos dados efetuado num *notebook Jupyter*, foi possível gerar o que chamamos de conjunto de dados final, constituído pela concatenação do que chamámos de variáveis de entrada e das variáveis objetivo ou de saída, tal como representado pelas tabelas em baixo.

Por se tratar de um conjunto de dados sintético, aberto e com origem académica, conclui-se que a qualidade dos dados é bastante aceitável para iniciar o processo de Modelação.

Uma imagem com mesa

Descrição gerada automaticamente

**Tabela 9 -** Captura de tela dos primeiros 10 registos do conjunto de dados final

# 5. Modelação

Os algoritmos de aprendizagem automática constroem um modelo baseado em dados de amostra, conhecidos como dados de treino, a fim de fazer previsões ou decisões sem serem explicitamente programados para o fazer. Cada um deste modelo tem características e capacidades diferentes e podem ser mais ou menos adaptados dependendo da situação. Nesta secção do trabalho descreveremos a escolha dos algoritmos, o desenho do teste de cada um dos algoritmos, o processo de construção e afinação de cada um dos modelos e, finalmente, a avaliação dos modelos.

## 5.1 Selecionar as Técnicas de Modelação

Para iniciar o processo de Modelação, começámos por definir quais os algoritmos que fariam mais sentido para a realização deste projeto. O primeiro aspeto a considerar é a natureza do problema: no nosso caso, estamos perante um problema de classificação – queremos que o nosso modelo devolva a categoria correta – com múltiplas saídas possíveis, ou seja, trata-se de um problema multiclasse, em que apenas se requer uma classe como resposta. Para além disto, como temos dados das entradas como das saídas desejadas, trata-se de um problema de aprendizagem supervisionada. Quanto ao conjunto de dados que vai ser como de base para a modelação, trata-se de dados tabulares.

Assim, considerando o nosso conhecimento e de acordo com o que foi abordado ao longo da unidade curricular de *Data Mining*, para desenvolver a fases seguinte deste projeto, decidimos utilizar os algoritmos relativos aos Modelos de:

1. Árvore de Decisão;
2. Máquina de Suporte de Vetores;
3. Floresta Aleatória;
4. Rede Neuronal.

Outras opções potencialmente promissoras foram consideradas (XGBoost ou LightGBM), mas optou-se pelo conjunto acima mencionado por termos uma melhor compreensão do seu funcionamento.

## 5.2 Conceção do Teste do Modelos

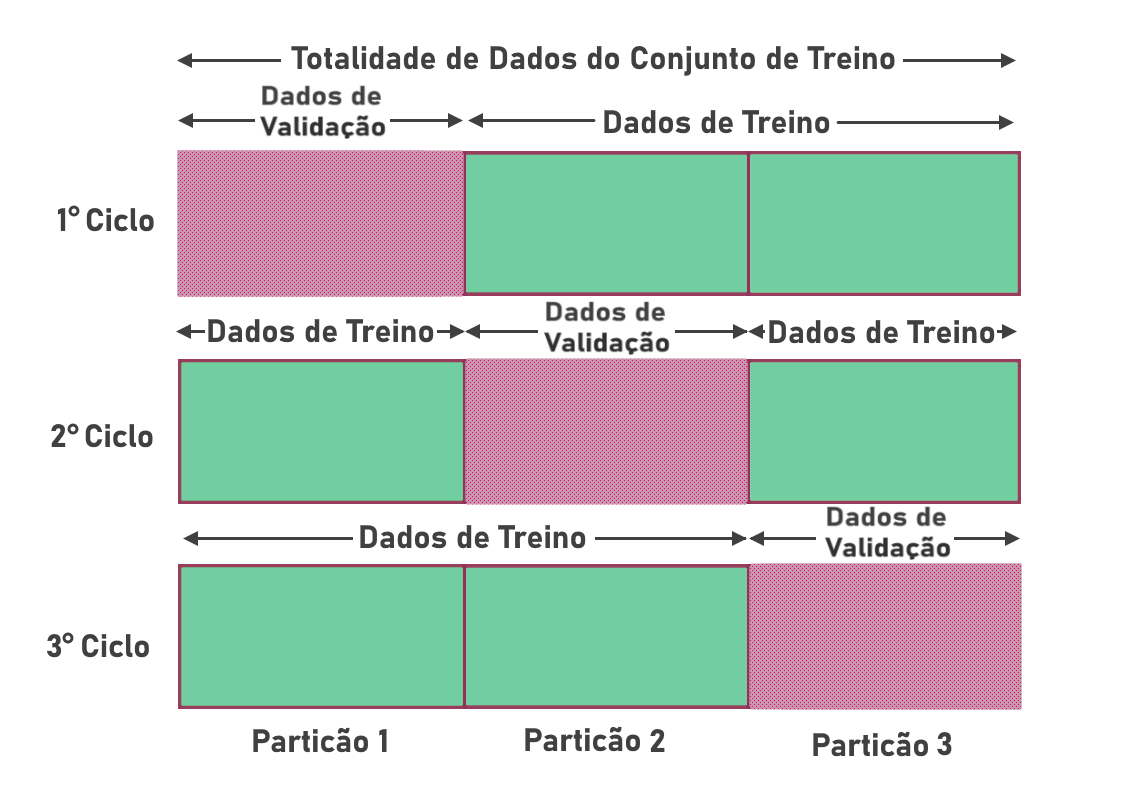
Depois de escolhidos os modelos, segue-se a fase de desenho do teste dos modelos. Nesta fase há uma série de questões a considerar:

- Como amostramos os dados de treino de forma adequada?

- Como colmatamos o problema dos dados desbalanceados?

- Quais são as métricas mais adequadas para o nosso problema?

