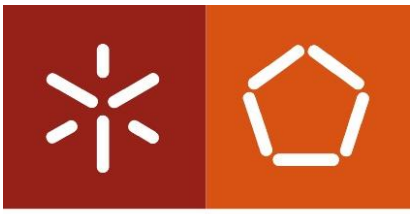


Universidade do Minho

Mestrado em Engenharia Informática

Dados e Aprendizagem Automática

Conceção e otimização de modelos de *Machine Learning*.

Grupo 22

Duarte Parente (PG53791) Gonçalo Pereira (PG53834)

José Moreira (PG53963) Santiago Domingues (PG54225)

Ano Letivo 2023/2024

Índice

1 Introdução **4**

2 Tarefa Dataset Grupo **5**

2.1 Business Understanding 5

2.2 Data Understanding 5

2.2.1 Ingestão dos Dados 5

2.2.2 Exploração e Visualização dos Dados 7

2.3 Data Preparation 8

2.3.1 Feature Importance 10

2.4 Modeling 10

2.4.1 Definição de Modelos e respetivos Hiperparâmetros 10

2.5 Evaluation 11

2.5.1 Avaliação do modelo sem *drop* de *features* 11

2.5.2 Avaliação do modelo com *drop* de *features* através da correlação 11

2.5.3 Avaliação do modelo com *drop* de *features* através da correlação e da *feature importance* 12

3 Tarefa Dataset Competição **13**

3.1 Business Understanding 13

3.2 Data Understanding 13

3.2.1 Ingestão dos Dados 13

3.2.2 Exploração e Visualização dos Dados 16

3.3 Data Preparation 18

3.3.1 Feature Importance 18

3.4 Modeling 19
3.4.1 Definição de Modelos e respetivos Hiperparâmetros 19
3.5 Evaluation 20

4 Conclusão **21**

Capítulo 1

Introdução

O presente relatório visa apresentar o projeto desenvolvido no âmbito da Unidade Curricular de **Dados e Aprendizagem Automática**. O trabalho prático encontra-se dividido em duas tarefas: de **Grupo** e de **Competição**, cada uma envolvendo um conjunto diferente de dados. O primeiro *dataset* foi escolhido pelo grupo através da plataforma **Kaggle** enquanto que o segundo foi atibuído pela equipa docente. O principal objetivo do projeto passa pela conceção de um projeto de **Machine Learning**, procurando extrair conhecimento relevante para o contexto dos problemas definidos, utilizando para o efeito as técnicas e modelos de aprendizagem abordados ao longo do semestre.

A metodologia adotada nas duas tarefas para a realização deste projeto foi o **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*). Desta forma, foi possível garantir uma abordagem sistemática e iterativa ao longo de todas as fases do projeto, proporcionando uma estrutura organizada para o desenvolvimento do mesmo e contribuindo para a eficácia e a qualidade dos resultados obtidos.

Neste relatório serão apresentados os resultados obtidos em cada uma das tarefas, bem como as metodologias e decisões tomadas para a construção dos modelos de *Machine Learning*. Serão também discutidos os desafios enfrentados durante cada processo, sendo por fim apresentadas as conclusões retiradas.

Capítulo 2

Tarefa Dataset Grupo

A primeira etapa relativa à tarefa de grupo deste trabalho prático consistia na consulta e escolha de um *dataset* para ser explorado. Dessa forma, recorreu-se à plataforma **Kaggle**, onde foi possível encontrar um conjunto de dados apropriado e suficientemente robusto para corresponder com qualidade ao objetivo de estudo definido.

2.1 Business Understanding

Como um dos desportos mais populares do planeta, o futebol sempre foi acompanhado de perto por um grande número de interessados. No entanto, com os avanços tecnológicos, nomeadamente no ramo da ciência de dados, testemunhamos o impacto crescente que esta ciência vai apresentando no seio deste desporto.

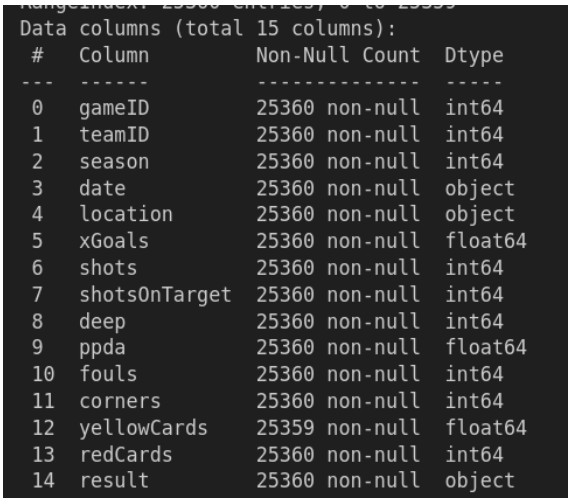
Um jogo de futebol tem três resultados possíveis para cada equipa: vitória, derrota ou empate. Consequentemente, este conjunto reduzido e limitado de opções pode levar à perceção equivocada de que prever o resultado de um jogo é praticamente direto. Uma das razões pelas quais este desporto é tão apreciado é o seu elemento inerente de imprevisibilidade, onde eventos aparentemente aleatórios ou improváveis podem ocorrer, o que significa que nem sempre as estatísticas de um jogo traduzem o seu resultado final. Para além disso, a coleção cada vez mais abundante de dados durante uma partida de futebol torna a sua utilização imperativa para realização de diversas análises sobre o próprio jogo e os seus intervenientes.

