Dados e Aprendizagem Automática

Relatório do Trabalho Prático

MEI - 2023/2024

Grupo 16

Hugo Martins
A95125

João Escudeiro
A96075

Afonso Bessa
PG53597

Ivo Ribeiro
PG53886

Universidade do Minho

Lista de Figuras

Lista de Tabelas

Resumo

Este documento apresenta de forma concisa os objetos de avaliação e análise de um projeto inserido na Unidade Curricular Dados e Aprendizagem Automática. Os principais objetivos deste projeto incluem a análise, o processamento e a previsão de dados a partir de dois conjuntos de dados distintos. Nos vários capítulos e secções deste relatório, são detalhadas todas as decisões tomadas pela equipa de trabalho em relação aos métodos escolhidos para atingir os objetivos do projeto.

Para ambos os *datasets*, foi adotada a seguinte estratégia, resumida pela figura seguinte:

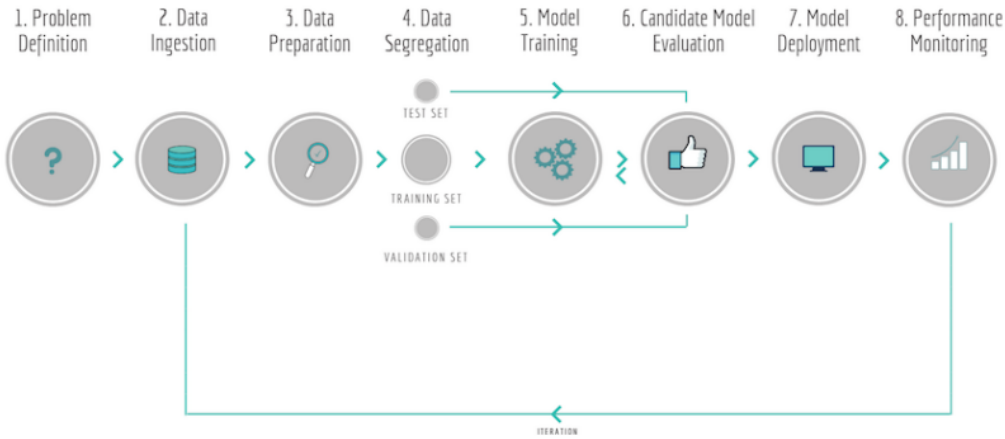


Figura 1: *Machine Learning Pipeline*

DataSet Grupo

Caracteristicas do DataSet SmokingAndDrinking

Após seleção e avaliação de diferentes conjuntos de dados, o grupo optou por um *dataset* cujo conteúdo aborda a temática de pessoas fumadoras e não fumadoras, bem como, os indicadores de fraqueza manifestados pelo corpo e a identificação de possíveis doenças.

O objetivo principal é analisar vários tipos de sinais apresentados pelo corpo, como hipertensão, fadiga, problemas de visão ou audição, procurando relacionar o surgimento desses problemas com o consumo regular de tabaco ou de bebidas alcoólicas.

Este *dataset* foi recolhido de *National Health Insurance Service* na Coreia. Antes de ser dado início a qualquer análise dos dados, realizou-se uma pesquisa prévia do problema com o objetivo de adquirir um conhecimento mais completo sobre o conjunto de dados selecionado. Para alcançar este objetivo, consultou-se a fonte do dataset, onde se encontraram as descrições detalhadas das colunas, o que facilita a interpretação das informações contidas neste ficheiro CSV.

Assim, ***SmokingAndDrinking*** apresenta 991346 linhas e 24 *features*, *features* essas que são explicadas através das tabela seguinte:

Tabela 1: Data Description SmokingAndDrinking

Number	Name	Description
0	sex	Sex of the individual (male or female)
1	age	Age of the individual (rounded up to 5 years)
2	height	Height of the individual (rounded up to 5 cm)
3	weight	Weight of the individual
4	waistline	Circumference of the individual's waist
5	sight_left	Visual acuity of the individual's left eye (ranging from 0.1 to 2.5, with values ¡0.1 set to 0.1)
6	sight_right	Visual acuity of the individual's right eye
7	hear_left	Hearing in the left ear of the individual (1 for normal, 2 for abnormal)
8	hear_right	Hearing in the right ear of the individual (1 for normal, 2 for abnormal)
9	SBP	Highest systolic blood pressure measured from the individual (mmHg)
10	DBP	Diastolic blood pressure measured from the individual (mmHg)
11	BLDS	Fasting blood glucose of the individual (mg/dL)
12	tot_chole	Total cholesterol in the individual (mg/dL)
13	HDL_chole	Cholesterol in the HDL (high density lipoprotein) region (mg/dL)
14	LDL_chole	Cholesterol in the LDL (low density lipoprotein) region (mg/dL)
15	triglyceride	Concentration of triglycerides in the individual's blood (mg/dL)
16	hemoglobin	Concentration of hemoglobin in the individual's blood (g/dL)
17	urine_protein	Amount of protein in the individual's urine (1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4))
18	serum_creatinine	Concentration of creatinine in the individual's serum (mg/dL)
19	SGOT_AST	SGOT (Glutamate-Oxaloacetate Transaminase) - AST (Aspartate Transaminase) value in IU/L
20	SGOT_ALT	SGOT (Glutamate-Oxaloacetate Transaminase) - ALT (Alanine Transaminase) value in IU/L
21	gamma_GTP	Gamma-Glutamyl Transpeptidase (y-glutamyl transpeptidase) value in IU/L
22	SMK_stat_type_cd	Smoking state of the individual (1(Never), 2(Used to Smoke but Quit), 3(Still Smoke))
23	DRK_YN	Drinker or Not

Nota 1:
1 dL, ou 1 decilitro, equivale a 100 mL (mililitros) ou 0,1 L (litros).

Nota 2:
IUs, ou unidades internacionais, são unidades de medida usadas para quantificar o efeito/atividade biológica de uma substância. (fonte: Wikipedia).

Análise Exploratória de Dados

Após carregarmos o *dataset*, foi realizado um amostragem dos dados, uma vez que, tal como foi mencionado anteriormente, o conjunto de dados possui cerca de um milhão de linhas. Assim, realizamos uma redução dos dados para 25.000 linhas.

De seguida, iniciamos a Análise Exploratória de Dados, que envolveu uma investigação mais detalhada das informações presentes nas colunas mencionadas anteriormente. Utilizando comandos como o `describe()` ou o `unique()` obtivemos estatísticas resumidas, como médias, desvios padrão, quartis e muitos outros indicadores relevantes, que nos ajudaram a compreender a natureza e distribuição dos dados.

Algumas das conclusões que conseguimos observar a partir dos dados são as seguintes:

- Duas *features* categóricas (*sexo* e *drinker_or_not*), enquanto todas as restantes são numéricas.
- A variável *Idade* varia entre 20 e 85 anos.
- A variável *Altura* varia entre 130 e 190 cm.
- A variável *Peso* tem um peso mínimo de 25 kg, sendo necessário corrigir esse valor posteriormente.
- A variável *Waistline* possui um valor máximo de 999 cm, o que parece claramente incorreto e impossível.
- As variáveis *sight_left* e *sight_right* têm valores máximos de 9.9, o que pode indicar algum tipo de erro ou valor adverso.
- Por fim, várias variáveis relacionadas com o corpo de cada utente, como *systolic_blood_pressure*, *diastolic_blood_pressure*, *BLDS*, *total_cholesterol*, e muitas outras apresentam valores máximos adversos e impossíveis de ocorrer em seres humanos.

Todas estas observações são essenciais para o Processamento de Dados, uma vez que as características em questão requerem correções e verificações adicionais antes de prosseguir para qualquer modelo de *Machine Learning*.

Encoding Inicial

Posteriormente, convertemos as duas *features* previamente identificadas como *object* em variáveis numéricas.

```
# Map the values in the "sex" column to numeric values using a dictionary.
# Replace 'Male' with 1 and 'Female' with 0.
df["sex"] = df["sex"].map({'Male': 1, 'Female': 0})

# Map the values in the "drinker_or_not" column to numeric values using a dictionary.
# Replace 'Y' with 1 and 'N' with 0.
df["drinker_or_not"] = df["drinker_or_not"].map({'Y': 1, 'N': 0})
```

Da mesma forma, transformamos a variável `smoking_state` de 1, 2 e 3, correspondendo, respetivamente, a *Nunca*, *Costumava Fumar mas Parou* e *Ainda Fuma*, para 1 e 2, onde 1 corresponde a *Nunca*, e reunimos toda a informação em 2, que corresponde a *Fumar*. Essa transformação foi realizada com o propósito de simplificar a análise dos dados, bem como, equilibrar os dados da variável dado que os mesmos estavam poucos balenceados.