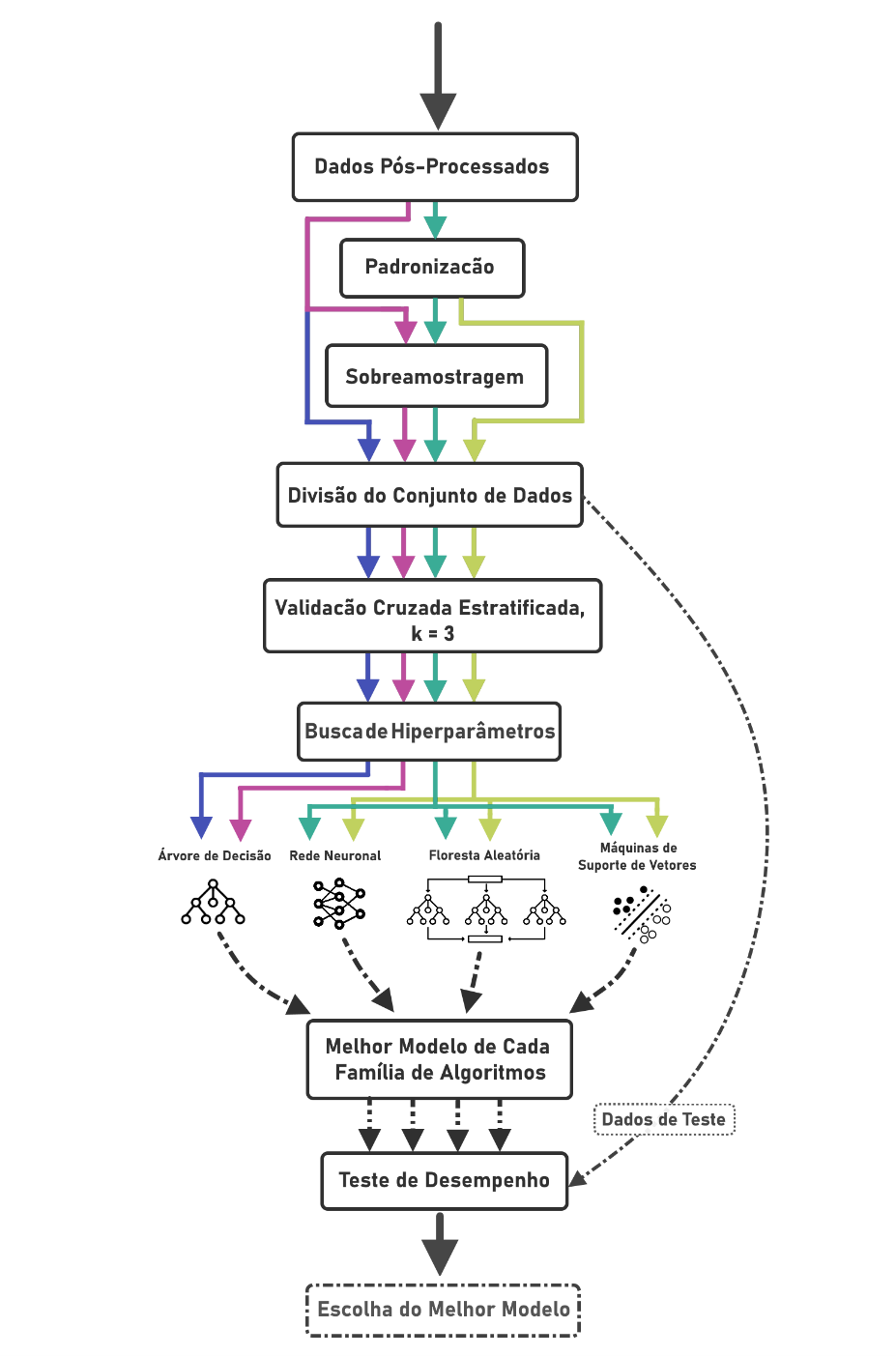
Uma forma possível de responder à primeira questão - e a forma que adotámos para este projeto – é a validação cruzada. No método de validação com k - partições, o conjunto de dados de teste é dividido em k subconjuntos de tamanho aproximadamente igual. Depois da divisão o modelo é treinado em k-1 partes e testado na partição que resta. Estre processo é repetido k vezes, utilizando em cada ciclo uma partição diferente para teste, tal como esquematizado na **Figura 17**, em baixo.



**Figura 17** - Método de amostragem por Validação Cruzada com k = 3

O desempenho final do modelo é dado pela média dos desempenhos observados sobre cada subconjunto de teste. Este processo de amostragem permite obter estimativas de desempenho mais confiáveis do que se utilizasse o mesmo conjunto de exemplos para teste e treino (Gama et al, 2017), método esse que é hoje geralmente visto com má prática.

No entanto, para medir o desempenho têm de ser primeiro definida uma métrica. Como já foi discutido em cima, teremos em conta um conjunto de métricas, mas daremos especial atenção às métricas que penalizam a existência de falsos-negativos, pois neste caso queremos detetar todos os positivos, mesmo que isso implique ter uma taxa de falsos-positivos um pouco mais elevada. Deste modo, os algoritmos foram treinados de forma a induzir-lhes uma Sensibilidade (ou *Recall*) o mais elevada possível. Concluindo, teremos em conta a Acurácia (ou Taxa de Acerto), a Precisão, mas a nossa atenção estará sobretudo na **Sensibilidade** e na **Medida-F**.