2.2 Data Understanding

2.2.1 Ingestão dos Dados

- Dataset Escolhido: **Football Database**

O conjunto de dados escolhido contém estatísticas jogos no período de 2014 a 2020, abrangendo todas os jogos das cinco principais ligas europeias ao longo do período de tempo especificada: **Premier League** (Liga Inglesa), **Serie A** (Liga Italiana), **Bundesliga** (Liga Alemã), **La Liga** (Liga Espanhola) e **Ligue 1** (Liga Francesa).



```
Data columns (total 15 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   gameID         25360 non-null  int64
 1   teamID         25360 non-null  int64
 2   season         25360 non-null  int64
 3   date           25360 non-null  object
 4   location       25360 non-null  object
 5   xGoals         25360 non-null  float64
 6   shots          25360 non-null  int64
 7   shotsOnTarget  25360 non-null  int64
 8   deep           25360 non-null  int64
 9   ppda           25360 non-null  float64
 10  fouls          25360 non-null  int64
 11  corners        25360 non-null  int64
 12  yellowCards    25359 non-null  float64
 13  redCards       25360 non-null  int64
 14  result         25360 non-null  object
```

Figura 1: Atributos do Dataset de Grupo

O *dataset* apresenta 15 atributos, dos quais 12 numéricas e 3 categóricas, contendo registos relativos a 25360 jogos.

- **gameID:** Identificador único para cada jogo.
- **teamID:** Identificador único para cada equipa.
- **season:** Temporada durante a qual o jogo ocorreu.
- **date:** Data do jogo.
- **location:** Jogo em casa ou fora.
- **xGoals:** Estimativa relativa ao número de golos que a equipe deveria ter marcado.
- **shots:** Número de remates.
- **shotsOnTarget:** Número de remates à baliza.
- **deep:** Passes efetuados a uma distância máxima de 20 metros da baliza (excluindo cruzamentos).
- **ppda:** Média de passes permitidos por ação defensiva na metade do campo adversário.
- **fouls:** Número de faltas cometidas pela equipa.
- **corners:** Número de cantos concedidas à equipa.
- **yellowCards:** Número de cartões amarelos recebidos pelos jogadores da equipa.
- **redCards:** Número de cartões vermelhos recebidos pelos jogadores da equipa.
- **result:** (**Variável Dependente**) Resultado do jogo.

2.2.2 Exploração e Visualização dos Dados

Balanceamento da Variável Dependente

A primeira análise realizada prendeu-se na verificação do balanceamento da variável dependente. Foi possível verificar que **25.02%** dos resultados correspondiam a **empates**, enquanto que os restantes **74.98%** (37.49%*2) estavam divididos entre vitórias e derrotas. Esta observação levou à conclusão de que os registos estavam balanceados pelas três categorias da variável dependente. Na figura abaixo encontra-se representada

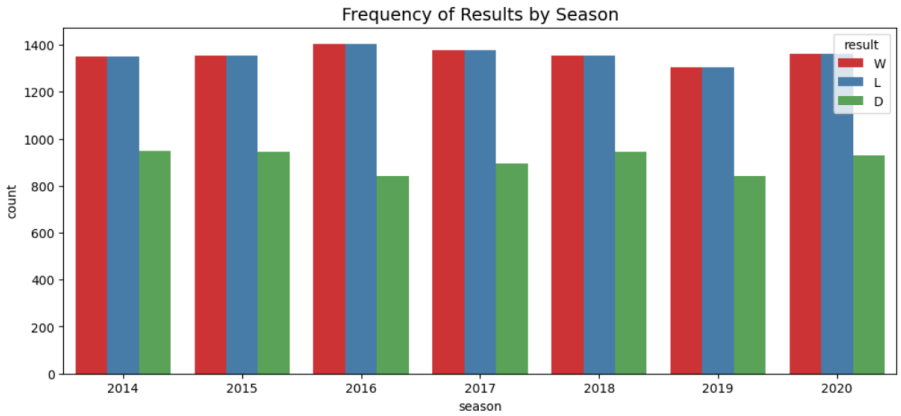


Figura 2: Distribuição da frequência de Resultados por Época

Influência e interação das variáveis independentes no Resultado

- **XGoals vs Game Location**: A figura abaixo demonstra claramente a influência no resultado no desfecho de um jogo. É possível observar que uma equipa que jogue fora apresenta uma maior tendência para a obtenção de um resultado negativo, num cenário onde não cria oportunidades suficientes para ganhar o jogo.

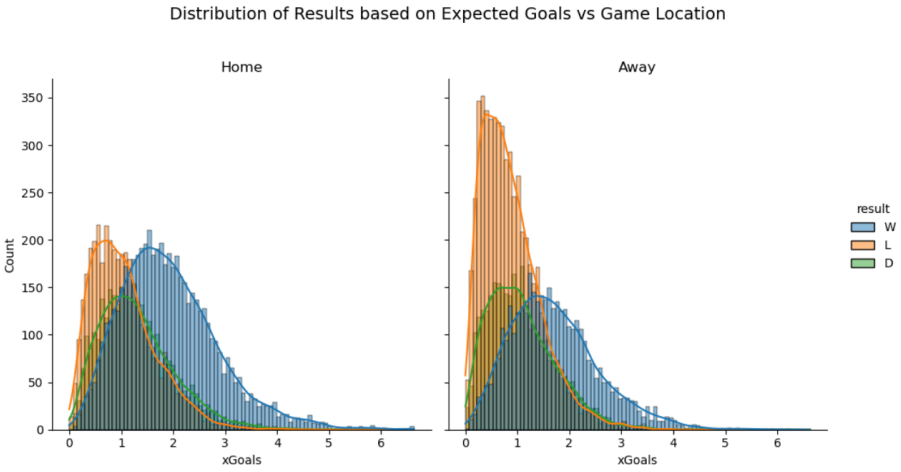


Figura 3: Distribuição de resultados com base interação dos golos esperados e localização do jogo

- **Influência das ações ofensivas/defensivas**: De forma a estudar a influência que os atributos relativos às ações ofensivas de uma equipa, isto é, **deep** e **shotsOnTarget**. Através da figura abaixo é possível observar a influência que um maior volume ofensivo apresenta no resultado do jogo. De forma análoga, mas não com o mesmo grau de influência, é também possível notar que um maior valor de **ppda** tende a conduzir a resultados mais negativos.

