

# DAA

# Trabalho Prático

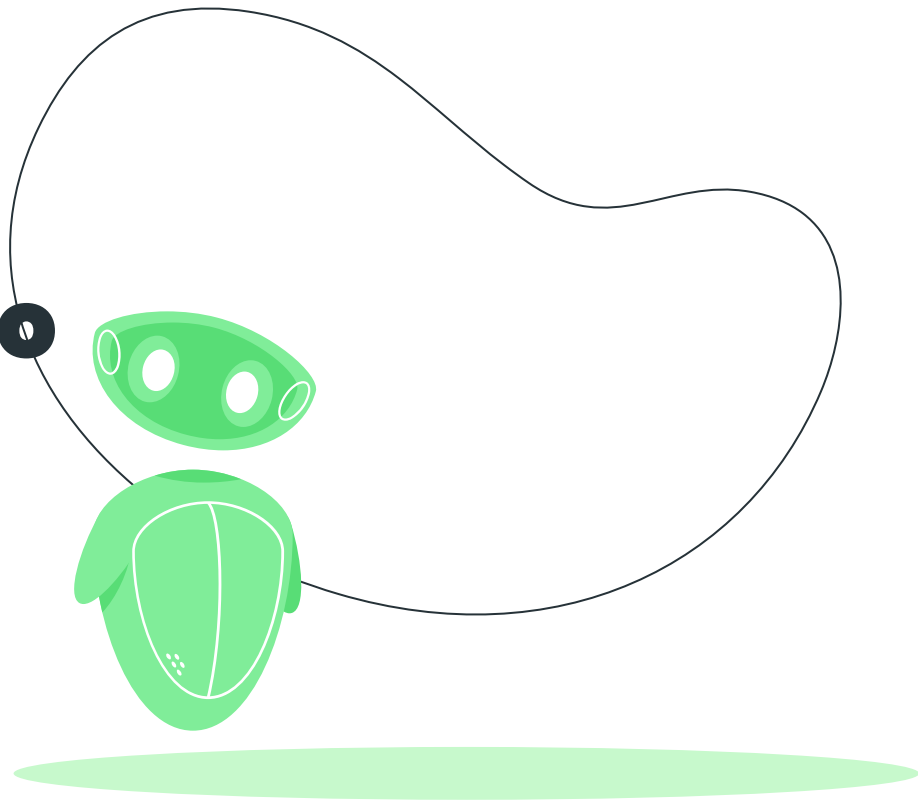
## Grupo 20

António Santos pg47031

Jorge Vieira pg47349

Manuel Moreira pg47439

Sara Dias pg47667



# Índice da Apresentação

1

## Dataset 1

- Exploração do dataset e tratamento de dados
- Desenvolvimento do modelo
- Análise dos resultados

2

## Dataset 2

- Exploração do dataset e tratamento de dados
- Desenvolvimento do modelo
- Análise dos resultados

# Dataset 1

A modelação do fluxo de tráfego  
rodoviário

# Exploração do dataset

- O dataset de treino fornecido pela equipa docente é composto por 14 **colunas** e 6812 **linhas**

Coluna	Categoria
city_name	Categórico
record_date	Numérico
<b>average_speed_dif</b>	Categórico
average_free_flow_speed	Numérico
average_time_diff	Numérico
average_free_flowtime	Numérico
luminosity	Categórico


Coluna	Categoria
average_temperature	Categórico
average_atmosp_pressure	Numérico
average_humidity	Numérico
average_wind_speed	Numérico
average_cloudiness	Categórico
average_precipitation	Numérico
average_rain	Categórico

Após o estudo do dataset o grupo chegou às seguintes conclusões:

# Tratamento de dados

- O dataset continha dados irrelevantes para o desenvolvimento do modelo.