Verificação da Qualidade dos Dados

Continuamente, realizamos uma Análise da Qualidade dos Dados para identificar a presença de valores em falta e linhas duplicadas. Essas verificações permitiram-nos concluir que o conjunto de dados está completo uma vez que não existem quaisquer registos duplicados. Além disso, não foram identificados quaisquer registos de *missing values*.

Divisão em Duas *Datasets*

Ao término da Análise Exploratória de Dados, optamos por dividir o conjunto de dados em dois (*Drinking.csv* e *Smoking.csv*). Isso deve-se ao facto de, após várias tentativas de combinar o grupo das duas variáveis, os resultados obtidos foram consideravelmente inferiores no que toca à correlação entre *feature* em relação à manutenção das variáveis separadamente. Dessa forma, decidimos realizar análises distintas para cada um dos novos conjuntos de dados, mantendo, no entanto, a mesma estrutura para ambas as análises.

As alternativas de *enconding* que o grupo optou mas não deram resultado foi:

1. Fumador e Bebedor
2. Fumador e Não Bebedor
3. Não Fumador e Bebedor
4. Não Fumador e Não Bebedor

1. Fumador e Bebedor
2. Fumador e Não Bebedor
3. Não Fumador e Bebedor
4. Não Fumador e Não Bebedor
5. Fumou mas Parou e Bebedor
6. Fumou mas Parou e Não Bebedor

Exploração dos Dados Antes da Pré-Processamento dos Dados

Análise de Correlação

Antes de passarmos para o Pré-Processamento de Dados, procedemos ao cálculo da matriz de correlação entre todas as variáveis do *dataset*, proporcionando assim uma compreensão das relações existentes entre as diferentes variáveis, e ao cálculo da *Mutual Information* de maneira a avaliar a dependência estatística entre as variáveis.

Nesta fase, foi possível verificar dependências entre *features*, bem como, o seu índice de correlação com a *target*.

Drinking No caso do Drinking, foi possível eliminar as seguintes colunas, pois apresentavam valores de correlação entre $]-0.1, 0.1[$ e *mutual information* entre menor que 0.005. Efetuamos imediatamente esta exclusão, reconhecendo que a remoção de variáveis com baixa correlação e Informação Mútua insignificante contribui para simplificar o conjunto de dados, facilitando uma análise mais eficaz e focalizada nas características mais relevantes para os objetivos do estudo.

```
Columns to Drop: ['hear_left', 'systolic_blood_pressure', 'BLDS', 'total_cholesterol', 'HDL_cholesterol', '
    LDL_cholesterol', 'urine_protein', 'SGOT_AST', 'SGOT_ALT']
```

Smoking No caso do Drinking, fizemos o mesmo processo porém eliminamos features com valores de correlação e *mutual information* de $]-0.1, 0.1[$ e $]-0.025, 0.025[$, respetivamente. As colunas removidas foram:

```
Columns to Drop: ['age', 'sight_left', 'sight_right', 'hear_left', 'hear_right', 'BLDS', 'total_cholesterol', '
    LDL_cholesterol', 'urine_protein', 'SGOT_AST']
```

Logo, após esta remoção inicial, identificamos algumas variáveis presentes em ambos os *datasets*, mas também observamos algumas diferenças, o que resultará em resultados distintos para cada uma das *targets*.

Análise Estatística Descritiva

Através do uso de diversas ferramentas de visualização de dados, como os gráficos *boxplot*, *scatter* e *pie charts*, importados da biblioteca *matplotlib*, conseguimos realizar uma análise estatística descritiva abrangente dos dados antes de qualquer etapa de Pré-Processamento. Essas visualizações permitiram-nos obter *insights* valiosos sobre a distribuição, tendências e características do conjunto de dados.

Por meio dos gráficos *boxplot*, identificámos a presença de potenciais valores extremos em várias variáveis, o que nos alertou para a necessidade de tratamento. Os gráficos *scatterplot* proporcionaram uma

representação visual das relações entre diferentes variáveis, destacando possíveis associações ou correlações entre elas. Além disso, os *pie charts* permitiram-nos observar a distribuição das categorias em variáveis categóricas, fornecendo uma visão geral da proporção de cada categoria. Adicionalmente, utilizámos *count-plots* para analisar a contagem de ocorrências em variáveis específicas, oferecendo uma perspetiva clara da distribuição dos dados em relação às diferentes categorias.

Esta análise estatística descritiva inicial foi fundamental para compreender a estrutura dos dados, identificando áreas que requerem intervenção. Destacam-se, por exemplo, a deteção de *outliers* e a análise de disparidades entre os valores associados aos sexos masculino e feminino, proporcionando *insights* valiosos.

Binning

Para aprofundar a nossa compreensão dos dados, recorremos à técnica de *binning*, que consiste na agrupação de valores contínuos em intervalos discretos, ou *bins*. Esta abordagem revela-se particularmente útil quando lidamos com variáveis numéricas contínuas e pretendemos simplificar a análise, destacando padrões ou tendências globais.

No contexto dos nossos conjuntos de dados, aplicamos o *binning* a variáveis específicas para categorizar faixas de valores, proporcionando uma visão mais clara e interpretável das distribuições. Esta técnica não só simplifica a análise, como também pode evidenciar padrões que poderiam passar despercebidos ao lidar com dados contínuos.

Pré-Processamento dos Dados

Esta etapa foi uma das fases que exigiu um maior tempo despendido e, consequentemente, uma taxa de trabalho mais elevada por parte do grupo, devido aos *outliers*, muitos deles encontrados nas *features*. Para tratar esses valores atípicos, foi essencial realizar uma pesquisa abrangente, que incluiu a exploração de informações científicas e a análise de dados disponíveis na *internet*. Esse processo meticuloso visava obter o melhor entendimento possível dos dados, possibilitando a produção de resultados mais precisos e confiáveis.

Drinking

Inicialmente, realizamos **Feature Engineering** ao incorporar *BMI*, que representa o Cálculo do Índice de Massa Corporal. O *BMI* foi determinado para cada entrada, utilizando as informações das colunas *weight* e *height*, com a conversão da altura de centímetros para metros. Da mesma maneira, foram criadas duas novas colunas, *eye_sight_left* e *eye_sight_right*, que classificam a visão com base nos valores das colunas *sight_left* e *sight_right*. As classificações são *Good* para visão melhor que 20/20, *Average* para visão melhor que 20/40, *Poor* para visão melhor que 20/80, e *Very Poor* para os demais casos.

Após isso, fizemos a filtragem e correção de dados:

1. Substituição dos valores de 9.9 nas colunas *sight_left* e *sight_right* por 0, representando a condição de cegueira. Esta escolha é baseada numa escala, onde 0.1 é considerado uma má visão, 1.0 representa uma visão média e 2.0 indica uma visão perfeita. Portanto, ao substituir 9.9 por 0, estamos a normalizar os dados para refletir a ausência de visão, contribuindo para uma representação mais precisa e coerente das condições visuais no conjunto de dados.
2. Foram mantidas apenas as entradas onde a coluna *waistline* é menor ou igual a 200, removendo os valores excessivos falados anteriormente.
3. Foram removidas as entradas onde a *diastolic_blood_pressure* ultrapassa 150.
4. Foram mantidas apenas as entradas onde a *hemoglobin* são maiores ou iguais a 10.
5. Foram mantidas apenas as entradas onde a *serum_creatinine* é menor ou igual a 1.5.
6. Foram mantidas apenas as entradas onde a *gamma-glutamyl_transpeptidase* é menor ou igual a 600.

Smoking

Da mesma maneira, realizamos **Feature Engineering** ao incorporar *BMI*, porém adicionamos, também, uma análise mais abrangente do sistema cardiovascular ao criar a variável *blood_pressure*, que categoriza os níveis de pressão arterial com base nos critérios estabelecidos. Além disso, introduzimos a variável *numerical_BP*, calculada como a razão entre a pressão sistólica e diastólica, fornecendo uma perspetiva numérica adicional sobre a saúde cardiovascular.

Após isso, fizemos a filtragem e correção de dados:

1. Foram mantidas apenas as entradas onde a coluna *waistline* é menor ou igual a 200, tal como no dataset anterior.
2. Foram removidas as entradas onde a *diastolic_blood_pressure* ultrapassa 150.
3. Foram mantidas apenas as entradas onde a *hemoglobin* são maiores ou iguais a 10.
4. Foram mantidas apenas as entradas onde a *serum_creatinine* é menor ou igual a 1.25.
5. Foram mantidas apenas as entradas onde a *gamma-glutamyl_transpeptidase* é menor ou igual a 100.
6. Foram mantidas apenas as entradas onde a *HDL_cholesterol* é menor ou igual a 175 e maior ou igual a 15.
7. Foram removidas as entradas onde a *systolic_blood_pressure* ultrapassa 200.
8. Foram removidas as entradas onde a *SGOT_ALT* ultrapassa 800.

