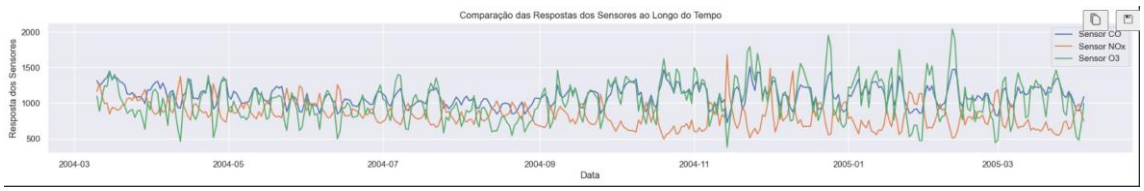
Cientista de Dados: Marina Lourenço Borges

1. Qualidade do Ar

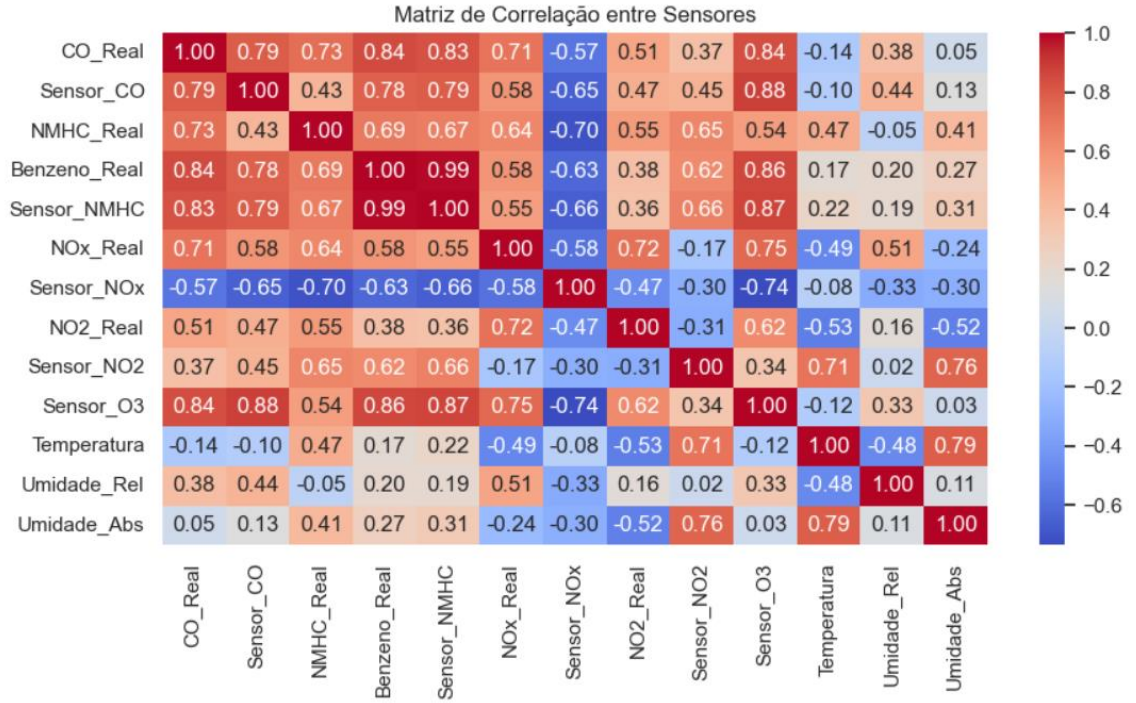
O DataFrame sobre Qualidade do Ar contém informações sobre a qualidade do ar em uma área poluída em uma cidade italiana, durante um ano. Ele inclui dados como a data, período do dia, concentrações médias horárias reais e sensoriais de vários poluentes atmosféricos, como monóxido de carbono (CO), hidrocarbonetos não metânicos (NMHC), benzeno, óxidos de nitrogênio (NOx) e dióxido de nitrogênio (NO2), bem como respostas dos sensores para esses poluentes. Além disso, o DataFrame também inclui informações sobre a temperatura, umidade relativa e umidade absoluta.

Inicialmente validamos que o dataset continha dados NaN mascarados com o número -200, sendo assim foi realizada a validação da informação e retirada as linhas que continham das três principais colunas: Temperatura, Umidade_Rel e Umidade_Abs.

Foi realizada avaliação as respostas dos sensores, conforme figura abaixo, sugerindo que pode haver alguma correlação entre as respostas dos sensores, visto que as linhas parecem seguir padrões semelhantes em alguns pontos, subindo e descendo juntas, o que poderia indicar uma resposta simultânea a certos fatores ambientais ou fontes de poluição.