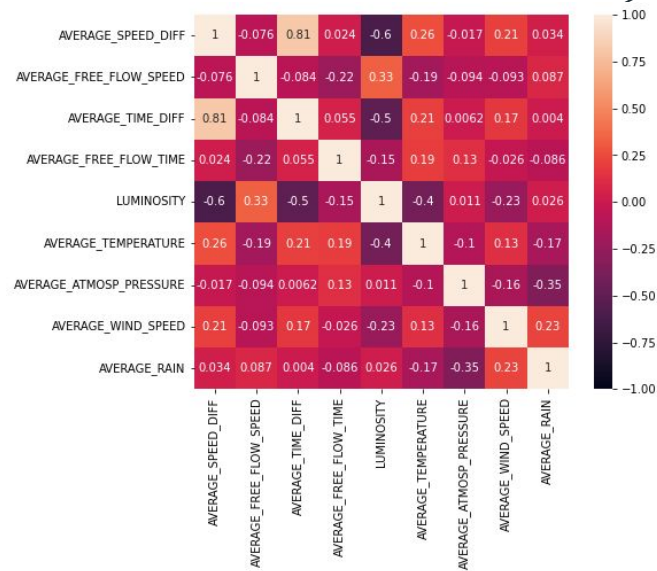
Após uma inspeção inicial o grupo concluiu que colunas **city\_name**, **average\_precipitation** e **record\_date** não apresentam nenhuma informação relevante para o desenvolvimento do modelo, logo estas colunas foram removidas.



# Tratamento de dados

- O dataset continha dados irrelevantes para o desenvolvimento do modelo.

Com o desenvolvimento e estudo de uma matriz de correlação foi possível concluir que a remoção das colunas **average\_humidity** e **average\_cloudiness** não afetariam o modelo.

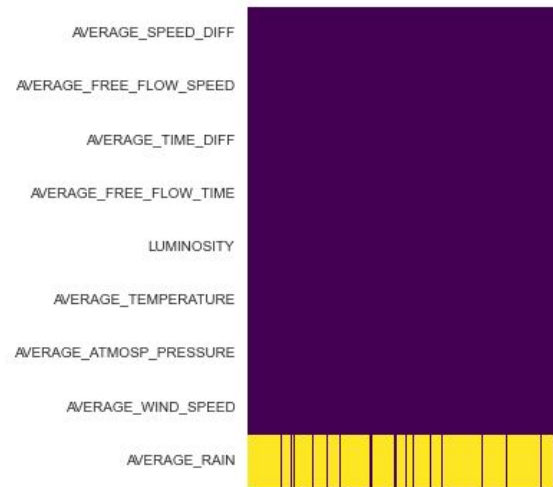


# Tratamento de dados

- **O dataset continha elementos com valores nulos.**

De forma a preencher todos os valores nulos presentes no dataset o grupo testou várias alternativas, tais como a remoção das colunas ou linhas que continham elementos nulos, o preenchimento dos valores com a média total...

**O grupo acabou por preencher todos os elementos nulos com o valor 0 visto que foi a solução que apresentou melhores resultados.**



# Tratamento de dados

- Os dados discretos presentes no dataset estavam representados como Strings.

Como os algoritmos de aprendizagem não são capazes de aprender com dados não numéricos foi necessário atribuir um valor numérico a cada elemento discreto.

O grupo adotou o uso de **Label Encoding** para substituir todos os valores não numéricos, atribuindo um valor mais alto quanto mais o valor contribuísse para o tráfego.

*"luminosity = {'DARK': 2, 'LOW\_LIGHT': 1, 'LIGHT': 0}"*



# Tratamento de dados

- **Os valores numéricos não estavam equilibrados.**

Devido à alta discrepância entre certas colunas (por exemplo uma coluna ter valores entre 0 e 1000 e outra ter entre 0 e 5), o grupo optou por acrescentar o **StandardScaler** ao pipeline de aprendizagem.

Isto resultou numa otimização do modelo, tal como na sua precisão, quer como no tempo de execução.

