Relatório - Projeto de Aprendizado de Máquina

Grupo:	**NUSP:**
Gabriel Santos Ribeiro	**9771380**
Matheus Resende Miranda	**9771696**
Luca Faria Curcio	**10295502**

1. Introdução

Anteriormente, o grupo havia escolhido trabalhar com uma base de dados referente à valores de ações na bolsa de valores, entretanto, após uma discussão com o professor em relação às técnicas necessárias para a realização dessa análise, o grupo optou por alterar o projeto para algo que melhor se encaixa no escopo da disciplina.

A base de dados escolhida foi obtida através do resultado de diversos testes de extinção de incêndio de quatro combustíveis diferentes com um sistema de extinção a ondas sonoras.

O sistema de ondas sonoras consiste em 4 subwoofers com uma potência total de 4,000 Watt posicionados numa cabine de colimador.

Um computador é usado como fonte da frequência, um anemômetro é usado para a medição do fluxo de ar resultado da onda sonora durante a fase de extinção e, um medidor de decibéis é usado para medir a intensidade da onda.

Um total de 17,442 testes foram executados para a formação da base de dados.

2. Propriedades da base de dados

	SIZE	FUEL	DISTANCE	DESIBEL	AIRFLOW	FREQUENCY	STATUS
0	1	gasoline	10	96	0.0	75	0
1	1	gasoline	10	96	0.0	72	1
2	1	gasoline	10	96	2.6	70	1
3	1	gasoline	10	96	3.2	68	1
4	1	gasoline	10	109	4.5	67	1

Figura 1 - Cabeçalho da base de dados.

A base de dados é composta por 17,442 linhas, cada uma referente a um experimento, totalizando o mesmo número de experimentos citado acima, além de 7 colunas.

Essas colunas são:

- **FUEL**
  - Indica qual dos quatro combustíveis foi utilizado, sendo eles, *gasolina*, *querosene*, *thinner* e *gás liquefeito de petróleo(GLP).*
- **SIZE**
  - Para *gasolina*, *querosene* e *thinner:*
    - Indica o tamanho da lata de combustível usada no experimento, os valores possíveis são, *7 cm, 12 cm, 14 cm, 16 cm* e *20 cm.*
    - Os valores são normalizados entre 1 a 5 na base de dados, sendo 1 referente à *7 cm*, 2 à *12 cm* e assim por diante.
  - Para *GLP:*
    - Os valores possíveis são 6 e 7, sendo 6 *half-throttle* e 7 *full-throttle.*
- **DISTANCE**
  - Indica a distância do equipamento de ondas sonoras até a lata de combustível, os valores vão de *10 cm* até *190 cm.*
- **DESIBEL**
  - Indicam os valores medidos de intensidade sonora, eles vão de *72 db* até *113 db.*
- **AIRFLOW**
  - Valores medidos de fluxo de ar, de *0* até *17 m/s.*
- **FREQUENCY**
  - Valores da frequência da onda gerada pelo computador, de *1 Hz* até *75 Hz.*
- **STATUS**
  - Estado final do experimento, *1* indica a extinção da chama enquanto *0* indica a não-extinção.

3. Análise dos dados

A análise dos dados foi feita buscando entender a influência dos atributos no experimento, além de entender como eles se relacionam. A primeira análise feita foi um gráfico com 10% dos dados gerados de forma pseudo aleatória dos atributos “DISTANCE”, “”SIZE” e “AIRFLOW” com os pontos verdes representando os incêndios apagados e os vermelhos os não apagados, uma vez que é intuitivo que um fluxo de ar maior sobre o incêndio implica em maior chance de apagar. Dessa forma, é possível ver como a distância entre o equipamento e o fogo, além do tamanho da chama influenciam no experimento.

O gráfico - Figura 1 - mostra que o tamanho da chama praticamente não tem influência quando comparada com a distância entre o equipamento e o incêndio e também fica perceptível que o fluxo de ar tem uma relação linear com essa distância.