A solução apresentada em cima para o Pré-Processamento das variáveis, nos dois casos, foi a que, após diversas tentativas, atingiu o melhor resultado. Algumas tentativas abordadas pelo grupo foram:

1. **Eliminação dos *Outliers***: Identificámos os valores discrepantes calculando o **IQR** (*Interquartile Range*) e eliminámos todos esses valores. Variámos o **IQR** de 1,5 para 3, 4,5 e 6; no entanto, os valores obtidos em todos os casos não ultrapassavam a solução apresentada.
2. **Transformação de Dados**: Aplicamos transformações matemáticas, como logaritmo ou raíz quadrada, não só para reduzir a influência de *outliers*, mas também para abordar a assimetria na distribuição dos dados, medida pelo *skewness*. Estas transformações ajudam a tornar a distribuição mais próxima de uma distribuição normal, facilitando análises estatísticas mais robustas, no entanto, o resultado não foi positivo.

Skew	Moderate	High	(Higher)	Extreme
Positive (right tail)	Square root transformation	Natural log transformation	Log base 10 transformation	Inverse transformation
Negative (left tail)	Reflect then square root transformation	Reflect then natural log transformation	Reflect then log base 10 transformation	Reflect then inverse transformation

Figura 2: *Skews*

3. **Substituição por Estatísticas Robustas**: Substituímos os valores *outliers* por estatísticas, como, por exemplo, a mediana ou a média mas, novamente, o resultado foi inferior.

Exploração dos Dados Depois da Pré-Processamento dos Dados

Seguidamente, após a adição de novas *features* e a realização de modificações em outras já existentes, torna-se imperativo avaliar o impacto dessas alterações no conjunto de dados. Dessa forma, reexaminamos a correlação e a *mutual information*, ao mesmo tempo em que visualizamos todos os dados disponíveis.

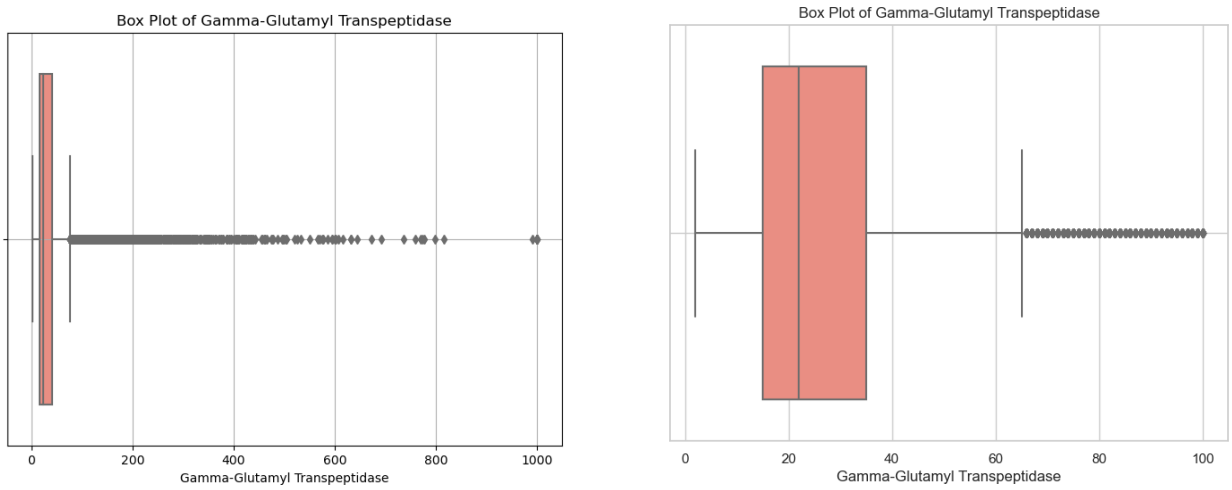
Análise de Correlação

Por um lado, verificamos que as novas features adicionadas não apresentaram um valor significativo de correlação com a variável alvo, seja ela *Drinking* ou *Smoking*. Logo, antes de aplicarmos modelos serão algumas das *features* que vão ser eliminadas do *dataset*.

Por outro lado, verificamos a subida de variadas *features* da sua correlação com a *target*. No caso do *dataset Drinking*, a variável *serum_creatinine* subiu de 0.089 para 0.17, no distinto *dataset*, *Smoking*, a variável *gamma-glutamyl_transpeptidase* subiu de 0.25 para 0.35.

Análise Estatística Descritiva

Ainda assim, efetuamos gráficos de modo a perceber qual o comportamento associado a essas novas variáveis, bem como, visualizamos as diferenças antes do Pré-Processamento e depois para algumas variáveis.



(a) BoxPlot

(b) CountPlot

Figura 3: Visualização da Correção da Variável *gamma-glutamyl_transpeptidase*

Modelos de *Machine Learning*

De seguida, iremos apresentar todos os modelos aplicados pelo grupo para cada um dos *datasets*, bem como, os resultados obtidos em cada um e os melhores entre todos.

Antes de testarmos uma lista de modelos para identificar os melhores para o nosso problema, foi necessário eliminar todas as variáveis com baixos valores de correlação e informação mútua, tal como foi efetuado anteriormente.

Eliminação de Variáveis

Drinking As variáveis eliminadas foram:

```
Columns to Drop: ['waistline', 'sight_left', 'hear_right', 'diastolic_blood_pressure', 'triglyceride', '
    eye_sight_left', 'eye_sight_right', 'bmi', 'sight_right', 'serum_creatinine', 'weight']
```

Ficamos assim com as seguintes *features* finais:

```
Index(['sex', 'age', 'height', 'hemoglobin', 'gamma-glutamyl_transpeptidase',
       'smoking_state', 'drinker_or_not'],
      dtype='object')
```

Smoking As variáveis eliminadas foram:

```
Columns to Drop: ['systolic_blood_pressure', 'diastolic_blood_pressure', 'HDL_cholesterol', 'SGOT_ALT', '
    blood_pressure', 'numerical_BP', 'bmi', 'triglyceride', 'waistline']
```

Ficamos assim com as seguintes *features* finais:

```
Index(['sex', 'height', 'weight', 'hemoglobin', 'serum_creatinine',
       'gamma-glutamyl_transpeptidase', 'smoking_state', 'drinker_or_not'],
      dtype='object')
```

Melhores Modelos

Com o objetivo de identificar os melhores modelos, testamos onze modelos diferentes, como, por exemplo, *RandomForest*, *GradientBoosting*, *Support Vector Machine*, entre outros, utilizando *cross validation* com 10 *folds* e repetição 3 vezes.

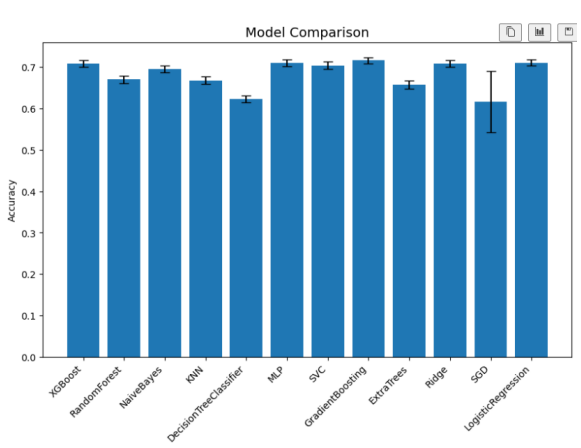
Os resultados obtidos foram:

Drinking

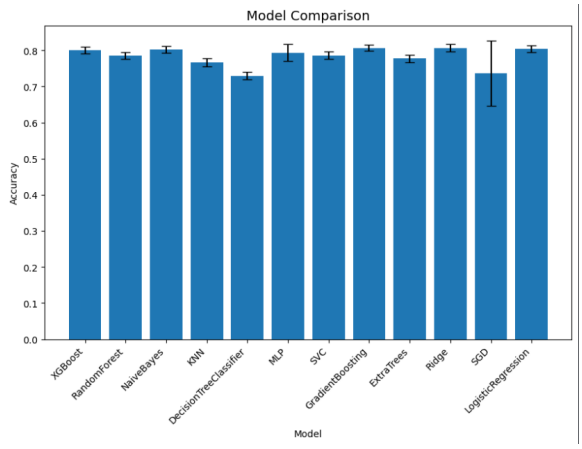
Model	CV	STD
XGBoost	0.70881	0.00800
RandomForest	0.66963	0.00855
NaiveBayes	0.69562	0.00822
KNN	0.66818	0.00830
DecisionTreeClassifier	0.62340	0.00765
MLP	0.70961	0.01222
SVC	0.70379	0.00902
GradientBoosting	0.71627	0.00774
ExtraTrees	0.65688	0.00992
Ridge	0.70798	0.00795
SGD	0.61465	0.07538
LogisticRegression	0.71070	0.00761