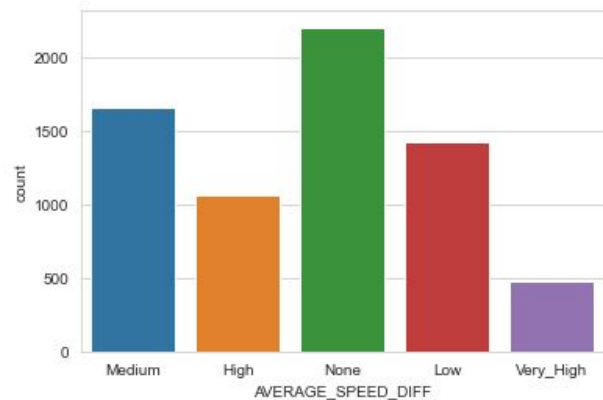
# Tratamento de dados

- **O dataset não está balanceado.**

Como dá para ver no gráfico à direita o dataset fornecido pela equipa docente não se encontra balanceado.

De forma a resolver este problema o grupo decidiu não alterar o dataset mas a acrescentar um parâmetro nos algoritmos de aprendizagem que faz com que o algoritmo dê mais ênfase às linhas que apresentam um resultado menos comum.

```
class_weight='balanced'
```



# Desenvolvimento do Modelo

Como temos acesso aos valores que pretendemos prever então podemos usar **algoritmos com supervisão**.

A coluna que pretendemos prever poderá ter 5 valores possíveis, logo podemos concluir que isto se trata de um problema de **classificação**.

Como ponto inicial o grupo decidiu comparar o desempenho de 4 algoritmos de classificação.

Estes são:

- Árvore de decisão
- Support Vector Machine (SVM)
- Regressão Logística
- Random Forest Classifier

# Desenvolvimento do Modelo

Após a execução inicial chegamos aos seguintes resultados:

Nome	Precisão (%)	Tempo de execução (ms)
Árvore de Decisão	70.5%	28
Support Vector Machine	56.85%	2605
Random Forest Classifier	77.7%	1173
Regressão Logística	76.57%	3764

Apesar da Árvore de Decisão apresentar um melhor tempo de execução o grupo decidiu tomar a precisão como prioridade e por isso será o algoritmo **Random Forest Classifier** que será utilizado.

# Desenvolvimento do Modelo

Após acrescentarmos o scaler ao pipeline temos os seguintes resultados:

Nome	Precisão (%)	Tempo de execução (ms)
StandardScaler+Random Forest Classifier	78.3%	889

# Desenvolvimento do Modelo

Com o dataset e o pipeline preparados basta afinar os parâmetros para atingirmos um modelo otimizado.

Para isso foi utilizada a função **gridSearchCV** para testar todas as combinações possíveis de um conjunto de parâmetros definido pelo grupo devolvendo a combinação que resulta numa maior precisão.

Este é um processo que demorou **mais de 10 minutos** a executar por isso o grupo decidiu copiar os parâmetros e remover o processo do ficheiro final.

Os parâmetros afinados são os seguintes:

- 'criterion': 'gini'
- 'max\_depth': 83
- 'max\_features': 'log2'
- 'n\_estimators': 200

# Análise crítica dos resultados

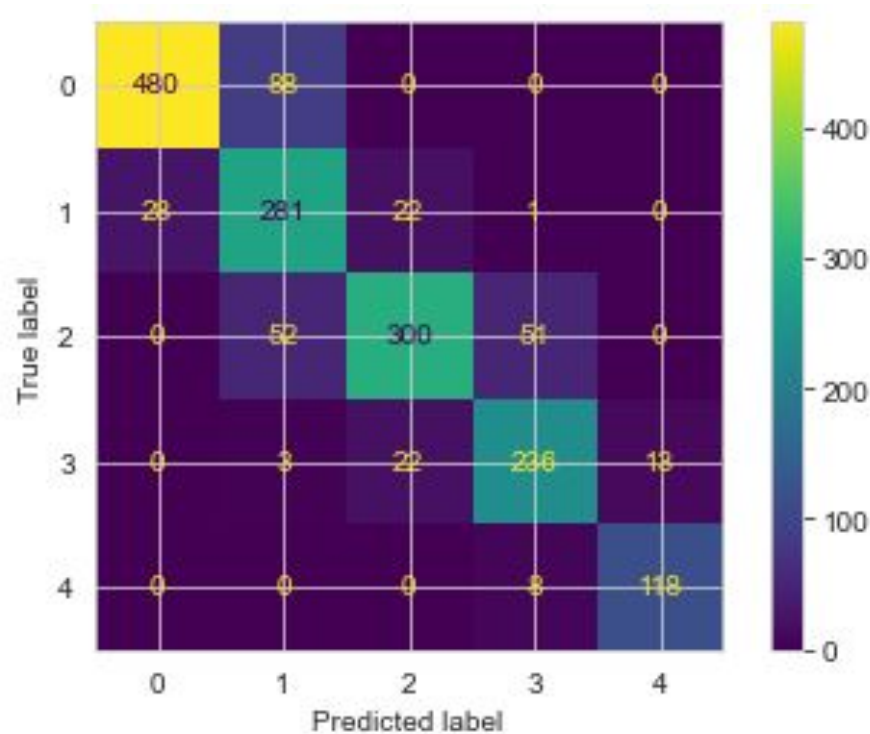
- Foi possível concluir que os atributos que mais influenciam os resultados são **average\_time\_diff** e **luminosity**, o que também é verdade pela nossa experiência do mundo real.
- A precisão do modelo atingiu os 78.8%.
- Como o modelo *Random Forest Classifier* não tem uma precisão fixa, o valor atingido resultou do melhor valor entre 10 execuções do modelo.