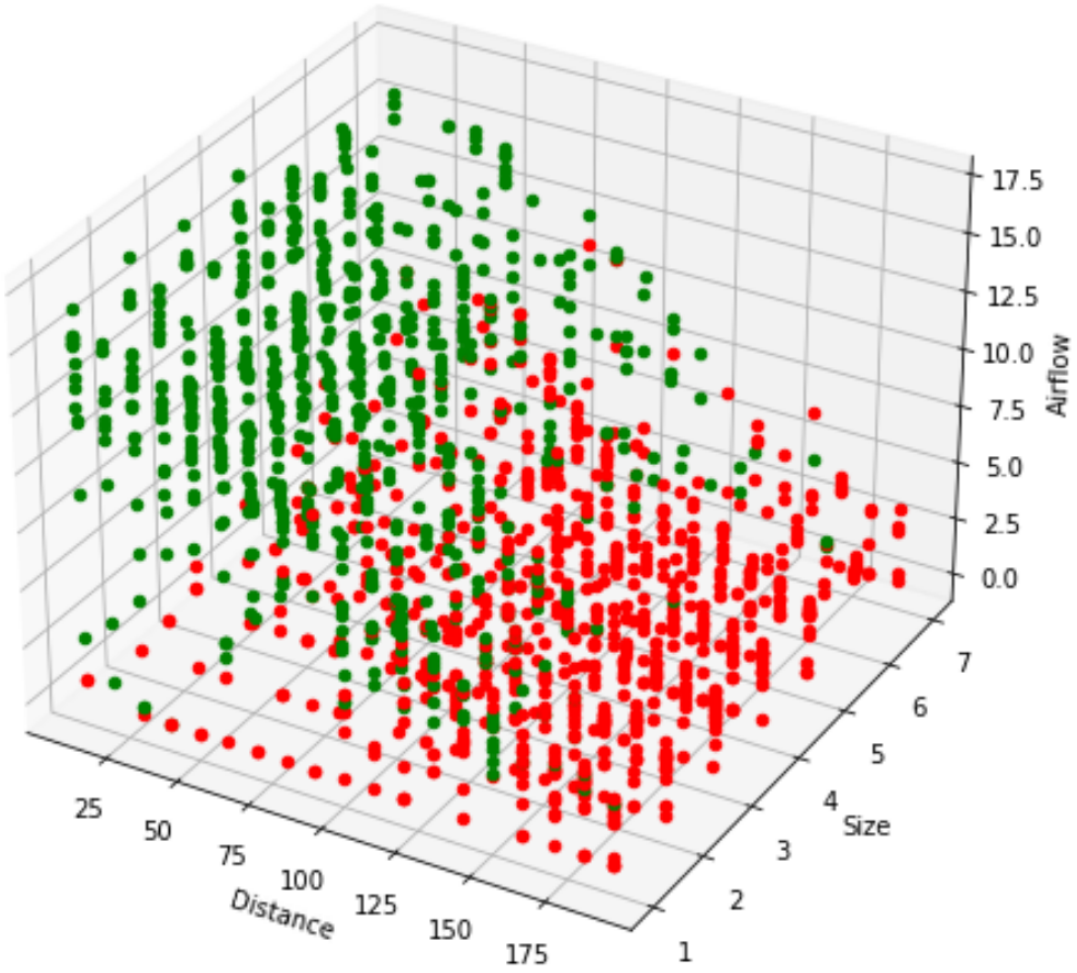

Figura 2 - Gráfico DISTANCE x SIZE x AIRFLOW.

Visto os resultados da primeira análise, uma segunda análise foi feita com os atributos referentes a onda sonora, com a finalidade de entender quais atributos

geram fluxos de ar maiores e consequentemente conseguem apagar o incêndio. Para que não houvesse influência da distância, que é um fator muito importante, foi fixada uma distância na moda do atributo “DISTANCE”, assim, todas as ondas analisadas estavam a uma distância equivalente de seus respectivos incêndios. O gráfico gerado - Figura 2 - que também apresenta os incêndios que foram apagados e os que não foram de maneira similar ao anterior, mostra que frequências médias e baixas com intensidades média-altas produzem fluxos de ar maiores a distâncias iguais.