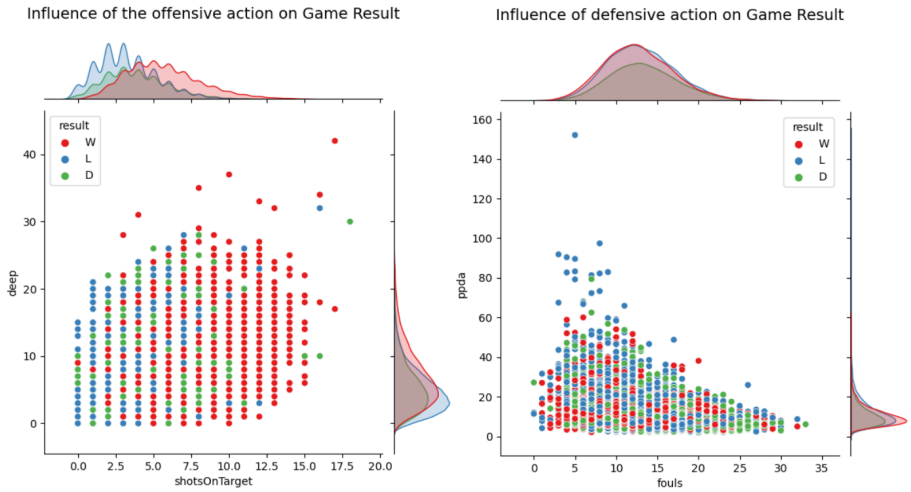


Figura 4: Influência das ações ofensivas e defensivas no Resultado do Jogo

- **Influência do número de cartões admoestados à equipa**: Analisando a figura abaixo, e validado pelo conhecimento do domínio, não é possível notar uma influência clara do número de cartões amarelos no desfecho do jogo. Pelo contrário, observa-se que o cartão vermelho tende a contribuir negativamente para a vitória de equipa.

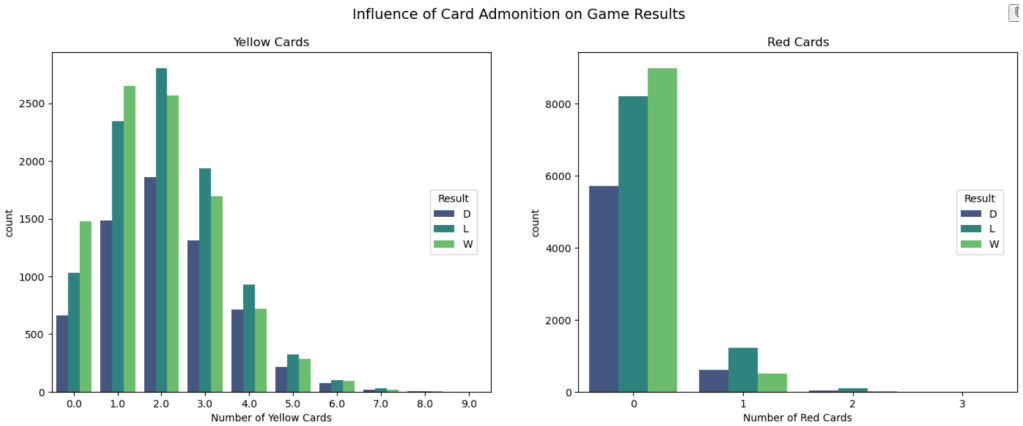


Figura 5: Influência do número de cartões admoestados à equipa

2.3 Data Preparation

A preparação dos dados, constituindo uma etapa determinante no desenvolvimento do modelo, possui um papel fundamental no sucesso ou fracasso do mesmo. Deste modo, apresenta-se os

vários processos aos quais foram sujeitos os dados ingeridos:

- **handling dates** - o atributo **date** que, inicialmente, possuía o formato **%Y%m-%%H:%M:%S**, foi subdividido nos seus elementos constituintes (**year**, **month**, **day**, **hour**, **minutes**, **seconds**), os quais foram, posteriormente, adicionados ao *dataset*, como novos atributos; por fim, decidiu-se remover o atributo que deu origem aos mesmos
- **handling missing data** - o dataset utilizado apenas possuía um *missing value* em todo o seu corpo; preencheu-se este mesmo valor em falta, relativo ao número de cartões amarelos recebidos pela equipa 95 (Roma), no jogo 4888 (vs Juventus), consultando fontes oficiais, mais especificamente o site **zerozero.pt**
- **unique values** - a verificação de valores únicos relativos a cada um dos atributos levou a que se verificasse que o recém adicionado atributo **seconds** apenas possuía um valor (0); esta constatação levou a que o mesmo atributo fosse retirado do *dataset*
- **handling categorical data** - através do método **Label Encoding**, conseguiu-se transformar os valores respetivos aos atributos **result** e **location** em elementos de caráter numérico; deste modo, verificou-se as seguintes alterações em **result**: $W \rightarrow 2$, $L \rightarrow 1$ e $D \rightarrow 0$; relativamente ao atributo **location**, observa-se as seguintes modificações: $h \rightarrow 1$ e $a \rightarrow 0$
- **correlation analysis** - através da análise atenta da matriz de correlação, conseguiu-se perceber a existência de um coeficiente de correlação alto entre a *feature* **year** e as *features* **season** e **gameID**, o que levou a que a primeira fosse removida; por outro lado e, baseada, em grande parte, no conhecimento do domínio, procedeu-se à eliminação do atributos **day** e **minutes**