A seguir foi realizada a análise da matriz de correlação entre todos os dados disponíveis:



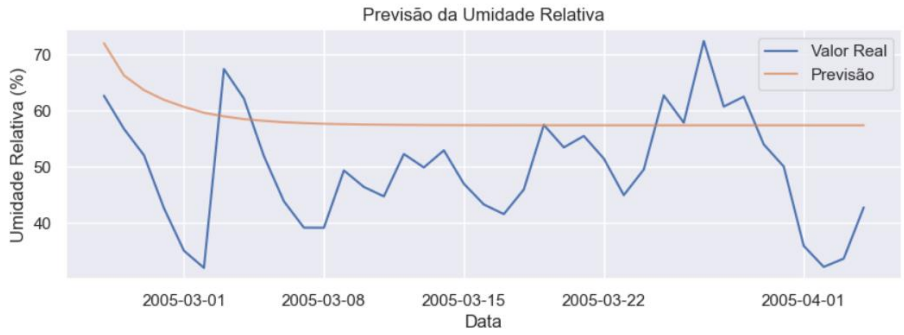
Conseguindo mapear as correlações acima de 0,7 das features:

```
[('Sensor_CO', 'CO_Real'),
 ('NMHC_Real', 'CO_Real'),
 ('Benzeno_Real', 'CO_Real'),
 ('Benzeno_Real', 'Sensor_CO'),
 ('Sensor_NMHC', 'CO_Real'),
 ('Sensor_NMHC', 'Sensor_CO'),
 ('Sensor_NMHC', 'Benzeno_Real'),
 ('NOx_Real', 'CO_Real'),
 ('NO2_Real', 'NOx_Real'),
 ('Sensor_O3', 'CO_Real'),
 ('Sensor_O3', 'Sensor_CO'),
 ('Sensor_O3', 'Benzeno_Real'),
 ('Sensor_O3', 'Sensor_NMHC'),
 ('Sensor_O3', 'NOx_Real'),
 ('Temperatura', 'Sensor_NO2'),
 ('Umidade_Abs', 'Sensor_NO2'),
 ('Umidade_Abs', 'Temperatura')]
```

Sendo assim é possível concluir que o gráfico e a matriz de correlação revelam um insight importante: a presença de correlações fortes entre as leituras dos sensores e as medições reais de poluentes, como CO e NOx, indicam que os sensores são eficazes em rastrear os níveis reais desses contaminantes.

Especificamente, a forte correlação entre 'Sensor_CO' e 'CO_Real' sugere que o sensor de CO é um indicador confiável da presença de monóxido de carbono no ambiente. Isso pode ser útil para o monitoramento da qualidade do ar em tempo real e para a implementação de medidas de controle de poluição mais eficazes.

Seguindo na estimativa da Umidade Relativa do Ar, realizado o primeiro modelo ARIMA (Autoregressive Integrated Moving Average) para a estimativa da Umidade Relativa do Ar, utilizando todas as variáveis disponíveis no banco de dados.



O gráfico mostra a comparação entre os valores reais de umidade relativa e as previsões geradas por um modelo de previsão. A linha azul representa os dados reais de umidade relativa ao longo do tempo, enquanto a linha laranja mostra as previsões do modelo.

Podemos ver que há algumas discrepâncias significativas entre as previsões e os valores reais, sugerindo que o modelo pode precisar ser mais afinado para melhor capturar a variação na umidade relativa. Isso pode envolver ajustar os parâmetros do modelo, incorporar outras variáveis preditoras ou usar um tipo diferente de modelo.

Para um segundo teste sabendo que, a umidade relativa a tem relação entre quantidade de água que existe no ar (umidade absoluta) e quantidade máxima de água que poderia existir na mesma temperatura (ponto de saturação).

Foi construído um segundo modelo com apenas as variáveis Temperatura e Umidade Absoluta, com o modelo SARIMA (Seasonal Autoregressive Integrated Moving Average), é uma extensão do ARIMA que adiciona suporte para padrões sazonais nos dados de séries temporais. Ele é capaz de modelar e prever séries que exibem sazonalidade, além de tendências e padrões não sazonais.

O conjunto de dados de treino (em azul) é usado para ajustar o modelo, e as previsões (em vermelho) são comparadas com os valores reais de teste (linha escura). A área sombreada em vermelho representa o intervalo de confiança das previsões. O modelo captura a tendência geral e algumas flutuações, mas existem desvios, especialmente em pontos de picos e vales.

O Erro Quadrático Médio (EQM) fornecido é uma medida da diferença entre os valores previstos e os valores reais, indicando a precisão das previsões do modelo. Um EQM de aproximadamente 9.93 sugere que as previsões estão, em média, a uma distância de cerca de 3.15 (a raiz quadrada do EQM) unidades percentuais dos valores reais.

O modelo SARIMA é uma escolha adequada para iniciar o estudo, dada a análise exploratória realizada. No entanto, é importante notar que ele utiliza apenas duas das 13 features disponíveis no dataset.