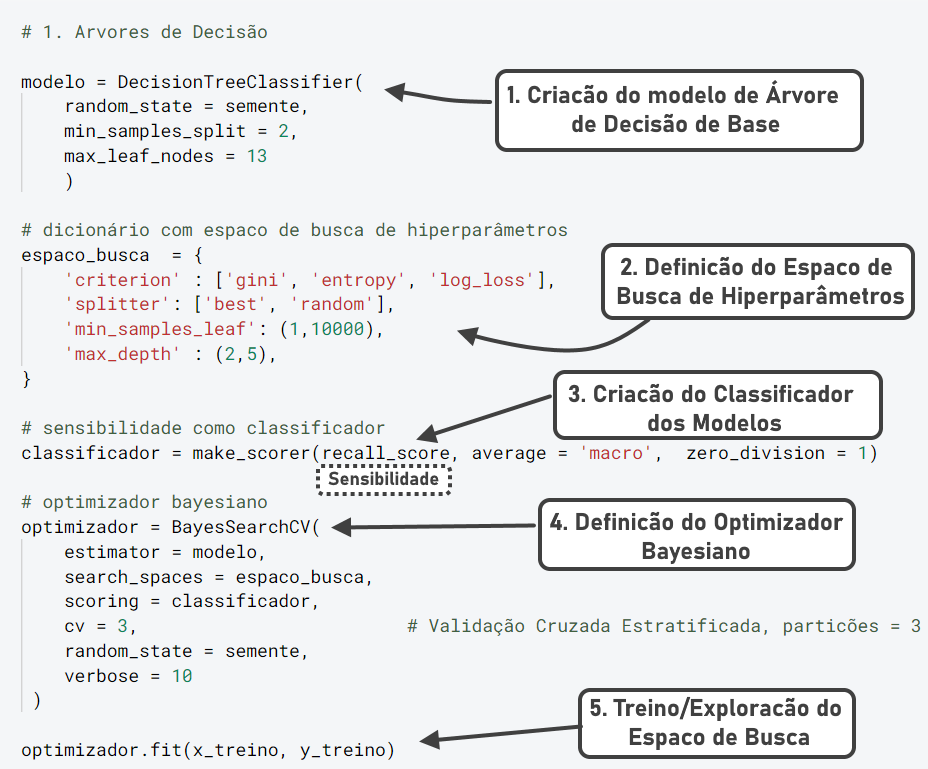
**Figura 18** – Esquema da sequência de passos seguidos para testar os modelos. A cor, a diferentes vias criadas para cada um dos algoritmos

Quanto à sequência dos passos da experiência, por termos algoritmos com características distintas e por não sabermos de antemão que espécie de impacto teria a sobreamostragem dos dados no desempenho dos modelos, optou-se por criar duas vias ou situações em que os algoritmos seriam testados: uma primeira sem Sobreamostragem de Dados, e um segunda com Sobreamostragem. Outro aspeto que consideramos foi a interpretabilidade do modelo de Árvore de Decisão: por ser a talvez a principal qualidade deste algoritmo, criou-se uma com o fim de preservar a distribuição das variáveis e ser mais fácil de ler o diagrama em árvore resultante. A Figura 18 procura capturar a sequência e raciocínio seguidos para testar os modelos.

## 5.3. Construção dos Modelo

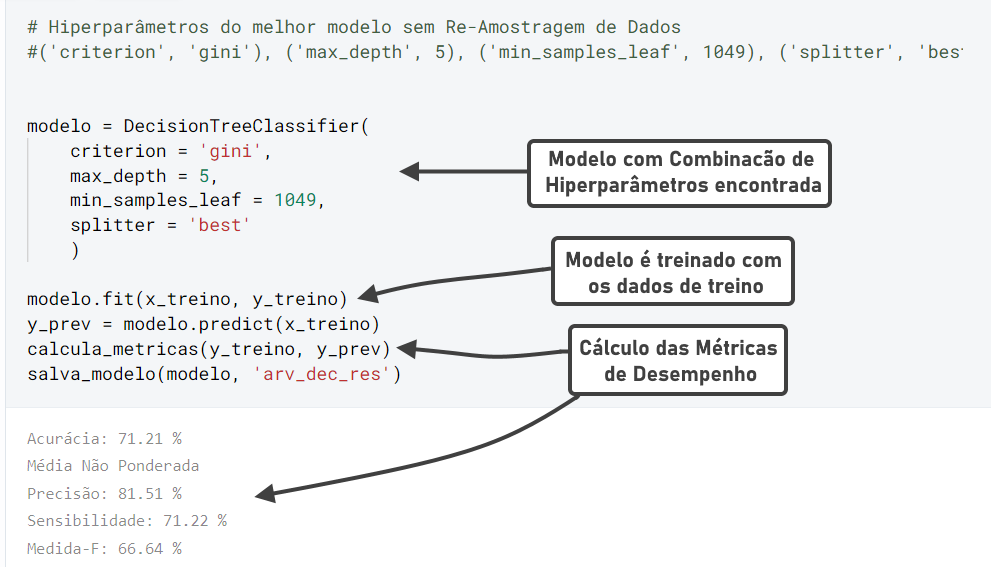
### Árvores de Decisão

Tal com esquematizado na Figura 18, um dos passos do teste de modelos consiste na busca de hiperparâmetros dos algoritmos. Entende-se por hiperparâmetros, os parâmetros cujo valor é utilizado ou tem influência no processo de treino do modelo. Em contraste, os outros parâmetros são “aprendidos” ou “descobertos” no processo de treino. Não sendo o nosso objetivo entrar em detalhes, existem muitas formas de abordar a busca dos hiperparâmetros entre os quais se incluem a busca em grelha, a busca aleatória, a busca baseada na evolução ou busca bayesiana, entre muitas outras. Alguns estudos indicam que a busca bayesiana obtém melhores resultados quando comparada com outras técnicas (Thornton et al, 2013) pelo que se optou por esta abordagem. A biblioteca Python *sklearn* providencia funções que facilitam a implementação deste algoritmo.