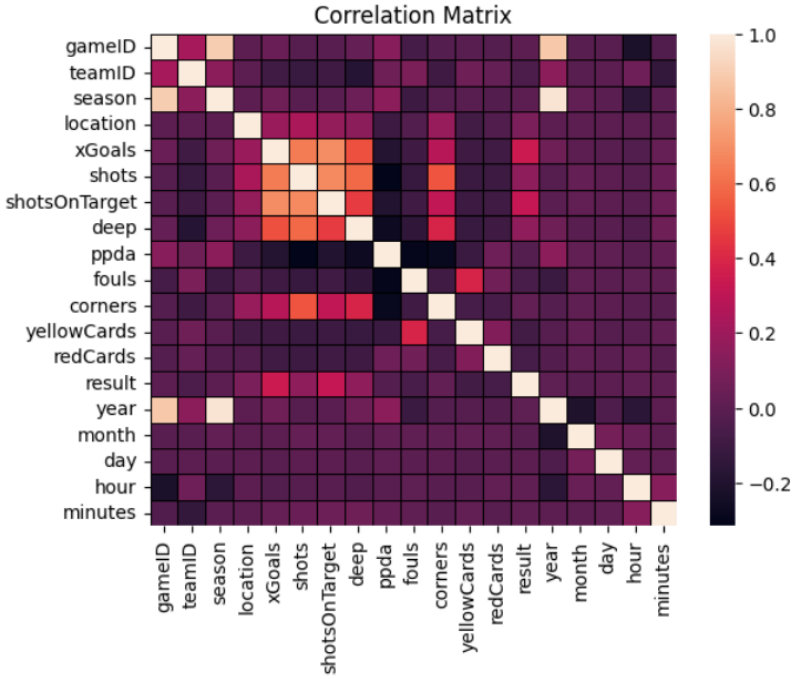


Figura 6: Matiz de Correlação

2.3.1 Feature Importance

De acordo com as características da variável dependente do problema em questão, isto é, a sua classificação como multi classe, foi implementada uma regressão logística multinomial de forma a calcular a importância dos coeficientes para a predição da variável dependente.

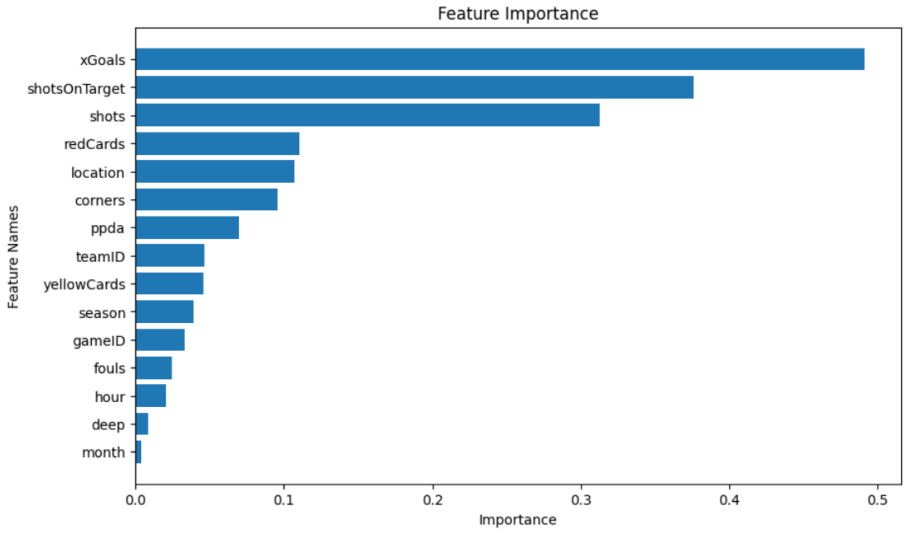


Figura 7: Feature Importance

Analisando a figura acima, é possível observar que as variáveis que apresentam um maior valor de importância na predição são precisamente as que estão relacionadas com a produção ofensiva de uma equipa. O atributo `month` é o que apresenta menor poder preditivo em comparação com os restantes e dessa forma serão efetuadas comparações de desempenho com e sem esta variável , tal como será demonstrado na etapa de avaliação. Realça-se ainda que apesar dos resultados obtidos para o atributo `hour`, o conhecimento do domínio permite concluir que este resultado poderá estar enviesado, uma vez que a hora da realização de um jogo não é um fator determinante para o resultado do mesmo.

2.4 Modeling

Relativamente à etapa relativa à implementação e configuração dos modelos a utilizar, primeiramente foi necessário decidir efetivamente quais os modelos utilizar e respetivos **hiperparâmetros** a serem ajustados para obter a melhor afinação.

2.4.1 Definição de Modelos e respetivos Hiperparâmetros

- **Decision Tree Classifier:**
  - **criterion:** Gini, Entropy
  - **max_depth:** [1,...,10]
  - **min_samples_split:** [2,...,10]
  - **min_samples_leaf:** [1,...,10]

- **Random Forest Classifier:**
  - **criterion:** Gini, Entropy
  - **n_estimators:** 16, 32, 100, 200
  - **max_depth:** [3,...,10]
  - **min_samples_split:** [2,...,10]
- **SVM:**
  - **kernel:** rbf
  - **C:** [1,...,10]
  - **gamma:** 0.1, 0.01, 0.001

Para o treino dos modelos e respetivo *hyperparameter tuning* foi usado o **GridSearch** e a técnica de pesquisa exaustiva com 5 *folds* de *cross validation*, confrontando todas as possibilidades na tentativa de encontrar a melhor combinação de hiperparâmetros que apresentassem melhor desempenho para cada modelo.

2.5 Evaluation

Tratando-se de um problema de classificação, utilizou-se uma **Matriz de Confusão** para avaliação dos resultados dos mesmos. Deste modo e, tendo-se dividido a avaliação do modelo em três fases, passa-se a apresentar os resultados relativos a cada uma das mesmas.