Smoking

Model	CV	STD
XGBoost	0.80056	0.01001
RandomForest	0.78571	0.00885
NaiveBayes	0.80266	0.01011
KNN	0.76668	0.01047
DecisionTreeClassifier	0.72860	0.00998
MLP	0.79649	0.02322
SVC	0.78641	0.01025
GradientBoosting	0.80698	0.00929
ExtraTrees	0.77796	0.01004
Ridge	0.80710	0.01074
SGD	0.76659	0.04242
LogisticRegression	0.80425	0.00886



(a) Model Comparison Drinking



(b) Model Comparison Smoking

Figura 4: Model Comparison

Avaliando os resultados produzidos para cada um dos modelos decidimos aplicar implementar os seguintes Modelos:

Drinking

- XGBoost
- RandomForest
- NaiveBayes
- KNN
- MLP
- SVC
- GradientBoosting
- ExtraTrees
- Ridge
- LogisticRegression

Smoking

- XGBoost
- RandomForest
- NaiveBayes
- KNN
- MLP
- SVC
- GradientBoosting
- ExtraTrees
- Ridge
- LogisticRegression

Resultados

Após aplicarmos hiperparâmetros a cada um dos Modelos identificados anteriormente iremos mostrar o melhor resultados obtido, bem como, os hiperparâmetros usado:

Drinking

Tabela 2: Model Performance and Parameters for Drinking

Smoking	Accuracy	Best Parameters	CV
ExtraTrees	77.79	{'criterion': 'entropy', 'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 500}	3
GradientBoosting	73.66	{'learning_rate': 0.01, 'max_depth': 5, 'min_samples_split': 2, 'n_estimators': 1000}	5
KNN	73.91	{'algorithm': 'ball_tree', 'leaf_size': 50, 'n_neighbors': 10, 'p': 1, 'weights': 'uniform'}	5
LogisticRegression	71.27	{'penalty': 'l2', 'solver': 'sag'}	10
NaiveBayes	69.79	{'priors': [0.2, 0.8], 'var_smoothing': 0.0001}	10
RandomForest	75.13	{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100}	5
Ridge	70.96	{'alpha': 0.1, 'solver': 'saga'}	10
SVC	71.39	{'C': 10, 'gamma': 'scale', 'kernel': 'linear'}	5
XGBoost	72.11	{'learning_rate': 0.01, 'max_depth': 5, 'min_child_weight': 4, 'n_estimators': 100}	5
MLP	71.05	{'activation': 'tanh', 'alpha': 0.0001, 'hidden_layer_sizes': (50, 50)}	5
LogisticRegression	71.27	{'penalty': 'l2', 'solver': 'sag'}	10

Smoking

Tabela 3: Model Performance and Parameters for Smoking

Smoking	Accuracy	Best Parameters	CV
ExtraTrees	81.40	{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 900}	5
GradientBoosting	81.82	{'learning_rate': 0.01, 'max_depth': 5, 'min_samples_split': 2, 'n_estimators': 500}	5
LogisticRegression	80.50	{'C': 0.1, 'penalty': 'l2'}	3
RandomForest	80.92	{'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 6, 'min_samples_split': 4, 'n_estimators': 400}	3
Ridge	80.71	{'alpha': 0.1, 'solver': 'saga'}	10
SVC	80.79	{'C': 10, 'gamma': 'scale', 'kernel': 'linear'}	5
XGBoost	81.21	{'learning_rate': 0.001, 'max_depth': 5, 'min_child_weight': 5, 'n_estimators': 1000}	5
Naive Bayes	80.66	{'activation': 'tanh', 'alpha': 0.01, 'hidden_layer_sizes': (100, 50, 100)}	3
MLP	80.66	{'activation': 'tanh', 'alpha': 0.01, 'hidden_layer_sizes': (100, 50, 100)}	3

Ensemble Learning

Finalmente, com o objetivo de tentar alcançar resultados ainda melhores, recorre-se à técnica de *Ensemble Learning*, mais especificamente, **Stacking**. Após termos obtido os melhores hiperparâmetros dos modelos individuais iremos aplicar a aprendizagem em conjunto. Isso assegura que cada modelo base tenha um desempenho sólido por si só. Após o refinamento dos modelos individuais, é possível combiná-los em um conjunto para tirar proveito da sua força coletiva.

Porém, após testarmos para o *dataset Smoking*, os modelos *ExtraTreesClassifier* e *XGBoost* como *estimators* e *final estimator GradientBoosting* e para o *dataset Drinking*, *RandomForest* e *XGBoost* como estimators e *ExtraTrees* como *final estimator*, não conseguimos melhorar os resultado já atingidos.

```
Cross-validated Accuracy of the Stacking Model: 80.68%
```

```
Cross-validated Accuracy of the Stacking Model: 71.11%
```

Em suma, podemos concluir que o melhor modelo para o conjunto de dados Smoking é o **GradientBoosting**, com 81.82% e para *Drinking*, **ExtraTrees** con 77.79% tal como é evidenciado pelas Tabelas 3 e 2.

DataSet Competição

Características dos Datasets de Energia e Meteorologia

Os conjuntos de dados utilizados para a competição foram disponibilizados pela Equipa Docente e estão centrados na interseção entre produção energética e meteorologia. O objetivo principal é desenvolver modelos capazes de antecipar como as condições meteorológicas podem alterar os níveis de produção energética.

Conjunto de Dados

Os conjuntos de dados estão segmentados por períodos de tempo em dois tipos:

1. **Dataset Energia**
2. **Dataset Metereologia**

Dataset Energia

O conjunto de dados de energia fornece uma visão abrangente do consumo e produção energética. Contém informações sobre a data e hora em que as medições foram realizadas, bem como, métricas cruciais relacionadas ao consumo de energia. Estas métricas incluem o consumo durante períodos normais, o consumo durante o horário económico, a quantidade de energia proveniente de fontes de autoconsumo (como painéis solares) e a quantidade de energia injetada na rede elétrica. Esses dados são fundamentais para analisar os padrões de consumo e avaliar o impacto das fontes de energia renovável no sistema elétrico.

Number	Name	Description	DType
0	Data	Date(YY-MM-DD-HH)	object
1	Hora	HH	int
2	Normal (KWH)	Consumo num período Normal	float64
3	Horario Economico (kWh)	Consumo no horário económico	float64
4	Autoconsumo (kWh)	Quantidade de Energia Gasta que é provenienete do painel Solar	float64
5	Injecao na rede (kWh)	Quantidade de energia injetada na rede	object

Tabela 4: Data Description Energia

Dataset Meteorologia

O conjunto de dados de meteorologia fornece informações abrangentes sobre as condições atmosféricas. Inclui dados temporais em formato *timestamp*, detalhes sobre a cidade em análise, e diversas métricas meteorológicas, tais como, temperatura atual, sensação térmica, temperatura mínima e máxima, pressão atmosférica, níveis de humidade, velocidade do vento, precipitação média, nebulosidade, e uma descrição do estado atmosférico. Esses dados são cruciais para compreender e analisar as variações climáticas ao longo do tempo.

Number	Name	Description	DType
0	dt	Timestamp	int64
1	dt_iso	Hora	object
2	city_name	Nome da Cidade	string
3	temp	Temperatura(Celsius)	float64
4	feels_like	Sensação Térmica	float64
5	temp_min	Temperatura Mínima	float64
6	temp_max	Temperatura Máxima	float64
7	pressure	Pressão atmosférica	int64
7	sea_level	Pressão atmosférica ao nivel do mar	float64
9	grnd_level	Local Atmospheric Pressure	float64
10	humidity	Humidade	int64
11	wind_speed	Velocidade do Vento	float64
12	rain1h	Precipitação Média	float64
13	clouds_all	Nível de nebulosidade	int64
14	weather_description	Descrição do estado atmosférico	object

Tabela 5: Data Description Metorologia

União dos Datasets de Energia e Meteorologia

Processo de Tratamento dos Dados de Treino

Após uma análise exaustiva dos diversos *datasets* fornecidos, verificamos a necessidade de integrar as informações dos *datasets* de treino de meteorologia e energia através da coluna de data. Para possibilitar essa integração, foi necessário efeturar modificações nos *datasets*. No caso dos *datasets* de energia, optou-se por consolidar as informações das colunas de hora e data, criando assim um identificador único. Este identificador único foi crucial para facilitar o processo de *merge*, possibilitando a união eficiente das linhas entre os dois conjuntos de dados.

No *dataset* de meteorologia, foi adotado um procedimento semelhante, envolvendo a aplicação de expressões regulares (*regex*) para remover informações excessivas da coluna de data e obter um formato uniforme do tipo `%Y-%m-%d %H:%M` .Ao seguir um padrão para a representação da data foi essencial para garantir a consistência durante a fusão dos *datasets*.

Durante esse processo, observamos que os *datasets* do ano de 2021 apresentavam discrepâncias em termos de quantidade de linhas. Por um lado, o *dataset* de energia continha 2256 linhas, abrangendo o período de 29 de setembro de 2021 a 31 de dezembro de 2021, enquanto o *dataset* de meteorologia possuía 2928 linhas, cobrindo o intervalo de 1 de setembro de 2021 a 31 de dezembro de 2021. Notavelmente, o *dataset* de meteorologia incluía informações referentes a quase um mês adicional.