**Figura 19** - Captura de ecrã de excerto do *Jupyter notebook* relativo à busca bayesiana de hiperparâmetros

Findo este processo de busca, é nos devolvido o conjunto de hiperparâmetros encontrados. Esta configuração serve de base para o treino do modelo com o conjunto de dados de teste e, de seguida, faz-se uma primeira medição do desempenho do algoritmo com os dados de treino da figura em baixo. Ficamos com uma boa indicação se o modelo foi capaz de extrair informação do conjunto de dado.



**Figura 20** - Treino e desempenho de algoritmo com os dados de treino

Depois deste processo, mais uma vez são calculadas a métricas de desempenho para o conjunto de dados de teste e é contabilizado o tempo que o algoritmo leva fazer a predição da classe para todos os dados de treino, como termo comparativo entre modelos.

Para cada um dos modelos, este processo é feito com dados sobreamostrados e originais e registados os resultados.

### 5.3.2 Máquinas de Suporte de Vetores

O processo descrito no modelo com Árvore de Decisão foi repetido para o modelo com o algoritmo da Máquina de Suporte de Vetores. A busca de hiperparâmetros revelou-se um pouco mais difícil pois nos casos em que o grau de liberdade do *kernel* é maior que 3, o tempo de treino crescia exponencialmente e, ao invés de levar segundos ou poucos minutos, levava algumas horas para cada ciclo da validação cruzada, o que se revelou incomportável. Por essa razão, estabelecemos como limite 3 graus de liberdade.

Uma imagem com texto

Descrição gerada automaticamente

**Figura 21** - Criação e definição do espaço de busca de hiperparâmetros

Considerando as particularidades do algoritmo de Máquina de Suporte de Vetores, o espaço de busca é naturalmente diferente do que vimos na Árvore de Decisão. Por uma questão de economia de tempo e por já o ter sido feito no decorrer da disciplina, não definiremos aqui os algoritmos de cada um dos modelos, mas faremos eco apenas dos resultados e eventuais dificuldades encontradas no decorrer dos testes.



**Figura 22** - Detalhe da função de busca bayesiana para o modelo com algoritmo de Máquina de Suporte de Vetores

Ainda não referido, após o processo de teste o melhor modelo é guardado em ficheiro de forma que possamos aceder ao modelo numa fase posterior, seja para proceder a testes, seja para exportar o modelo para outras plataformas que não *Jupyter notebooks*.



**Figura 23** – Modelo encontrada após busca de hiperparâmetros com dados sobreamostrados

### 5.3.3 Floresta Aleatória

De forma muita sucinta, o algoritmo de Floresta Aleatória tem em comum com o algoritmo de Árvore de Decisão o facto de construir várias Arvores de Decisão durante o seu treino, tendo como saída a classe a moda das classes, para o caso de ser utilizada num problema de classificação.

Todos os modelos foram implementados tenho como base a biblioteca *sklearn*, sendo que para a floresta aleatória a função pré-definida é a *RandomForestClassifier*, tal como é possível verificar na Figura 24.

Uma imagem com texto

Descrição gerada automaticamente

**Figura 24** – Criação do modelo base de através

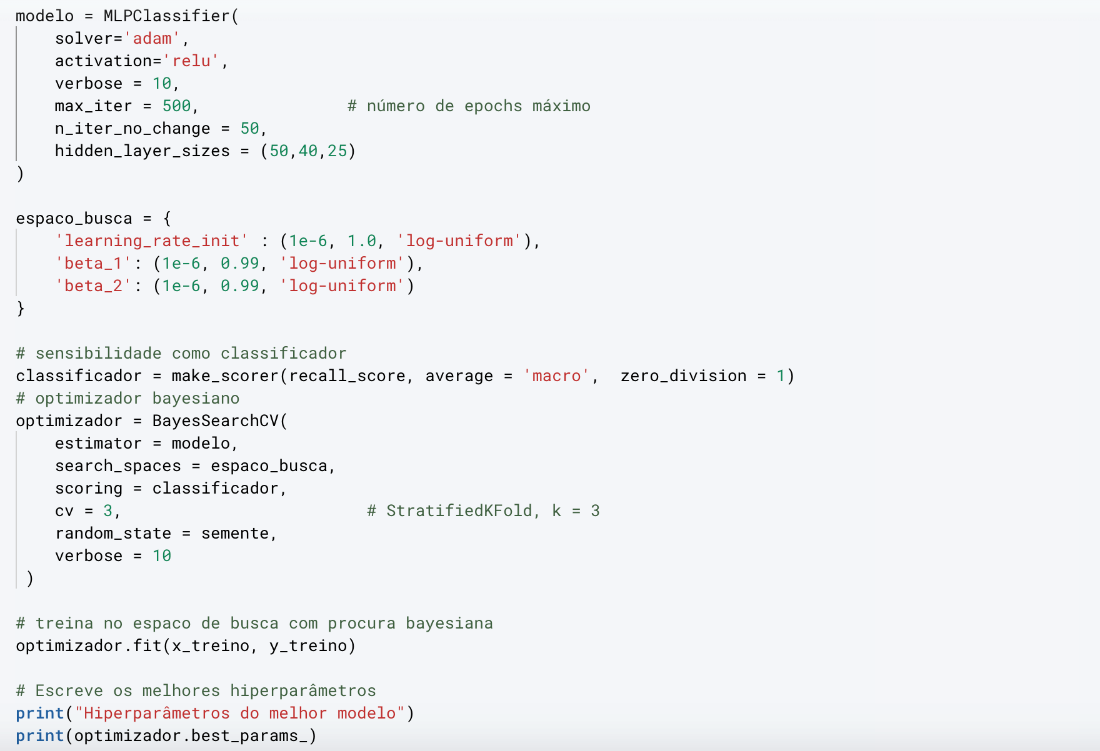
Olhando para o espaço de busca criado, em retrospetiva, talvez se pudesse ter deixado fazer a pesquisa numa profundidade maior, isto é, o número de patamares de decisão máximo. Ainda que o espaço de busca retire, até certo ponto, a responsabilidade a quem desenha o modelo, pois, em teoria, sintoniza o modelo ao conjunto de dados através do hiperparâmetros, a escola do espaço de busca fica ao critério do cientista de dados.