Nome	Precisão (%)	Tempo de execução (ms)
StandardScaler+RFC (Afinado)	78.8%	1402

Name	Submitted	Wait time	Execution time	Score
predictions1.csv	19 hours ago	1 seconds	0 seconds	0.78888

Complete

# Análise crítica dos resultados





# Dataset 2

*Performance de alunos nos exames*

# Exploração do dataset

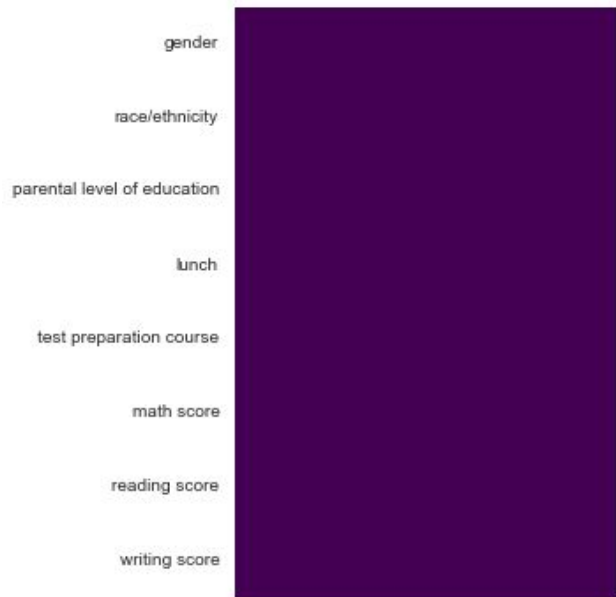
- O dataset de treino fornecido pela equipa docente é composto por 8 **colunas** e 1000 **linhas**

Coluna	Categoria
gender	Categórico
race/ethnicity	Categórico
parental level of education	Categórico
lunch	Categórico
test preparation course	Categórico
math score	Numérico
reading score	Numérico
<b>writing score</b>	<b>Numérico</b>

# Tratamento de dados

- **O dataset não contém elementos com valores nulos.**

Tal como se pode ver na imagem à direita, o *dataset* não apresenta valores nulos, logo não é necessário qualquer tipo de tratamento.



# Tratamento de dados

Após o estudo do dataset o grupo chegou às seguintes conclusões:

- **Os dados discretos presentes no dataset estavam representados como Strings.**

Como os algoritmos de aprendizagem não são capazes de aprender com dados não numéricos foi necessário atribuir um valor numérico a cada elemento discreto.

O grupo adotou o uso de **Label Encoding** para substituir todos os valores não numéricos.