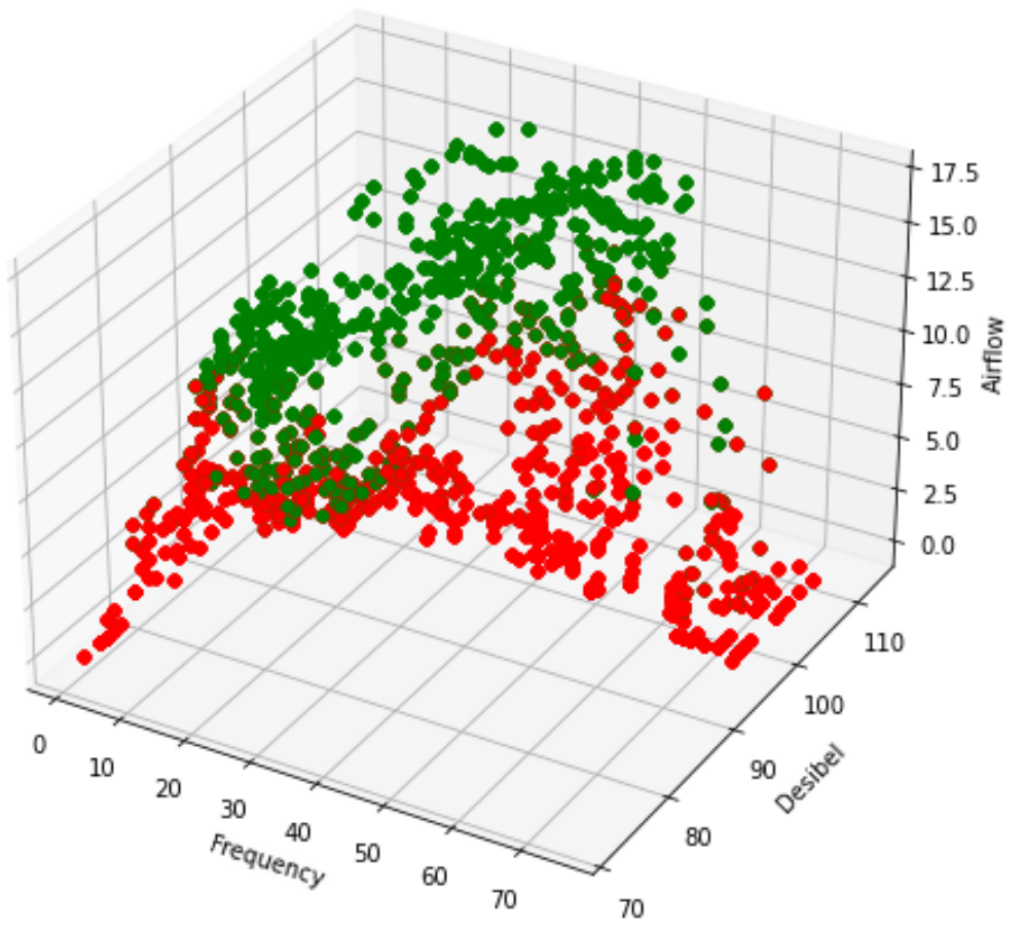

Figura 3 - Gráfico FREQUENCY x DESIBEL x AIRFLOW.

Uma terceira e última análise foi feita visando ver se o tipo de combustível utilizado na criação do incêndio tinha grande influência no experimento. Para isso, foram gerados gráficos relacionando “AIRFLOW” e “DISTANCE”, visto que esses foram os atributos mais importantes até então. O gráfico - Figura 3 - foi feito utilizando 80% das amostras de cada combustível e retorna a ideia de que os combustíveis fazem o incêndio ter praticamente o mesmo comportamento durante o experimento, com uma pequena exceção do querosene que parece ser mais difícil de apagar mesmo em distâncias menores e do GLP que tende a ser mais fácil de apagar mesmo a distâncias um pouco maiores.