**Figura 25** -Código relativo ao treino do modelo de Floresta Aleatória

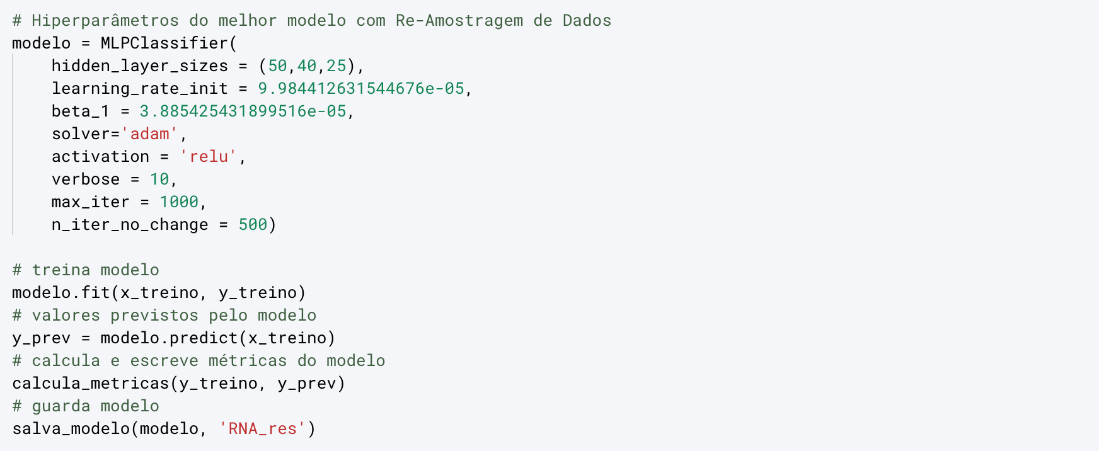
### Rede Neuronal

De todos os algoritmos testados as redes neuronais são as que mais parâmetros, o que lhe confere uma particular dificuldade no momento de treino. Foram várias a iterações até se chegar à arquitetura do modelo base. Inicialmente, tentou-se treinar a rede com apenas uma camada oculta, mas verificou-se que o valor da função de custo atingia um *plateau.* Deste modo, de forma iterativa, optou-se por 3 camadas ocultas, como se pode ver na Figura 26.



**Figura 26** - Detalhe do desenho do teste do modelo com rede Neuronal

Outro aspeto de relevo é o tempo que esta rede neuronal leva a ser treinada quando comparada com a Árvore de Decisão ou a Florestas Aleatória, por exemplo: a primeira leva horas enquanto as segundas apenas segundos.



**Figura 27** – Configuração do melhor modelo de Rede Neuronal encontrado

Desta forma, o espaço de procura não pode ser muito vasto sob pena de o algoritmo de busca demorar demasiado tempo a encontrar hiperparâmetros adequados. Assim, optou-se por buscar apenas os hiperparâmetros do optimizador (*solver*).

## 5.3 Avaliação do Modelo

Após serem treinados todos os modelos dos diferentes algoritmos, passámos à fase de avaliação da sua performance, de forma a conseguirmos selecionar o melhor modelo com base nas métricas previamente definidas, e para posteriormente ser colocado em produção.

Uma imagem com mesa

Descrição gerada automaticamente

**Tabela 10** – Resultados do Desempenho dos Modelos

Na Tabela 10 são apresentados todos os resultados de cada modelo, relativamente à fase de treino e à fase de teste, de acordo com os dados originais e com sobreamostragem, respetivamente.

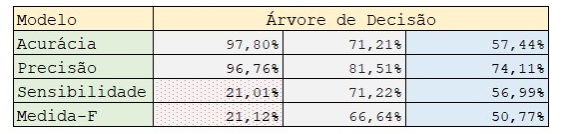
Na próxima etapa serão avaliados os resultados de cada modelo em específico. Para já, podemos afirmar e concluir que à priori, o modelo que obteve melhores resultados a nível dos dados originais, e relativamente aos valores da sensibilidade, a métrica que indica a possível falha ou não, foi o modelo da Rede Neuronal. Já relativamente aos dados com reamostragem, verificamos que o modelo que obteve melhor resultado, avaliando a mesma métrica, foi o Modelo Máquinas de Suporte de Vetores. Já na fase de teste, fase final, concluímos que também o Modelo Máquinas de Suporte de Vetores foi o que obteve melhores resultados.

# 6. Avaliação

## 6.1 Avaliação dos Resultados

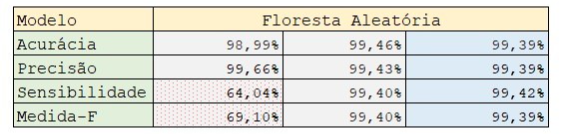
Como referido anteriormente, nesta fase do projeto faremos a avaliação mais específica dos resultados de cada um dos modelos testados, verificando a sua qualidade e precisão.

Relativamente ao Modelo da Árvore de Decisão, verificamos que, na fase de treino dos dados com reamostragem, existe uma melhoria a nível da Sensibilidade e Medida-F em relação aos dados originais, no entanto, verifica-se um grande decréscimo nos valores da Acurácia e Precisão. Já na fase teste, podemos concluir que os valores não são tão positivos em comparação com os restantes modelos que serão avaliados de seguida.