*race = {'group A': 0, 'group B': 1, 'group C': 2, 'group D': 3, 'group E': 4}*

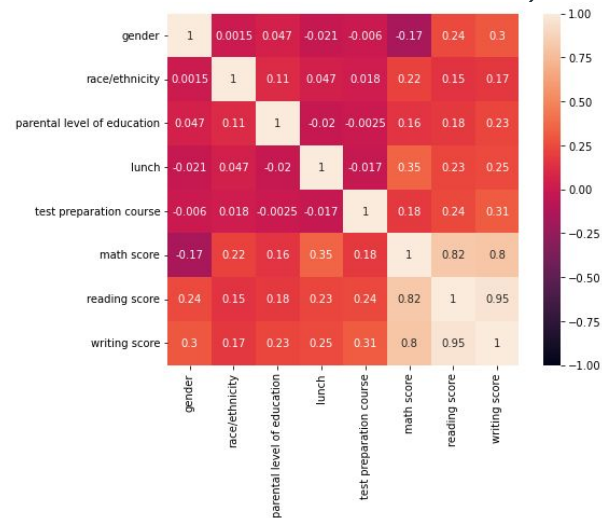
# Tratamento de dados

- Matriz de correlação

Com o desenvolvimento e estudo de uma matriz de correlação, foi possível concluir que o modelo não teria um bom desempenho com a previsão das notas de todos os exames.

Assim, decidimos escolher o *writing score* como o atributo a prever.

Também podemos verificar que as notas dos exames apresentam uma maior influência do que os restantes atributos.



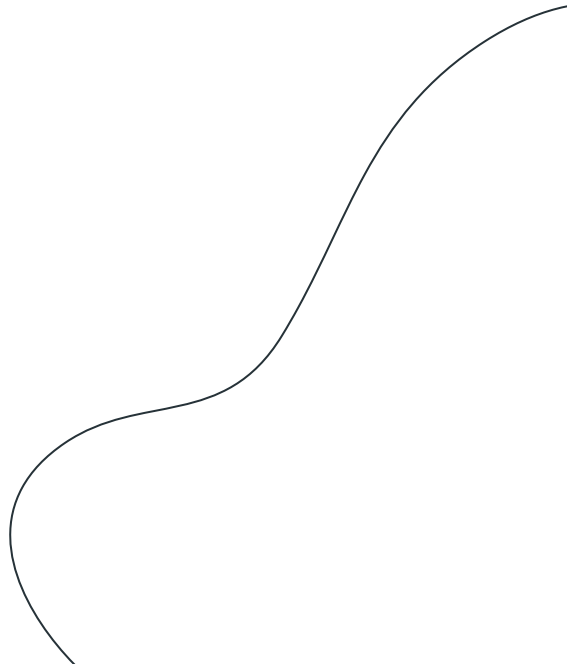
# Desenvolvimento do Modelo

Como temos acesso aos valores que pretendemos prever então podemos usar **algoritmos com supervisão**.

Este problema pode ser tratado como um problema de **classificação ou de regressão**, e como tal, de forma a variar a solução, decidimos tratar o problema como um de **regressão**.

Como ponto inicial o grupo decidiu comparar o desempenho de 4 algoritmos de regressão:

- Árvore de decisão
- Regressão Ridge
- Regressão Linear
- Rede Elástica



# Desenvolvimento do Modelo

Ao executar o excerto de código apresentado conseguimos desenhar a seguinte tabela:

Nome	Erro médio (%)	Tempo de execução (ms)
Ridge	2.87%	2.99
Decision Tree	4.20%	3.99
Linear Regression	2.86%	3.99
Elastic Net	3.65%	2.99

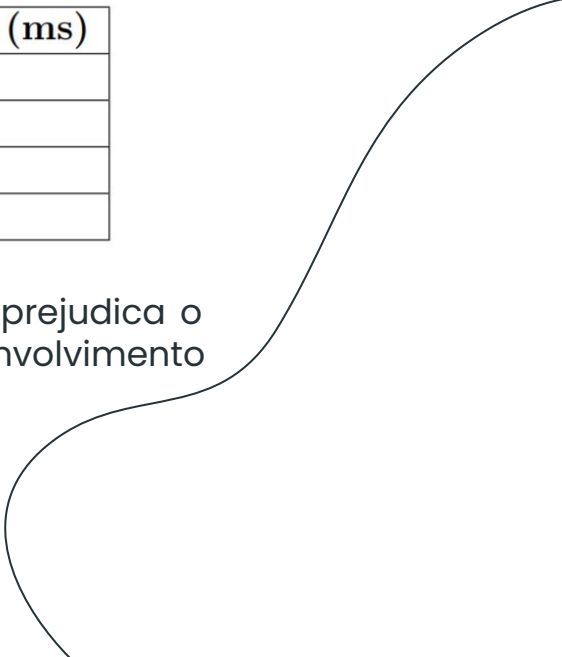
Como é possível ver, o **Ridge Regression** parece ser a melhor opção (sem ajustes) em termos de erro médio, porém a sua execução pode ser otimizada.