A fim de mitigar essa discrepância, foi antecipado que o resultado do *merge* consistiria apenas nas linhas cujas datas estivessem contidas na interseção dos dois intervalos, ou seja, entre 29 de setembro de 2021 e 31 de dezembro de 2021.

Posteriormente a essa etapa, o método *merge* foi aplicado com sucesso, permitindo a fusão dos dois conjuntos de dados com base no parâmetro de data e hora, que atuou como um identificador único.

Processo de Tratamento dos Dados de Teste

Através de uma análise detalhada dos dados de teste de meteorologia e energia, identificamos algumas inconsistências significativas. O *dataset* de energia continha informações relativas ao período entre 01-01-2023 e 04-04-2023, enquanto o *dataset* de meteorologia correspondente fornecia dados meteorológicos apenas para o intervalo entre 01-01-2023 e 14-03-2023. Reconhecendo a importância e o peso considerável dos dados de energia, tornou-se evidente que descartá-los não seria uma abordagem viável. Face a essa discrepância temporal, decidimos procurar soluções para obter as informações de energia necessárias para o período ausente. Após pesquisas em várias plataformas, identificamos a plataforma *Weather Forecast* que

apresentava dados compatíveis com as datas e as colunas que precisávamos. Essa abordagem permitiu-nos fortalecer a integridade dos dados de teste, garantindo a plenitude das informações e preservando a robustez do conjunto de dados como um todo.

Análise da Qualidade dos Dados

Após a consolidação de todos os *datasets* num único ficheiro, avançamos para a verificação da qualidade dos dados, focando nossa atenção em *Missing Values* e *Duplicate Values*. Baseando-nos no facto de as colunas *sea_level* e *grnd_level* possuírem tantos *Missing Values* quanto o número total de linhas total, tomamos a decisão de remover ambas as colunas, visto que não continham qualquer informação relevante.

Verificamos, também, que os valores da coluna *injection* cujo conteúdo era **None** estavam a ser considerados como *Missing Values*. Para tratar essa ambiguidade, utilizamos a *flag* $na\_filter = False$, evitando assim que fossem contabilizados como valores em falta.

Quanto à coluna **rain_1h**, optamos por substituir os valores nulos por 0, visto que faz sentido no contexto do problema quando não há precipitação, o valor associado é 0.

Decidimos renomear algumas colunas cujo nome era bastante extenso e pouco sugestivo para os nomes mencionados de seguida:

$$
\begin{aligned}
dt &\rightarrow \text{timestamp} \\
dt\_iso &\rightarrow \text{date} \\
city\_name &\rightarrow \text{city} \\
Data &\rightarrow \text{date} \\
Hora &\rightarrow \text{hour} \\
Normal(\text{kWh}) &\rightarrow \text{normal\_consume} \\
HorarioEconomico(\text{kWh}) &\rightarrow \text{consumption\_in\_period} \\
Autoconsumo(\text{kWh}) &\rightarrow \text{autoconsume} \\
Injecao \text{ na rede(kWh)} &\rightarrow \text{injection}
\end{aligned}
$$

Após isso realizamos *encoding* das colunas *injection* e *weather_description* com o objetivo de representar categorias qualitativas por valores numéricos.

None	0
Low	1
Medium	2
High	3
Very High	4

sky is clear	0
few clouds	1
broken clouds	2
overcast clouds	3
scattered clouds	4
light rain	5
moderate rain	6
heavy intensity rain	7

Análise Exploratória de Dados

Na fase da análise exploratória de dados, concentramo-nos na observação e análise geral dos dados usando métodos como *describe*, entre outros. O principal objetivo foi compreender a consistência e estrutura dos dados, destacando características estatísticas fundamentais e observando os primeiros e últimos registros. Essa abordagem inicial proporcionou *insights* iniciais sobre a natureza dos dados, estabelecendo uma base para análises mais aprofundadas e tomada de decisões subsequentes.

A **variável-alvo** em questão era a **Injeção de Energia na Rede**. O nosso objetivo principal era desenvolver modelos capazes de prever como as condições meteorológicas e o consumo de energia em uma residência poderiam influenciar essa quantidade específica de energia injetada na rede elétrica.

Análise antes do Pré-Processamento

Análise de Correlação

Antes de avançarmos para o Pré-Processamento de Dados, realizámos o cálculo da matriz de correlação entre todas as variáveis do conjunto de dados. Este passo proporcionou uma compreensão das relações existentes entre as diferentes variáveis. Adicionalmente, calculámos a Informação Mútua para avaliar a dependência estatística entre as variáveis.

Análise Estatística Descritiva

Uma das fases críticas e fundamentais no estudo do problema consiste na análise estatística dos dados, onde é possível compreender a dispersão dos dados, identificar relações entre variáveis e antecipar possíveis impactos na fase subsequente de modelação. Optamos por iniciar esta etapa com a inclusão de gráficos simples, como *boxplots*, para todas as features, com o propósito de examinar a distribuição dos dados e identificar a presença de *outliers*. Ao realizar essa análise, observamos a existência de alguns *outliers* em

determinadas colunas, os quais serão abordados de maneira adequada durante a fase de pré-processamento dos dados.

Além dos gráficos exploratórios básicos, implementamos visualizações mais complexas com o objetivo de compreender relações entre *features* dado o problema de **Time Series**. Essas práticas visam não apenas à compreensão profunda dos dados, mas também à criação de novas variáveis que possam potencialmente melhorar o desempenho dos modelos.

Através da análise de gráficos conseguimos obter informações como:

- Em cerca de 11 mil registos, cerca de 71% dos mesmos representam valores nulos para a injeção;
- Os valores de Consumo Normal, Consumo no período económico e Autoconsumo variam entre 0 e 3 KWH na sua grande parte;
- O nível de nebulosidade varia entre 20 e 100 de uma forma igualmente distribuída;
- A distribuição de valores para descrição do estado meteorológico apresenta uma ligeira tendência para ser "Céu Limpo"(27%);

Construímos também alguns gráficos elaborados que nos permitissem observar a variação e a média da injeção no período de um mês, num dia, num feriado, num fim de semana entre outro.

Observamos uma tendência notável que sugere um aumento nos valores de injeção durante os meses de verão, em contraste com uma diminuição correspondente no consumo durante esse período. Ao que já era espectável por parte do grupo.

Pré-Processamento dos Dados

Após analisarmos detalhadamente o problema e realizar algumas pesquisas sobre meteorologia e energia, chegamos à conclusão de que seria vantajosa a inclusão de novas colunas (**Feature Engineering**) que contivessem informações relevantes para o contexto do problema. Essas novas variáveis poderiam ser úteis na previsão da variável-alvo durante a fase de modelação.

Decidimos incorporar características relacionadas ao contexto do problema, permitindo-nos obter informações adicionais. Essas características incluem, por exemplo, indicadores de períodos de férias, fins de semana, períodos de sol, entre outras. A adição dessas variáveis tem como objetivo capturar padrões e

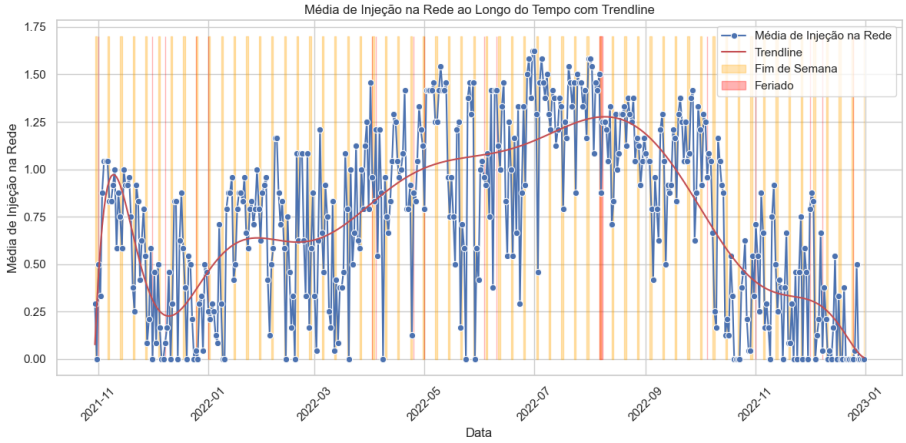


Figura 5: *Variação da injeção de Energia ao longo do Tempo*

correlações que podem não ser evidentes nas variáveis originais, mas que têm o potencial de influenciar as variações na injeção de energia na rede.

Ao considerar aspetos como períodos de férias ou fins de semana, conseguimos capturar comportamentos específicos associados a esses eventos, que afetam diretamente a injeção. Da mesma forma, a inclusão de informações sobre períodos de sol foi crucial pois a injeção em diferentes períodos do ano depende do sol.