2.5.1 Avaliação do modelo sem *drop* de *features*

Tabela 2.1: *Dataset* sem drop de features

Modelo \ Métrica	Precision	Recall	F1-score	Accuracy	Accuracy (Treino)
Decision Tree	0.54	0.50	0.43	0.56	0.55
Random Forest	0.47	0.50	0.44	0.57	0.57
SVM	0.48	0.51	0.44	0.45	0.57

2.5.2 Avaliação do modelo com *drop* de *features* através da correlação

Tabela 2.2: *Dataset* com *drop* de *features* através da correlação

Modelo \ Métrica	Precision	Recall	F1-score	Accuracy	Accuracy (Treino)
Decision Tree	0.54	0.50	0.43	0.56	0.55
Random Forest	0.47	0.50	0.44	0.57	0.57
SVM	0.48	0.51	0.44	0.45	0.57

2.5.3 Avaliação do modelo com *drop* de *features* através da correlação e da *feature importance*

Tabela 2.3: *Dataset* com *drop* de *features* através da correlação e da *feature importance*

Modelo \ Métrica	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Accuracy (Treino)**
Decision Tree	0.54	0.50	0.43	0.56	0.55
Random Forest	0.47	0.50	0.44	0.57	0.57
SVM	0.48	0.51	0.45	0.57	0.58

Capítulo 3

Tarefa Dataset Competição

Como referido na introdução, o *dataset* para a competição foi atribuído pela equipa docente, diferindo assim da fase anterior, na qual o grupo teve a liberdade de escolher.

3.1 Business Understanding

A energia solar é uma das principais fontes de energias renováveis, desempenhando não só um papel fundamental na transição para fontes de energia limpa e renovável, mas também na promoção da sustentabilidade ambiental. Para além de ser crucial otimizar o uso da energia solar, a relação entre o gasto e a produção energética é essencial para permitir um planeamento eficaz do consumo energético e a integração harmoniosa de sistemas de energia solar em redes elétricas existentes.

Uma vez que a quantidade de energia produzida por estas fontes de energia pode variar consoante vários fatores, como dados meteorológicos ou geográficos,é bastante importante conseguir desenvolver modelos que permitam prever qual a quantidade de energia produzida consoante esses mesmos fatores. Esta previsão permite, não só otimizar o desempenho destas fontes, aproveitando o máximo de energia possível, mas também a redução de custos, que se mostra também crucial, uma vez que a escolha de alternativas sustentáveis tem tido um grande aumento nos últimos anos.

3.2 Data Understanding

3.2.1 Ingestão dos Dados

- Dataset Fornecido: **Produção Energética**

O *dataset* selecionado pelos docentes contém dados referentes à **produção energética de determinados painéis solares na cidade de Braga**, cobrindo um período que vai desde setembro de 2021 até abril de 2023.

Estes dados estão representados em **dois tipos de *datasets***, um tipo com dados energéticos e outro tipo com dados meteorológicos. Além disso, cada tipo de *dataset* está dividido por data, um cobre o período de 2021 e outro cobre todo o ano de 2022.

Datasets de Energia

```
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Data                    2256 non-null   object
 1   Hora                    2256 non-null   int64
 2   Normal (kWh)            2256 non-null   float64
 3   Horario Economico (kWh) 2256 non-null   float64
 4   Autoconsumo (kWh)       2256 non-null   float64
 5   Injecao na rede (kWh)   566 non-null    object
```

Figura 8: Atributos dos Datasets de Energia

Como observável na figura acima, estes *datasets* apresentam 6 atributos, dos quais 4 são numéricos e 2 são categóricos.

- **Data:** *Timestamp* associado ao registo, ao dia.
- **Hora:** Hora associada ao registo.
- **Normal (kWh):** Quantidade de energia elétrica consumida, em kWh, e proveniente da rede elétrica, num período considerado normal em ciclos bi-horário diários (horas fora de vazio).
- **Horario Economico (kWh):** quantidade de energia elétrica consumida, em kWh e proveniente da rede elétrica, num período considerado económico em ciclos bi-horário diários (horas de vazio).
- **Autoconsumo (kWh):** quantidade de energia elétrica consumida, em kWh, proveniente dos painéis solares.
- **Injecao na rede (kWh):** (**Variável Dependente**) quantidade de energia elétrica injetada na rede elétrica, em kWh, proveniente dos painéis solares.

Datasets Meteorológicos

```
Data columns (total 15 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   dt                   2928 non-null   int64
 1   dt_iso               2928 non-null   object
 2   city_name            2928 non-null   object
 3   temp                 2928 non-null   float64
 4   feels_like           2928 non-null   float64
 5   temp_min             2928 non-null   float64
 6   temp_max             2928 non-null   float64
 7   pressure             2928 non-null   int64
 8   sea_level            0 non-null      float64
 9   grnd_level           0 non-null      float64
 10  humidity             2928 non-null   int64
 11  wind_speed           2928 non-null   float64
 12  rain_1h              537 non-null    float64
 13  clouds_all           2928 non-null   int64
 14  weather_description  2928 non-null   object
```

Figura 9: Atributos dos Datasets Meteorológicos

Relativamente aos *datasets* meteorológicos, estes apresentam 15 atributos, sendo 12 numéricos e 3 categóricos

- **dt:** *Timestamp* associado ao registo.
- **dt_iso:** Data associada ao registo, ao segundo.
- **city_name:** Local em causa.
- **temp:** Temperatura em ºC.
- **feels_like:** Sensação térmica em ºC.
- **temp_min:** Temperatura mínima sentida em ºC.
- **temp_max:** Temperatura máxima sentida em ºC.
- **pressure:** Pressão atmosférica sentida em atm.
- **sea_level:** Pressão atmosférica sentida ao nível do mar em atm.
- **grnd_level:** Pressão atmosférica sentida à altitude local em atm.
- **humidity:** Humidade em percentagem.
- **wind_speed:** Velocidade do vento em metros por segundo.
- **rain_1h:** Valor médio de precipitação.
- **clouds_all:** Nível de nebulosidade em percentagem.
- **weather_description:** Avaliação qualitativa do estado do tempo.