# Desenvolvimento do Modelo

De forma a otimizar o desempenho, tentamos criar uma *pipeline* que torne os valores mais uniformes utilizando o *StandardScaler*.

Nome (com scale)	Erro médio (%)	Tempo de execução (ms)
Ridge	2.87%	4.99
Decision Tree	4.18%	5.98
Linear Regression	2.86%	4.986
Elastic Net	4.24%	3.99

Curiosamente e ao contrário do *dataset* anterior, o uso do *scaler* prejudica o desempenho do algoritmo, por isso não vai ser usado no desenvolvimento desta etapa.





# Desenvolvimento do Modelo

Para otimizar mais o modelo, foi utilizado **gridSearchCV** para encontrar os parâmetros de *tuning* que resultam num menor erro médio.

O **gridSearchCV** irá executar o algoritmo com todas as combinações possíveis de parâmetros dados no dicionário e irá devolver os parâmetros resultantes da execução que retornou um menor erro médio.

Isso dá-nos os seguintes parâmetros:

- 'alpha': '0.001'
- 'tol': 0.0001

# Desenvolvimento do Modelo

Com os novos parâmetros, o desempenho do nosso modelo é o seguinte:

Nome	Erro médio (%)	Tempo de execução (ms)
Ridge (Afinado)	2.86%	1224

De notar que o tempo de execução tem em conta a procura dos parâmetros ótimos.



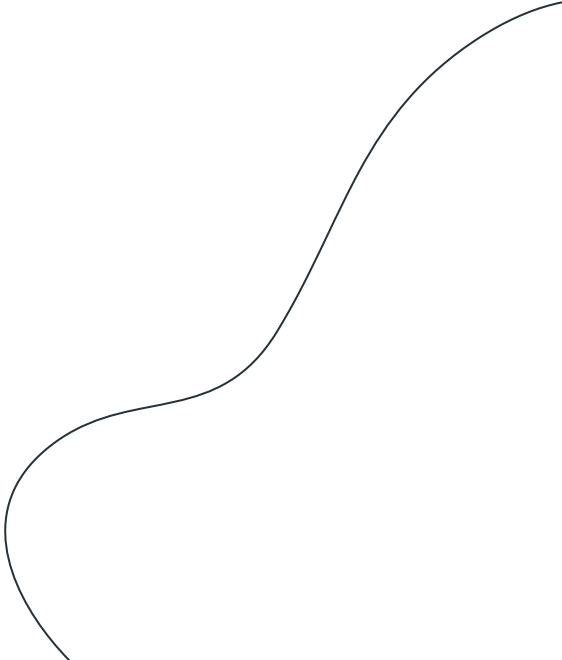
# Análise crítica dos resultados

Como o *dataset* tinha uma baixa quantidade de informação útil, focamo-nos mais na escolha e afinação dos algoritmos do que na preparação dos dados.

Como se trata de um problema de regressão, não faz sentido calcular a precisão na escolha dos resultados, e por isso optamos por utilizar o erro médio como medida de desempenho.

Quando, por exemplo, o modelo prevê o valor 83.65, podemos dizer que este acha mais provável o valor ser 84 do que 83.

O modelo tem um erro médio de 2.86, de onde se conclui que tem um desvio de 2.86% da realidade, o que é bom, considerando que as notas variam de 0 a 100.



# **DAA**

# **Trabalho Prático**

António Santos pg47031

Jorge Vieira pg47349

Manuel Moreira pg47439

Sara Dias pg47667