Em resumo, o *dataset* encontrava-se, neste ponto, com as seguintes *features* novas:

Number	Name	Description	DType
0	Consumo_Total	Representa a soma total dos consumos por hora	int64
1	Taxa de Autoconsumo	Percentagem da energia total consumida que é proveniente dos painéis solares	int64
2	is_Weekend	Indica se é ou não fim de semana	int64
3	is_Feriado	Indica se é ou não feriado	int64
4	injection_Tentativa	Variável que tenta prever a injeção energia na rede	int64
5	vacaciones	Indica que está no período de férias	int64
6	estação do ano	Indica qual é a estação do ano.	int64
7	hora_partida	Indica se é ou não período de sol.	int64
8	dew	Indica a existência ou não de Orvalho	int64
9	injection_tentativa	Tentativa de verificar a injeção	int64
10	is_sunny	Indica se naquela hora estava ou não sol	int64

Tabela 6: *Feature Engeneering*

Análise após o Pré-Processamento

Análise de Correlação

Após o Pré-Processamento, decidimos criar uma matriz de correlação para avaliar as relações entre as variáveis numéricas (incluindo as novas). Rapidamente identificamos a presença de um quadrado vermelho, indicativo de correlações muito elevadas, e reconhecemos colunas que poderiam ser removidas.

Posteriormente, procedemos à eliminação de colunas com correlação inferior a 0.24 e maior que -0.24 e informação mútua inferior a 0.1. Esta decisão teve como objetivo eliminar variáveis que apresentassem correlações mais fracas ou informações redundantes, simplificando o conjunto de dados.

A escolha desses valores de correlação e informação mútua foi estratégica, visando manter no conjunto de dados apenas as variáveis que contribuem de forma mais significativa para a previsão da **target**.

Ficamos com as seguintes colunas :

```
'normal_consume', 'consumption_in_period', 'autoconsume', 'injection',
'temp_max', 'humidity', 'hour', 'taxa_autoconsumo', 'hora_partida',
'injection_tentativa', 'is_sunny'
```

Nota: É importante salientar que os procedimentos realizados no conjunto de dados de treino foram igualmente aplicados ao conjunto de dados de teste. Esta abordagem garantiu a coerência entre os conjuntos de treino e teste, possibilitando uma avaliação precisa do desempenho dos modelos nos dados de teste.

Análise Estatísitica Descritiva

As únicas mudanças em relação à Análise Estatística Descritiva, antes do Pré-Processamento, basearam-se na criação de novos gráficos mais complexos e na tentativa de compreender o comportamento de todas as novas *features* criadas anteriormente.

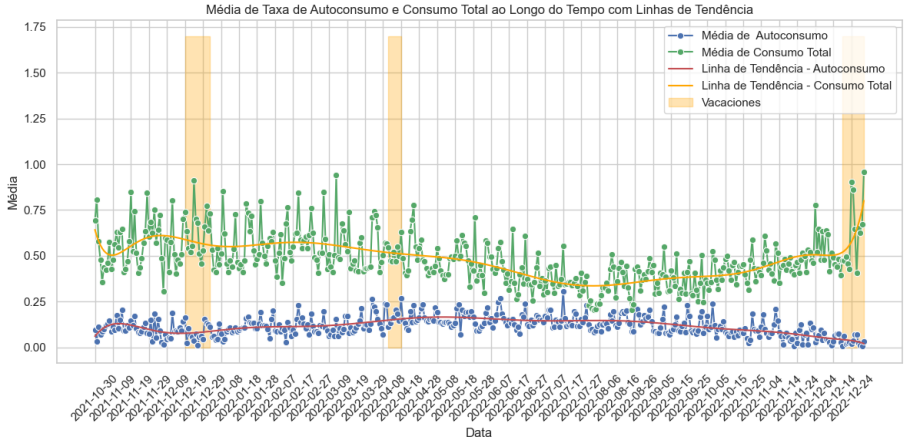


Figura 6: Média de Taxa de Autoconsumo e Consumo Total ao Longo do Tempo com Linhas de Tendência

Modelação

Após a exclusão das colunas com correlações e informações mútuas mais baixas, avançamos para a fase de modelação. O objetivo principal era prever a coluna *injection* nos dados de teste usando os dados de treino. De seguida iremos apresentar os resultados obtidos no **Kaggle** para os diversos modelos que implementamos. Vale ressaltar que a estratégia adotada envolveu a realização de testes locais para ajuste dos hiperparâmetros, visando obter o melhor desempenho possível antes de submeter as previsões finais no **Kaggle**.

Com o objetivo de identificar os melhores modelos, testamos onze modelos diferentes, como, por exemplo, *RandomForest*, *GradientBoosting*, *Support Vector Machine*, entre outros, utilizando *cross validation* com 10 *folds* e repetição 3 vezes. Daí resultaram os quatro modelos aos quais dedicamos todo o nosso esforço:

1. **RandomForest**
2. **GradientBoosting**
3. **ExtraTrees**
4. **XGBoost**

Os valores obtidos resultaram da aplicação destes modelos aos dados de **treino**, com o objetivo de obter algum feedback sobre o desempenho antes de os submeter ao Kaggle. Esta abordagem permitiu avaliar a eficácia dos modelos em dados já conhecidos, possibilitando ajustes e refinamentos antes da submissão final para avaliação externa.

Modelo	Precisão CV	Desvio Padrão
XGBoost	0.87625	0.00775
RandomForest	0.87973	0.00614
GradientBoosting	0.88167	0.00611
ExtraTrees	0.87749	0.00640
NaiveBayes	0.69315	0.01134
KNN	0.79541	0.00892
DecisionTreeClassifier	0.84568	0.00749
MLP	0.85603	0.00814
SVC	0.73787	0.00396
Ridge	0.81349	0.00612
SGD	0.77201	0.03849
LogisticRegression	0.71070	0.00761

Os melhores resultados obtidos na Competição Pública do **Kaggle** estão presentes na figura abaixo:

Modelo	Accuracy	Hyperparameters
XGBoost	0.87573	learning_rate = [0.09], n_estimators = [150], max_depth = [5], objective =['multi:softprob']
RandomForest	0.87278	n_estimators=1000; max_depth=60
GradientBoosting	0.87573	n_estimators=700; max_depth=20; learningRate=0.01
ExtraTrees	0.86242	n_estimators: 400 max_depth: 50, min_samples_split: 7, criterion:entropy, min_samples_leaf: [1],
GradientBoosting	0.86982	n_estimators=200; max_depth=5; learningRate=0.01; min_samples_split=15
RandomForest	0.86686	n_estimators=1000; max_depth=25; criterion=Entropy

Tabela 7: Model Performance in Kaggle

Ensemble Learning

Na parte final do nosso trabalho, incorporamos técnicas de *Ensemble Learning*, como o *Stacking*. Esta técnica foi utilizada em todos modelos para efetuar previsões. O *stacking* consiste em combinar as previsões individuais destes modelos através de um meta-modelo, permitindo criar uma abordagem mais robusta e poderosa.

O processo de *stacking* desenrola-se em duas fases. Na primeira fase, cada modelo é treinado de forma independente com os dados de treino, gerando previsões para o conjunto de dados de validação. Na segunda fase, essas previsões do conjunto de validação são utilizadas como entrada para um meta-modelo, o qual é treinado para prever a variável alvo final. Este meta-modelo é capaz de combinar as forças individuais de cada modelo base, resultando numa previsão mais precisa e eficaz.

Modelo	**Accuracy**
XGBoost	0.87573
RandomForest	0.86982
GradientBoosting	0.86686
ExtraTrees	0.85946

Tabela 8: Stacking Model Performance in Kaggle

Em suma, podemos concluir que o melhor modelo para o conjunto de dados de maneira a prever a *injection* é o **GradientBoosting**, com 87.573%

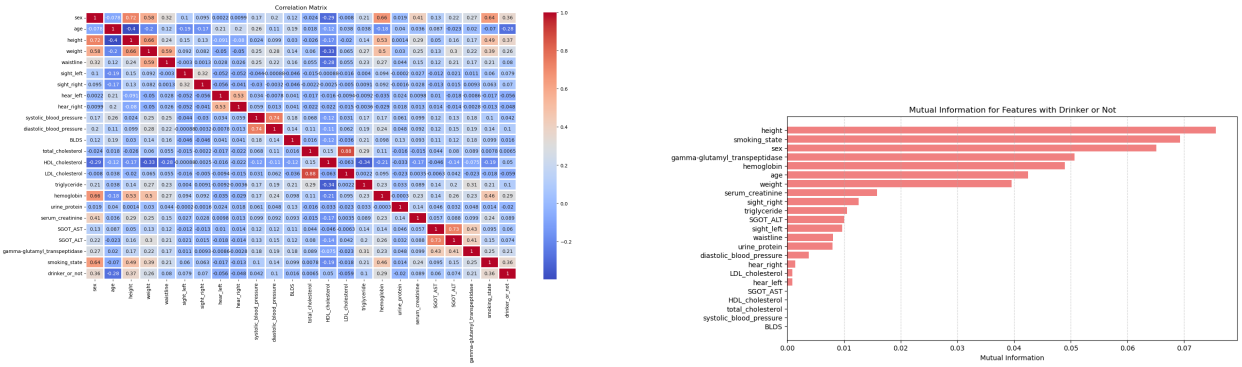
Conclusão

Em suma, com este projeto conseguimos aplicar todos os conceitos abordados ao longo do semestre na Unidade Curricular, desde a Análise de Dados ao *Ensemble Learning* passando pela Correlação, Pré-Processamento e os Modelos.