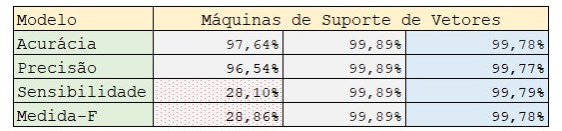
**Tabela 11** – Métricas de Desempenho do Modelo de Árvore de Decisão

A nível do modelo da Floresta Aleatória, verifica-se que os dados de reamostragem são visivelmente mais positivos que os dados originais, o que nos leva a um resultado bastante positivo na fase de teste, com cerca de 99,42% de sensibilidade e 99,39% de precisão.



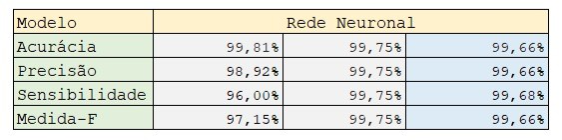
**Tabela 12** - Métricas de Desempenho do Modelo de Floresta Aleatória

Em relação à Máquina de Suporte de Vetores, podemos analisar que os dados originais eram significativamente baixos em relação à métrica de Sensibilidade e Medida-F, verificando-se uma grande melhoria nos dados com reamostragem. Por fim, na fase teste, verifica-se que todos as métricas de erro, nomeadamente, a Sensibilidade, Precisão, Acurácia e Medida-F, se encontram acima dos 99%, à semelhança do modelo anterior.



**Tabela 13** - Métricas de Desempenho do Máquinas de Suporte de Vetores

Por fim, analisando os resultados da Rede Neuronal, verificamos que, os valores do modelo com e sem reamostragem são igualmente positivos, verificando-se uma ligeira melhoria nos dados com reamostragem. Na fase de teste, o modelo alcançou valores superiores a 99,6%.

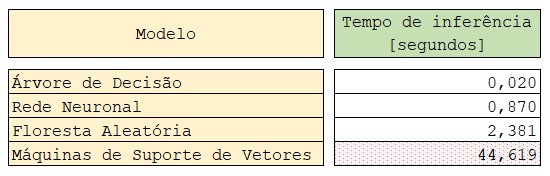


**Tabela 14** - Métricas de Desempenho do Máquinas de Suporte de Vetores

Nesta fase torna-se também relevante comparar os tempos de inferência de cada um dos modelos, pois um modelo pode ser superior a outro relativamente aos resultados que obtém, mas, no entanto, ser bastante mais lento e isso pode não ser o ideal no dia-a-dia do negócio.

Uma vez que a empresa em estudo pretende prever avarias específicas das máquinas que são utilizadas no dia a dia, de forma a evitar perdas de tempo com tarefas de reparação e substituição, é bastante relevante que o modelo em uso seja o mais rápido e eficiente possível, e que o sistema que vier a implementar o modelo de classificação seja responsivo.

Assim, podemos verificar pela Tabela 15 que em relação aos seus tempos de predição, em segundos, a Máquina de Suporte de Vetores é o modelo que demora mais de 30 vezes mais tempo que os restantes. A Árvore de Decisão e a Rede Neuronal são os modelos que têm menos tempo de inferência.



**Tabela 15** – Tempo de cálculo de predição de todos os dados do conjunto de treino

A fim de concluir a análise efetuada, podemos afirmar que o modelo que melhor se ajusta em resposta à problemática em estudo corresponde ao algoritmo da Rede Neuronal. No entanto, verificou-se, ao longo do projeto e de todos os testes efetuados, que apesar dos seus excelentes resultados, e em comparação com a Floresta Aleatória (que também obteve resultados satisfatórios), esse modelo demorou sensivelmente 10 vezes mais tempo a treinar. Ou seja, enquanto a Floresta Aleatória leva aproximadamente 3 minutos, a Rede Neuronal levou pelo menos 3 horas.

## 6.2 Processo de Revisão

O resultado final do processo de Modelação que aqui descrevemos é apenas uma primeira iteração dum sistema que se pretende seja de melhoria e vigilância contínuas. Existe um largo espaço para melhoria no que toca à afinação dos modelos que desenvolvemos assim como à experimentação doutros modelos e técnicas.

Ainda assim, consideramos que com este projeto foi possível a construção de um modelo viável, capaz de corresponder às necessidades do negócio, com alta capacidade de deteção de falhas, alta sensibilidade e potencialmente capaz de dar resposta aos problemas do dia-a-dia da empresa. Assim, podemos verificar que, apesar de possíveis falhas, os riscos são baixos, e que todos os requisitos propostos do projeto foram cumpridos.

## 6.3 Determinar os Próximos Passos

Para ações e projetos futuros, destacamos a possibilidade e interesse em testar outros modelos igualmente interessantes, como é o caso do XGBoost e LightBGM, através de um cenário real e dados reais de uma indústria em específico.

Para cumprir essas ações, seria importante melhorar o sistema atual, inclusive a melhoria da aplicação, desenvolvida em Streamlit, que será mencionada mais à frente neste projeto. Gostaríamos ainda de destacar, que para que este estudo fosse mais completo e realista, seria relevante a existência de dados em tempo real provenientes de sensores de máquinas industriais e a criação de um sistema capaz de os processar síncrona ou assincronamente. ao contrário do conjunto de dados que nos baseamos para este estudo que não tem omissões, *outliers* ou falhas com muito relevo, conhecendo a realidade da indústria local, estamos em crer que os conjuntos de dados seriam muito desafiantes de pré-processar e limpar.