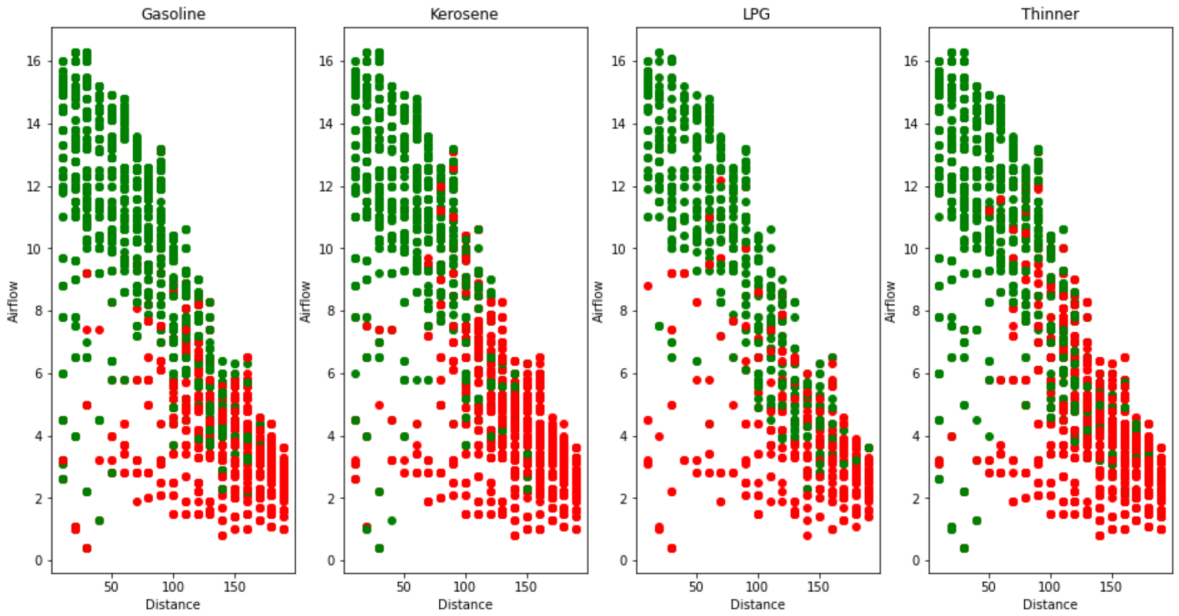

Figura 4 - Gráficos de DISTANCE x AIRFLOW.

4. Tratamento dos Dados

O tratamento dos dados para geração dos modelos começa com a deteção e remoção de *outliers* dos atributos “AIRFLOW” e “DESIBEL”, visto que esses são dados medidos com equipamentos de medição que podem fazer leituras errôneas, enquanto que os demais atributos não são medidos, mas sim setados no começo do experimento. Dessa forma foram identificados 629 *outliers* que foram retirados da base. Logo depois foi feita uma varredura em busca de dados incompletos ou vazios, mas nenhum foi encontrado.

Após essas remoções foi feita uma conversão dos dados qualitativos do atributo “FUEL” em numéricos usando a técnica one-hot encoding através do método get_dummies no pandas.

Por fim, para que os diferentes dados numéricos não ocasionassem em erros devido às diferentes magnitudes, foi feita uma normalização geral na tabela utilizando o método Min-Máx com intervalo entre 0 e 1, resultando na base da seguinte forma:

	gasoline	kerosene	lpg	thinner	SIZE	DISTANCE	DESIBEL	AIRFLOW	FREQUENCY	STATUS
0	1.0	0.0	0.0	0.0	0.0	0.0	0.466667	0.000000	1.000000	0.0
1	1.0	0.0	0.0	0.0	0.0	0.0	0.466667	0.000000	0.959459	1.0
2	1.0	0.0	0.0	0.0	0.0	0.0	0.466667	0.159509	0.932432	1.0
3	1.0	0.0	0.0	0.0	0.0	0.0	0.466667	0.196319	0.905405	1.0
4	1.0	0.0	0.0	0.0	0.0	0.0	0.900000	0.276074	0.891892	1.0
...	...	...	...	...	...	...	...	...	...	...
17434	0.0	0.0	1.0	0.0	1.0	1.0	0.300000	0.116564	0.094595	0.0
17435	0.0	0.0	1.0	0.0	1.0	1.0	0.300000	0.098160	0.081081	0.0
17436	0.0	0.0	1.0	0.0	1.0	1.0	0.166667	0.153374	0.067568	0.0
17437	0.0	0.0	1.0	0.0	1.0	1.0	0.133333	0.134969	0.054054	0.0
17438	0.0	0.0	1.0	0.0	1.0	1.0	0.066667	0.122699	0.040541	0.0

16813 rows × 10 columns

Figura 5 - Base de dados após tratamento de dados.