Concatenação Datasets

Após analisados os atributos de cada tipo de *dataset* o grupo começou por realizou a concatenação dos quatro *datasets*. No processo de concatenação, o atributo **dt_iso** foi excluído, ficando então o úunico *dataset* constituído pelos atributos apresentados na imagem seguinte.

```
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Data                    2256 non-null   object
 1   Hora                    2256 non-null   int64
 2   Normal (kWh)            2256 non-null   float64
 3   Horario Economico (kWh) 2256 non-null   float64
 4   Autoconsumo (kWh)       2256 non-null   float64
 5   Injecao na rede (kWh)   566 non-null    object
 6   dt                      2256 non-null   int64
 7   city_name               2256 non-null   object
 8   temp                    2256 non-null   float64
 9   feels_like              2256 non-null   float64
 10  temp_min                2256 non-null   float64
 11  temp_max                2256 non-null   float64
 12  pressure                2256 non-null   int64
 13  sea_level               0 non-null      float64
 14  grnd_level              0 non-null      float64
 15  humidity                2256 non-null   int64
 16  wind_speed              2256 non-null   float64
 17  rain_1h                 386 non-null    float64
 18  clouds_all              2256 non-null   int64
 19  weather_description     2256 non-null   object
```

Figura 10: Atributos do Dataset resultante da concatenação

3.2.2 Exploração e Visualização dos Dados

Uma primeira análise baseou-se na visualização do balanceamento dos dados, particularmente a distribuição do atributo **Injeção na rede (kWh)** nos dois anos que se estudaram (2021,2022). A análise do seguinte gráfico permite assegurar que os dados não se encontram totalmente balanceados, sendo notória uma maior presença de valores **None** em relação aos demais.

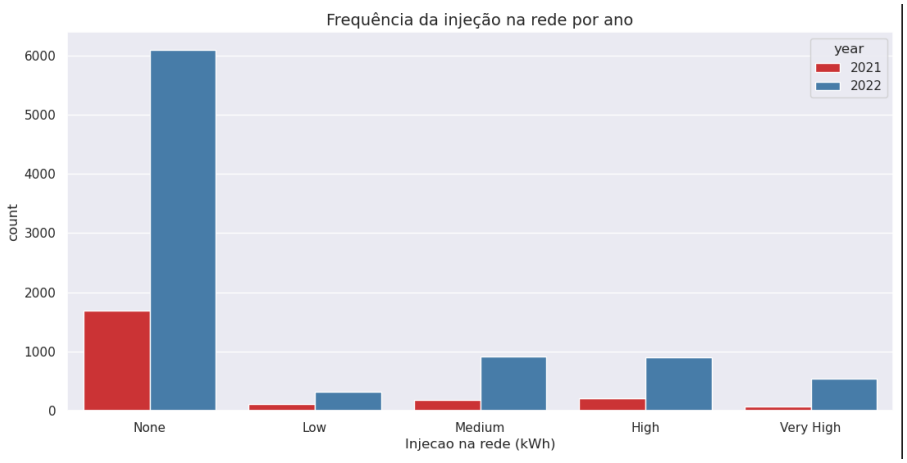


Figura 11: Distribuição da frequência de injeção na rede por ano

Posteriormente, foi necessário verificar a relação e influência de algumas variáveis independentes com a variável dependente.

Influência e interação das variáveis independentes na Injeção na rede

Relação entre a distribuição da nebulosidade e a injeção na rede ao longo dos meses: Através da análise do seguinte gráfico é percetível a influência do nível de nebulosidade com a injeção na rede. Tornando-se num facto trivial dado que esta é efetuada através de painéis solares.

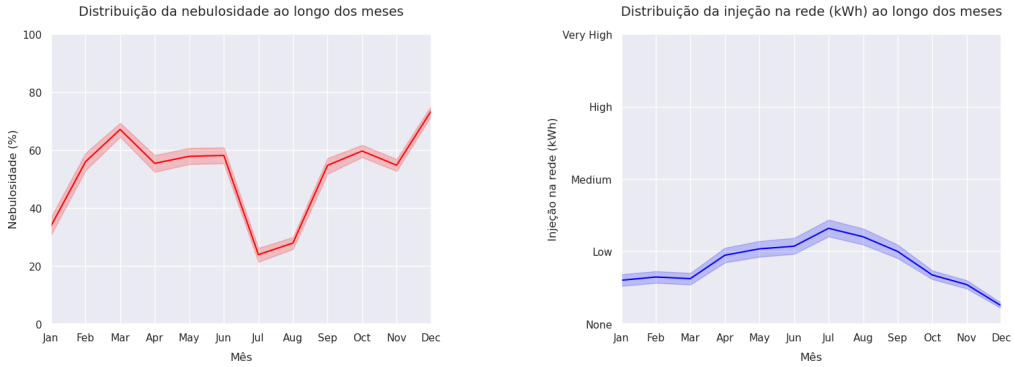


Figura 12: Distribuições da nebulosidade e injeção na rede ao longo dos meses

Influência da nebulosidade e humidade na injeção na rede

Tendo conhecimento do gráfico previamente descrito, seria possível prever que a nebulosidade e a humidade influenciariam de um modo significativo a injeção na rede. O facto de serem duas características meteorológicas que condicionam a injeção na rede através de painéis solares, é fácil entender que o seu comportamento em relação à injeção na rede será bastante semelhante. Como se pode observar na seguinte figura, os níveis de nebulosidade e humidade são inversamente proporcionais aos valores de injeção na rede, por exemplo, quanto menor a percentagem de nebulosidade, maior será a injeção na rede.