# 7. Desenvolvimento

## 7.1 Plano de Implementação

Para dar início ao processo de implementação do projeto, foi necessário executar um estudo sobre as plataformas de implementação de modelos de aprendizagem automática, onde acabámos por optar pela ferramenta Streamlit, pela maior facilidade de implementação e adaptação ao nosso propósito (Shin, 2021).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Simplicidade | Maturidade | Flexibilidade | Principal Utilização |
| **Streamlit** | Alta | Baixa | Média | Demonstrações |
| Gradio | Alta | Baixa | Média | *Dashboards* |
| Dash | Média | Média | Média | *Dashboards* |
| Flask | Baixa | Alta | Alta | Interfaces Web |

**Tabela 16** – Síntese das características das plataformas de implementação de modelos

O Streamlit é uma plataforma de código aberto, que nos permitirá colocar em produção o projeto, sem qualquer necessidade de conhecer ferramentas de desenvolvimento front-end. Por meio desta plataforma será possível criarmos uma aplicação interativa que vai reunir todos os dados essenciais para a empresa em estudo conseguir facilmente ter acesso à previsão de avarias das máquinas, que será partilhada com a organização através de um URL partilhável.

Para implementação do modelo nesta plataforma, optámos pela utilização do modelo da Rede Neuronal, conforme avaliada como a melhor opção para o projeto.

O protótipo desenvolvido é composto por duas páginas. Uma das páginas é de Exploração de Dados, onde o utilizador tem oportunidade de compreender melhor o conjunto de dados. Esta página foi desenvolvida mais como complemento e por interesse em explorar a ferramenta.

Uma imagem com mesa

Descrição gerada automaticamente

**Figura 28** – Página de Exploração de Dados desenvolvida com a plataforma Streamlit

A outra página, é a que implementa um sistema de Predição simples, onde o utilizador insere os valores das variáveis de entrada e, pressionando o botão “Calcula Estado da Máquina”, o modelo baseado na Rede Neuronal devolve uma mensagem de texto com a categoria que correspondente àquele sinal de entrada, de acordo com o esquematizado na Tabela 10.

Como próximo passo de desenvolvimento, cremos que seria interessante no futuro criar um programa capaz de prever o estado da máquina mediante um sinal real ou simulado, sem que o fosse necessário a inserção manual de valores.



**Figura 29** - Aspeto gráfico do protótipo desenvolvido para *deployment* do modelo preditivo

## 7.2 Plano de Monitorização e Manutenção

A manutenção, monitorização e atualização do modelo em estudo ficará à responsabilidade do técnico contratado para implementação deste projeto, ficando este responsável por todo o processo e implementação do modelo de *Data Mining*.

## 7.3 Elaboração do Relatório Final

Este projeto de *Data Mining* consistiu na elaboração de um projeto de *Data Mining* numa empresa fabril, que permitisse que esta não sofresse problemáticas por culpa de falhas em maquinaria. Criou-se assim as várias opções de modelos de *Data Mining*, por forma a testar qual seria a melhor solução para corresponder a todos os objetivos da empresa e todas essas opções foram avaliadas ao longo deste relatório. Em suma, verificámos que a implementação deste projeto seria viável num cenário real, verificando-se assim o sucesso do projeto.

## 7.4 Revisão do Projeto

Ao longo do projeto foi possível deparar-nos com vários desafios que nos levaram a questionar e a procurar respostas às dificuldades encontradas sempre de forma autónoma. Procurámos sempre ultrapassar essas dificuldades com base na experimentação e nos vários testes efetuados. No fim, optámos por utilizar o algoritmo que melhor correspondeu à necessidade do projeto, mas verificando a sua demora, acreditamos que poderão existir melhorias a fazer, como em qualquer projeto piloto.

# 8. Conclusão

Este projeto permitiu-nos implementar tudo o que aprendemos ao longo da unidade curricular de *Data Mining* do Mestrado em Ciência de Dados do Instituto Politécnico de Leiria. Com a realização deste trabalho prático, começámos por fazer a recolha e preparação do conjunto de dados para o projeto, que foi utilizado para realização dos treinos e testes dos modelos em estudo, de forma a avaliar qual seria o melhor algoritmo para dar resposta à necessidade da empresa em causa.

A primeira parte do projeto tinha como objetivo executar as primeiras duas seções do modelo de referência CRISP-DM que consiste na Compreensão do Negócio e na Compressão dos Dados. Na segunda parte deste trabalho, faremos a Preparação dos Dados, seguida da Modelação, Avaliação dos Modelos e, finalmente, a fase de Desenvolvimento.

Para além da capacidade técnica e analítica nas disciplinas de aprendizagem automática ou estatística, uma das valências importantes de um cientista de dados é a habilidade para perceber problemas concretos das mais variadas indústrias, formular hipóteses e soluções para esses problemas utilizando o manancial de técnicas que tem ao seu dispor e, finalmente, conseguir comunicar eficazmente com leigos nas ciências de dados. No entanto, sendo este trabalho baseado num conjunto de dados relativamente realista, mas fictício, nem sempre nos foi fácil contextualizar o nosso caso de estudo e, sobretudo, obter o enquadramento e descrever as condicionantes com o realismo desejável.

Na segunda parte do projeto foi um desafio mitigar explorando as soluções que existem para a implementação realista de modelos de aprendizagem automática em ambientes industriais de pequenas e médias empresas que abundam na região centro.

Podemos afirmar, com a finalização deste projeto, que o modelo que obteve melhores resultados foi a Rede Neuronal, que com a presença dos dados de reamostragem atingiu valores bastante superiores a 99,6%. Podemos ainda afirmar que utilizar o modelo da Floresta Aleatória seria também uma boa aposta face aos resultados obtidos, como já referimos anteriormente.

Concluímos que, apesar de fictício, este projeto poderia realmente ser implementado numa indústria semelhante àquela que estudámos ao longo do trabalho, e que seria sem dúvida uma mais-valia para evitar erros e falhas de máquinas.