Anexos Dataset Grupo

Análise de Correlação antes do Pré-Processamento

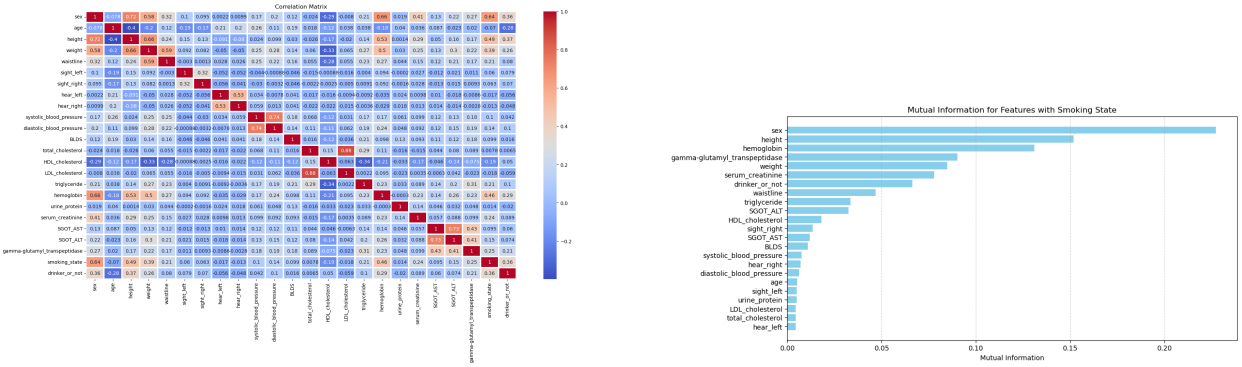
Drinking



(a) Correlação (b) Mutual Information

Figura 7: Correlação e Mutual Information Drinking

Smoking



(a) Correlação (b) Mutual Information

Figura 8: Correlação e Mutual Information Smoking

Análise Estatıstica Descritiva antes do Pré-Processamento

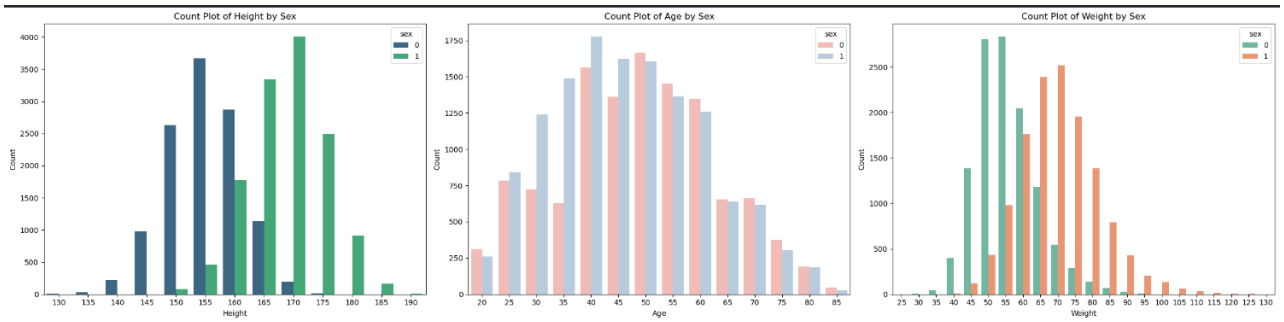


Figura 9: *CountPlot*

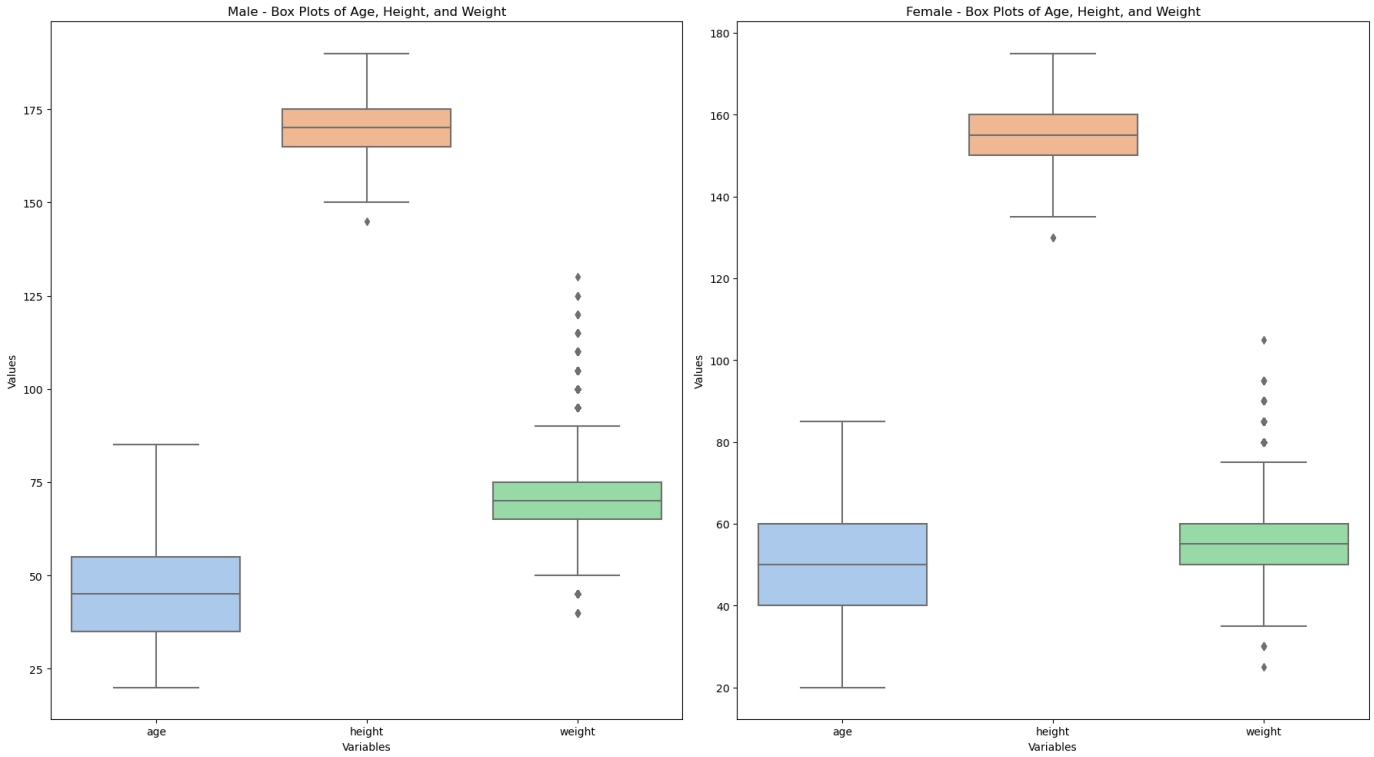
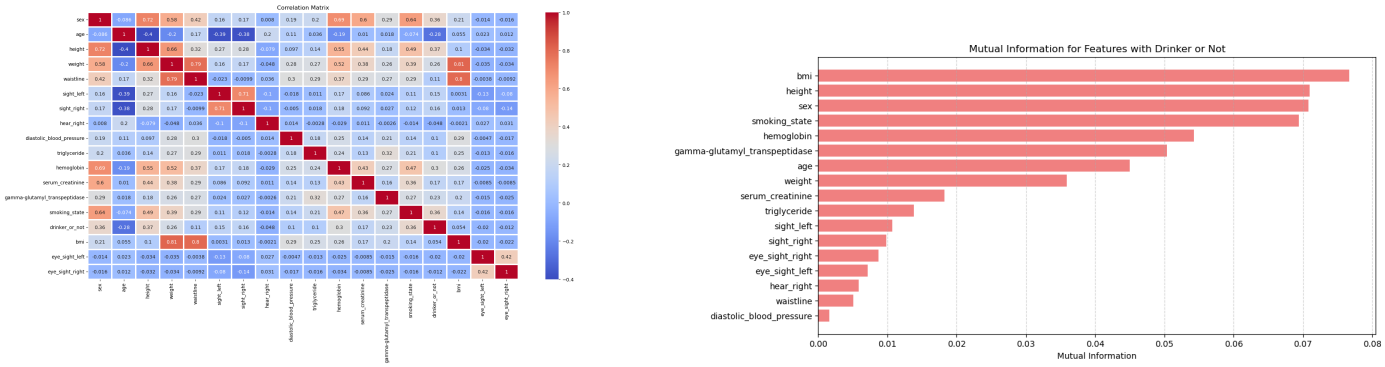