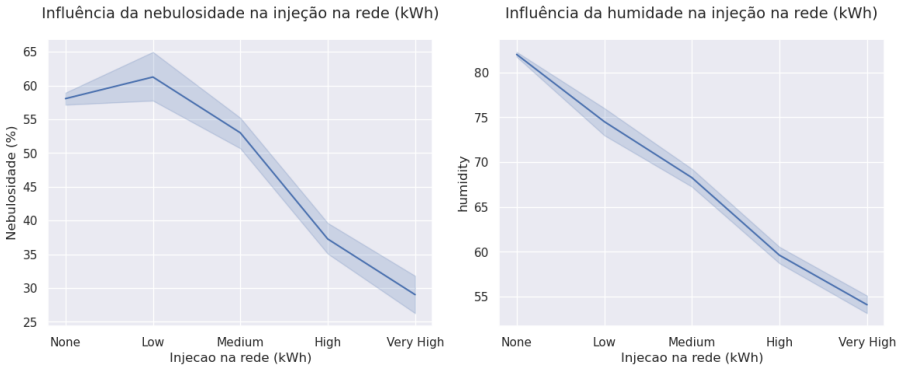


Figura 13: Influência no nível de nebulosidade e humidade na Injeção na rede

Distribuição de Injeção na rede por ano de acordo com variações de temperatura

No seguinte gráfico, o principal objetivo passou por analisar a influência da temperatura na injeção na rede, nos dois anos de estudo em separado. Analisando o gráfico pode-se observar que temperaturas mais amenas (entre os 20^{o} e 25^{o} graus) favorecem de uma forma mais positiva a injeção na rede do que temperaturas muito baixas ou muito altas.

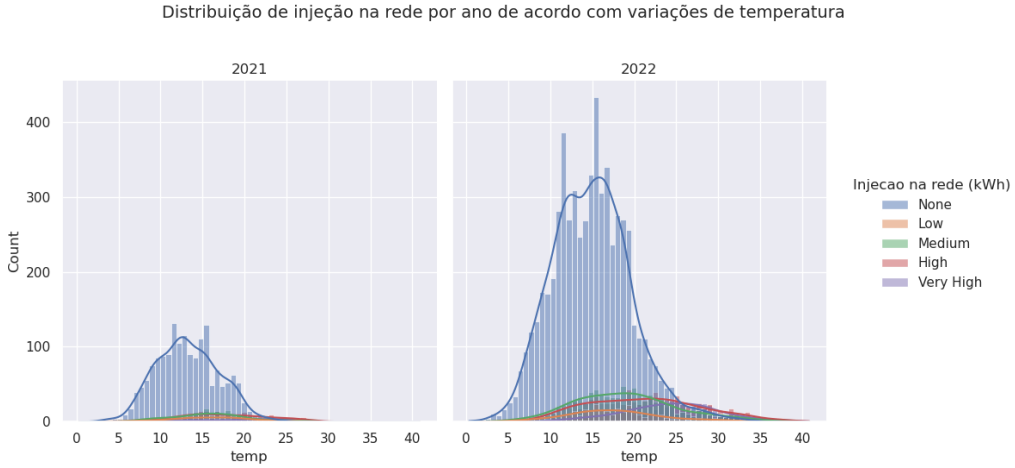


Figura 14: Distribuição de Injeção na rede por ano de acordo com variações de temperatura

3.3 Data Preparation

Após a visualização dos dados e a sua análise detalhada, foi altura de passar para a preparação dos mesmo. Uma vez mais, esta processo constitui um papel fundamental para o desenvolvimento do(s) modelo(s), apresentando grande influência no resultado final.

Antes de explicar os processos realizados nesta fase, é importante notar que as informações que foram descritas no capítulo da ingestão de dados em relação à concatenação dos *datasets* pode também ser consideradas neste capítulo porém, para tentar detalhar a informação de forma coerente e de maneira a que a interpretação do problema seja mais fácil e intuitiva, essas modificações foram apresentadas nesse capítulo.

Deste modo, apresentam-se os processos realizados:

- **handling missing data** - todos os valores das colunas **sea_level** e **grnd_level** são nulos, pelo que as colunas foram excluídas. O atributo **Injecao na rede (kWh)** continha 7777 *missing values*, devido ao facto de terem sido interpretados como tal aquando da leitura, que foram substituídos por *None*. Por último, preencheram-se os *missing values* do atributo **rain_1h** com o valor 0.0.
- **unique values** - a verificação de valores únicos relativos a cada um dos atributos levou a que se excluísse o atributo **city_name**, sendo que apenas tinha 1 *unique value*.
- **feature engineering** - Foi calculado o consumo total de energia (**Consumo Total (kWh)**), sendo a soma do autoconsumo com o consumo em horário económico e com o consumo em horário normal. Foi calculada também a taxa de autoconsumo (**Taxa Autoconsumo**) como sendo a divisão do autoconsumo pelo consumo total, no caso do consumo total não ser 0 (se for 0, então a taxa também é 0). Foi criado também o atributo **is_weekend** para identificar os fins de semana. Por último, a previsão da injeção (**Previsao Injecao**) foi calculada, utilizando a diferença entre o autoconsumo e a soma dos consumos em horário normal e económico.

3.3.1 Feature Importance

Tendo em conta as características da variável em estudo, foi importante estabelecer a sua relação com as restantes variáveis independentes de modo a verificar o nível de influência que estas demonstravam ter na variável dependente. Para tal foi implementada uma regressão logística multinomial de forma a calcular a importância dos coeficientes para a predição da variável dependente.