Após o tratamento, as classes (“STATUS”) e atributos foram separados para o treinamento dos modelos implementados.

5. Modelos

O primeiro modelo utilizado para classificar os dados foi o KNN (K-Nearest-Neighbours), dentro do qual foi variado o parâmetro K (número de vizinhos mais próximos) entre os valores 1, 3, 5, 7, 11 para a obtenção dos resultados, que nesse caso é o “STATUS” da chama, ou seja, se foi apagada ou não. Para cada variação, o modelo foi treinado e validado, utilizando tanto o método de cross-validation como também o método holdout, e a acurácia do modelo foi calculada em todas as situações. Os resultados mostraram-se similarmente muito altos.

```
Acurácia com 1 K-NN: 0.9623 +/- 0.0030
Acurácia com 3 K-NN: 0.9610 +/- 0.0032
Acurácia com 5 K-NN: 0.9606 +/- 0.0042
Acurácia com 7 K-NN: 0.9607 +/- 0.0034
Acurácia com 11 K-NN: 0.9596 +/- 0.0036
```

Figura 6 - Resultados obtidos com cross-validation.

```
Acurácia com 1 K-NN: 0.965
Acurácia com 3 K-NN: 0.961
Acurácia com 5 K-NN: 0.961
Acurácia com 7 K-NN: 0.964
Acurácia com 11 K-NN: 0.96
```

Figura 7 - Resultados obtidos com holdout.

Um outro modelo, o de regressão múltipla, também foi utilizado para que fosse utilizado com esses dados, e nesse caso uma predição do valor da distância (entre a emissão e a chama) foi feita. O modelo foi treinado de três modos diferentes, de uma maneira inicial com todos os dados relevantes incluídos, em seguida sem o tipo de combustível relevante, e em um último momento retirando ainda o atributo “SIZE”, referente ao tamanho da chama. Para cada um desses casos o erro quadrático médio foi calculado assim como a pontuação R2, e apesar das mudanças, os resultados foram bem similares e com uma pontuação relativamente baixa.

```
Número de amostras: 8592
Erro Médio Quadrático: 0.02
R Score: 0.54

Número de amostras: 8592
Erro Médio Quadrático: 0.02
R Score: 0.53

Número de amostras: 8592
Erro Médio Quadrático: 0.02
R Score: 0.53
```

Figura 8 - Resultados obtidos de todas as variações, em sequência.

6. Referências

1: KOKLU M., TASPINAR Y.S., (2021). Determining the Extinguishing Status of Fuel Flames With Sound Wave by Machine Learning Methods. IEEE Access, 9, pp.86207-86216, Doi: 10.1109/ACCESS.2021.3088612
Link: https://ieeexplore.ieee.org/document/9452168 (Open Access)
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9452168

2: TASPINAR Y.S., KOKLU M., ALTIN M., (2021). Classification of Flame Extinction Based on Acoustic Oscillations using Artificial Intelligence Methods. Case Studies in Thermal Engineering, 28, 101561, Doi: 10.1016/j.csite.2021.101561
Link: https://www.sciencedirect.com/science/article/pii/S2214157X21007243 (Open Access) https://www.sciencedirect.com/sdfe/reader/pii/S2214157X21007243/pdf

3: TASPINAR Y.S., KOKLU M., ALTIN M., (2022). Acoustic-Driven Airflow Flame Extinguishing System Design and Analysis of Capabilities of Low Frequency in Different Fuels. Fire Technology, Doi: 10.1007/s10694-021-01208-9
Link: https://link.springer.com/content/pdf/10.1007/s10694-021-01208-9.pdf