Figura 10: *BoxPlot*

Análise de Correlação depois do Pré-Processamento

Drinking

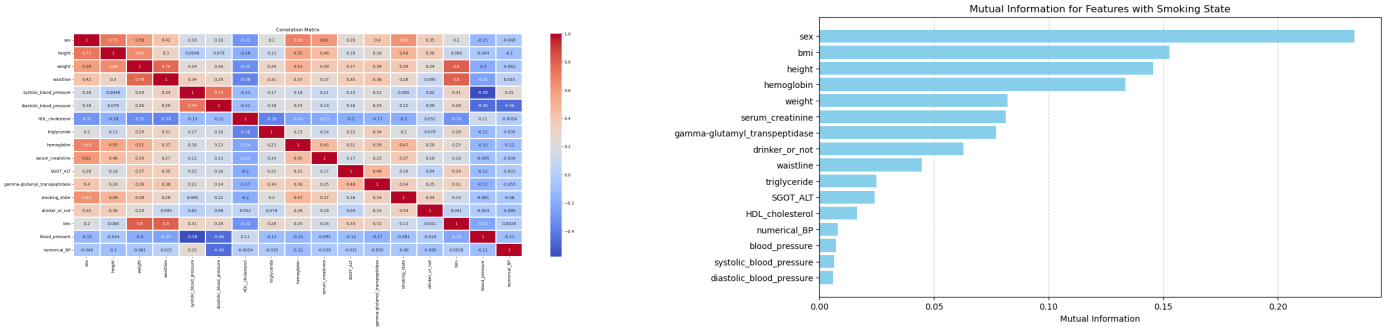


(a) Correlação

(b) Mutual Information

Figura 11: Correlação e Mutual Information Drinking

Smoking

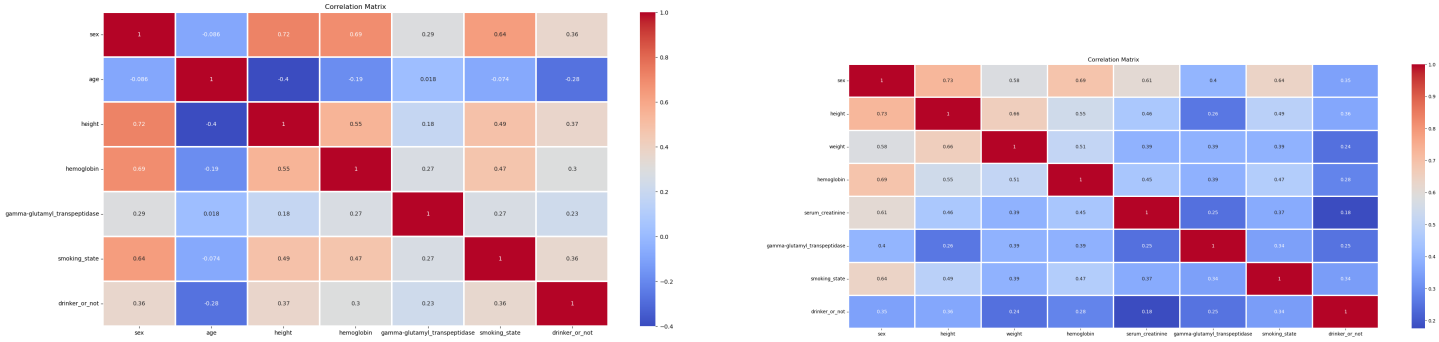


(a) Correlação

(b) Mutual Information

Figura 12: Correlação e Mutual Information Smoking

Análise de Correlação Final



(a) Correlação Final Drinking

(b) Correlação Final Smoking

Figura 13: Correlação Final

Anexos Dataset Competição

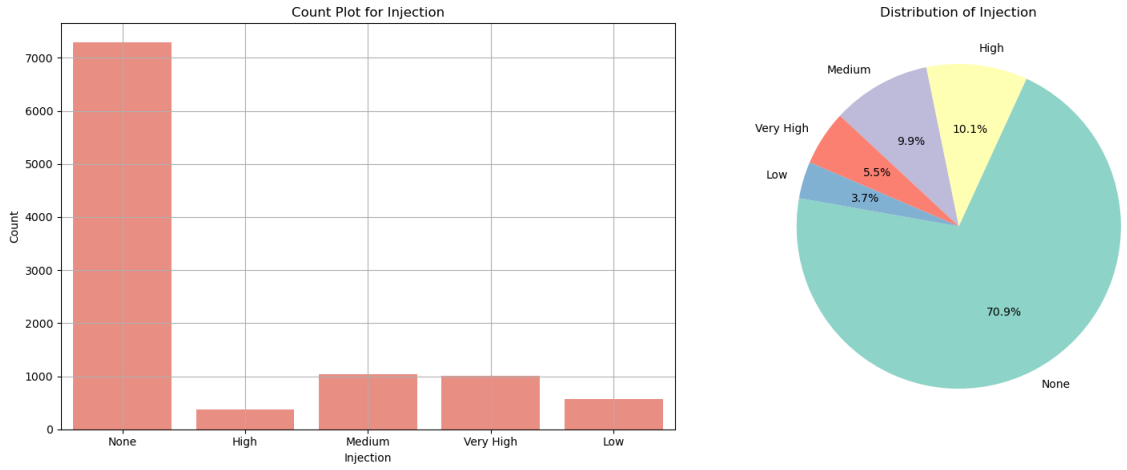


Figura 14: *Distribuição dos Valores de Injection*

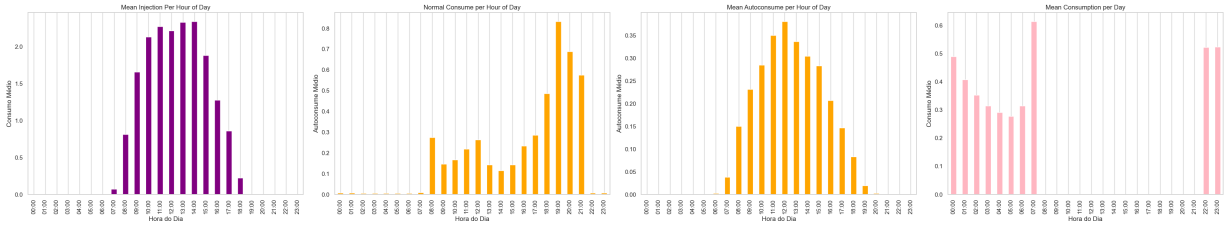
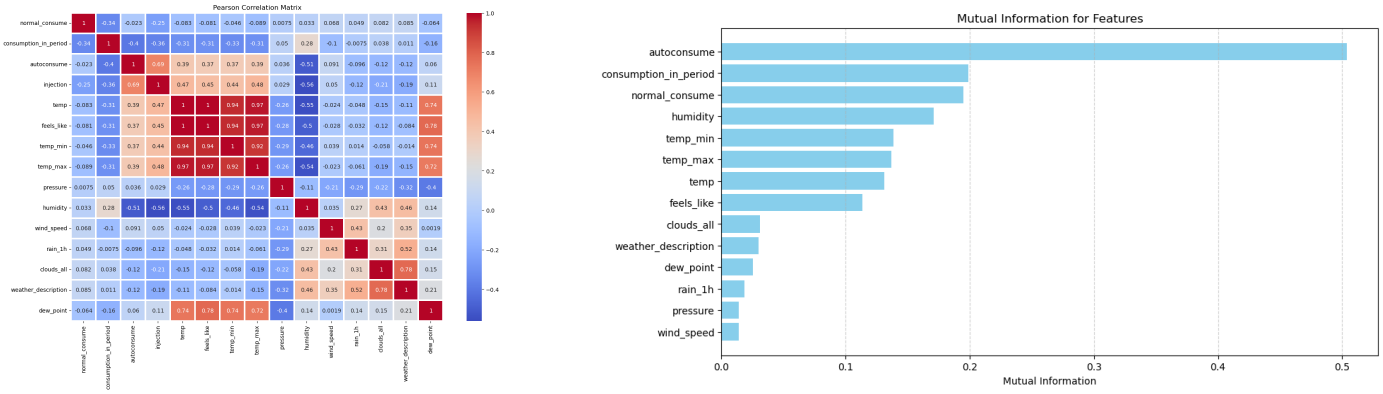


Figura 15: *BarCharts Energia/Hora do dia*

Análise da Correlação Inicial



(a) Correlação

(b) Mutual Information

Figura 16: Correlação e Mutual Information antes do PP