# 9. Referências

1. *AI4I 2020 Predictive Maintenance Dataset Data Set.* (30 de Agosto de 2020). Obtido em 27 de Abril de 2022, de UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset
2. Amruthnath, N., & Gupta, T. (2018). A Research Study on Unsupervised Machine Learning Algorithms for Fault Detection in Predictive Maintenance. Obtido em 16 de 05 de 2022, de https://www.researchgate.net/publication/322869981\_A\_Research\_Study\_on\_Unsupervised\_Machine\_Learning\_Algorithms\_for\_Fault\_Detection\_in\_Predictive\_Maintenance
3. Barella, V. H. (2015). Técnicas para o problema de dados desbalanceados em classificacão hierárquica. Obtido de https://www.teses.usp.br/teses/disponiveis/55/55134/tde-06012016-145045/publico/VictorHugoBarella\_dissertacao\_revisada.pdf
4. Caetano, L., & Gama, R. (2004). Industrialização, desindustrialização e desenvolvimento - a indústria na Região Centro. *21/23*(Cadernos de Geografia), 259-268. Obtido em 16 de 05 de 2022, de https://www.uc.pt/fluc/depgeotur/publicacoes/Cadernos\_Geografia/Numeros\_publicados/CadGeo21\_23/artigo35
5. Comissão Europeia. (2009). *Cosmetic product notification portal.* Obtido em 16 de May de 2022, de European Commission: http://ec.europa.eu/growth/sectors/cosmetics/cpnp/
6. Doran, G. (1981). There's a SMART Way to Write Management's Goals and Objectives. *Management Review*. Obtido em 16 de May de 2022, de https://www.scirp.org/(S(czeh2tfqyw2orz553k1w0r45))/reference/ReferencesPapers.aspx?ReferenceID=1459599
7. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. Obtido em 16 de 05 de 2022, de https://arxiv.org/pdf/1702.08608.pdf
8. Forum Estudante. (25 de 06 de 2019). *A Indústria da região de Leiria em números.* Obtido em 16 de May de 2022, de Forum Estudante: https://www.forum.pt/leiria-in/a-industria-da-regiao-de-leiria-em-numeros
9. Gama et al, J. (2017). *Extracão de Conhecimento de Dados - Data Mining.* Lisboa: Sílabo.
10. International Organization for Standardization. (2007). ISO 22716:2007(en) Cosmetics — Good Manufacturing Practices (GMP) — Guidelines on Good Manufacturing Practices. *ISO - International Organization for Standardization*. Obtido em 16 de 05 de 2022, de https://www.iso.org/obp/ui/#iso:std:iso:22716:ed-1:v2:en
11. Jornal de Leiria. (5 de 12 de 2019). 250 Maiores Empresas do Distrito de Leiria. *Jornal de Leiria*. Obtido em 16 de May de 2022, de https://www.insignare.pt/m/7212/250-maiores-empresas-de-leiria-2019.pdf
12. Loyola-González, O. (5 de 10 de 2019). *Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View.* Obtido em 16 de 05 de 2022, de IEEEAccess: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8882211
13. Lunardon, N., Menardi, G., & Torelli, N. (06 de 2014). ROSE: A Package for Binary Imbalanced. *The R Journal Vol. 6/1*, pp. 79-89.
14. Marcorin, W., & Lima, C. (2003). Análise dos Custos de Manutenção e de Não-manutenção de Equipamentos Produtivos. *Revista de Tecnologia e Ciência, 11*(22), 35-42.
15. Matzka, S. (2020). Explainable Artificial Intelligence for Predictive Maintenance Applications. *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, 69-74. Obtido de Research Gate: https://www.researchgate.net/publication/344327789\_Explainable\_Artificial\_Intelligence\_for\_Predictive\_Maintenance\_Applications
16. Ministério da Saúde. (24 de 09 de 2008). Decreto-Lei n.º 189/2008, de 24 de setembro. *Diário da República*. Obtido em 16 de 05 de 2022, de https://dre.pt/dre/detalhe/decreto-lei/189-2008-452215
17. Ministério do Ambiente. (11 de 12 de 2017). Decreto-Lei n.º 152-D/2017. *Diário Da Républica Eletrónico*. Obtido em 16 de May de 2022, de https://dre.pt/dre/detalhe/decreto-lei/152-d-2017-114337042
18. Parlamento Europeu, C. (22 de 12 de 2009). Regulamento (CE) no 1223/2009 do Parlamento Europeu e do Conselho de 30 de Novembro de 2009 relativo aos produtos cosméticos. *Jornal Oficial da União Europeia*. Obtido em 16 de 5 de 2022, de https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32009R1223&from=de
19. Shin, T. (23 de 02 de 2021). *Gradio vs Streamlit vs Dash vs Flask*. Obtido de Towards Data Science: https://towardsdatascience.com/gradio-vs-streamlit-vs-dash-vs-flask-d3defb1209a2
20. Thornton et al, C. (2013). *Auto-WEKA: Combined Selection and Hyperparameter.* Vancouver, Canada: University of British Columbia.

1. Em essência, a indústria 4.0 é a tendência para a automação e intercâmbio de dados em tecnologias e processos de fabrico que incluem sistemas ciberfísicos (CPS), Internet das coisas industrial, computação em nuvem, computação cognitiva, e inteligência artificial. [↑](#footnote-ref-2)
2. um conjunto de dados é dito desbalanceado quando nele existe uma clara desproporção entre o número de exemplos de uma ou mais classes em relação às demais [↑](#footnote-ref-3)
3. Passeio aleatório é um processo estocástico que descreve um caminho que consiste numa sucessão de passos aleatórios num determinado espaço matemático. [↑](#footnote-ref-4)