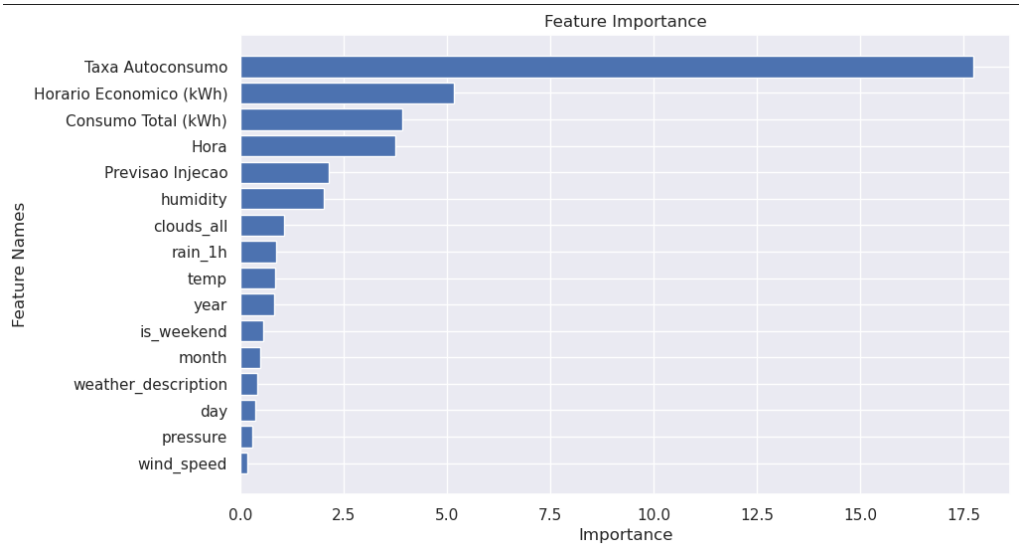


Figura 15: Feature Importance

Através da análise do gráfico acima referido é possível observar que as variáveis independentes que apresentam um valor mais significativo em termos de importância são aqueles que estão diretamente ligados com o consumo da energia. Por outro lado, o atributo **wind_speed** é o que apresenta menor poder preditivo em comparação com os demais.

3.4 Modeling

Nesta fase, um dos pontos fundamentais foi a escolha de quais modelos usar e que **hiperparâmetros** utilizar em cada modelo, de modo a obter a melhor afinação possível.

3.4.1 Definição de Modelos e respetivos Hiperparâmetros

- **Decision Tree Classifier:**
  - **criterion:** Gini, Entropy
  - **max_depth:** [1,...,10]
  - **min_samples_split:** [2,...,10]
  - **min_samples_leaf:** [1,...,5]
- **Random Forest Classifier:**
  - **criterion:** Entropy
  - **n_estimators:** 16, 32, 100
  - **max_depth:** [1,...,10]
  - **min_samples_split:** [2,...,10]
- **SVM:**
  - **kernel:** rbf,poly,sigmoid

– **C:** [0.1,1,10,100]
– **gamma:** 1,0.1,0.01,0.001

Para o treino dos modelos e respetivo *hyperparameter tuning* foi usado o **GridSearch** e a técnica de pesquisa exaustiva com um número variado de *folds* de *cross validation*, confrontando todas as possibilidades na tentativa de encontrar a melhor combinação de hiperparâmetros que apresentassem melhor desempenho para cada modelo.

3.5 Evaluation

Relativamente à medição do desempenho dos modelos implementados, irão ser apresentados os resultados mais significantes relativos ao treino e à respetiva submissão na plataforma da competição. Para além disso, e apesar da definição de vários modelos, o **Random Forest Classifier** foi o que desde início apresentou melhores resultados e dessa forma apenas serão mostrados os resultados associados a este modelo.

Tabela 3.1: Medição de Desempenho

Modelo \ Métrica	**Accuracy (Treino)**	**Accuracy (Teste)**	**Nº Folds**
(1)	0.85785	0.84023	10
(2)	0.86166	0.8565	10
(3)	0.86475	0.84171	10
(4)	0.86620	0.84763	10
(5)	0.86865	0.86834	10
(6)	0.86892	0.83727	10
(7)	0.86992	0.85059	10

De seguida são descritas as alterações efetuadas no *dataset* relativamente ao drop de atributos.

- **(1)** `wind_speed, pressure, day, feels_like, temp_min, temp_max, dt`
- **(2)** `wind_speed, pressure, day, weather_description, year, feels_like, temp_min, temp_max, dt`
- **(3)** `win_speed, pressure, day, weather_description, year, month, feels_like, temp_min, temp_max, dt`
- **(4)** `win_speed, pressure, day, weather_description, year, feels_like, temp_min, temp_max, dt`
- **(5)** `igual ao (6) mas os dados não foram escalados`
- **(6)** `feels_like, temp_min, temp_max, dt, Autoconsumo (kWh), Normal (kWh), wind_speed, pressure, day, weather_description, month`
- **(7)** `wind_speed, pressure, day, year, month, feels_like, temp_min, temp_max, dt`

Capítulo 4

Conclusão

Em conclusão, o desenvolvimento deste projeto de **Machine Learning**, que envolveu a análise de dois conjuntos de dados distintos - um escolhido pelo grupo através da plataforma Kaggle e outro atribuído pela equipa docente - proporcionou uma experiência rica e desafiadora. Ao longo do processo, aplicaram-se técnicas e conceitos aprendidos ao longo das aulas, desde a seleção e preparação dos dados até à implementação e avaliação de diversos modelos de aprendizagem automática. A metodologia adotada, seguindo o **CRISP-DM** (Cross-Industry Standard Process for Data Mining), permitiu uma abordagem sistemática e iterativa, garantindo uma estrutura organizada para o desenvolvimento do projeto. Enfrentaram-se desafios significativos, como a escolha e engenharia de recursos relevantes, a gestão de valores ausentes e a seleção adequada de modelos para cada tarefa proposta. A análise dos resultados foi fundamental para ajustar e melhorar continuamente os modelos, proporcionando uma compreensão mais profunda dos conjuntos de dados. Este projeto destacou a importância da adaptação às características específicas dos dados e a forma como se conseguiu lidar com elas, ganhando conhecimento no domínio do problema. Além disso, reforçou a relevância de uma abordagem metodológica, não apenas para atingir resultados precisos, mas também para garantir a eficácia e a qualidade ao longo de todo o processo. Em suma, ao explorar e aplicar as técnicas de **Machine Learning** aprendidas, conseguiu-se extrair *insights* valiosos e enfrentar desafios complexos. Este projeto representa não apenas uma aplicação prática do conhecimento adquirido, mas também um passo significativo no desenvolvimento de habilidades e compreensão na área de **Data Science** e **Machine Learning**. Deste modo, encerra-se o atual relatório relativo ao projeto desenvolvido no âmbito da unidade curricular de **Dados e Aprendizagem Automática**.