A umidade relativa, apesar de não apresentar correlações fortes com outras variáveis, uma vez que a umidade absoluta e a temperatura têm uma relação com outras variáveis, essas podem ser usadas indiretamente para estimar a umidade relativa.

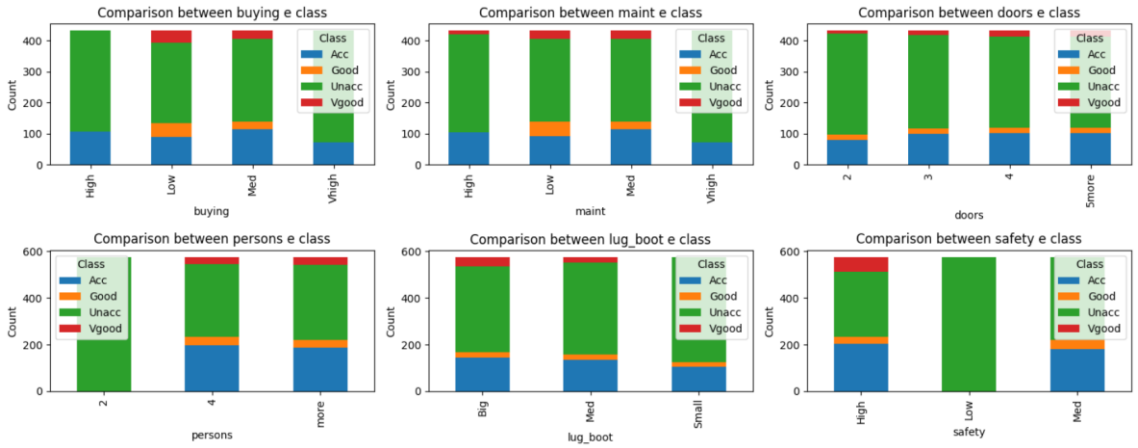
Entender como a temperatura e os poluentes afetam a umidade absoluta poderia, portanto, ser útil na modelagem e na estimativa da umidade relativa do ar. Por exemplo, se a temperatura aumenta e a umidade absoluta não muda, a umidade relativa diminuirá, pois o ar mais quente pode conter mais vapor d'água.

Sendo assim o estudo deve seguir para a próxima sprint com o conceito de realizar a resposta para as validações de como a temperatura e os poluentes afetam a umidade absoluta sendo uma inferência direta a umidade relativa no ar. Seguindo para a construção de outro modelo que traga as demais variáveis disponíveis como features para a realização do estudo.

2. Evolução dos Carros.

O DataFrame sobre a evolução de carros contém informações sobre diferentes aspectos dos veículos, como preço de compra, preço da manutenção, número de portas, capacidade de passageiros, tamanho do porta-malas e nível estimado de segurança. Com uma coluna que traz o nível de avaliação dos carros, classificando-os como inaceitáveis, aceitáveis, bons ou muito bons.

Primeiramente, realizando as séries de gráficos de barras abaixo foi possível construir uma comparação visual imediata entre as categorias e suas respectivas.



Nota-se que veículos com um preço de compra e custos de manutenção elevados são frequentemente associados a avaliações negativas, sendo muitas vezes classificados como inaceitáveis.

Em contrapartida, características como a capacidade de acomodar passageiros e um alto nível de segurança mostram-se como atributos valorizados, conduzindo a classificações mais positivas, tais como aceitável ou bom. De maneira semelhante, um porta-malas de dimensões generosas é um fator que contribui positivamente para a classificação de um veículo, reforçando a preferência por carros mais espaçosos e versáteis.

Curiosamente, o número de portas dos veículos, uma característica que poderia ser considerada como indicativa do tamanho ou do tipo do carro, aparenta ter uma influência menor nas avaliações quando comparado com os outros fatores mencionados, sugerindo que as prioridades dos consumidores podem estar alinhadas com aspectos mais funcionais e de segurança em detrimento do design ou da forma.

Ao avaliar a eficácia de diferentes modelos de machine learning na classificação de veículos com base em características como preço, manutenção, capacidade e segurança, observou-se uma variação significativa no desempenho.

- Regressão Logística, apesar de ser um método amplamente utilizado para problemas de classificação, alcançou uma acurácia de apenas 66.47%, com baixa precisão e recall para a maioria das classes, indicando uma adequação limitada a este conjunto de dados.

- Floresta Aleatória demonstrou uma melhoria substancial, com uma acurácia de 96.72%, evidenciando a sua capacidade de capturar a complexidade e as nuances dos dados.

- Árvore de Decisão exibiu um desempenho ligeiramente superior, com uma acurácia de 97.11%, e apresentou um recall particularmente mais alto para a classe rotulada como 'good'.

Esses resultados sugerem que, para este conjunto de dados específico, tanto a Floresta Aleatória quanto a Árvore de Decisão são modelos robustos, com a Árvore de Decisão oferecendo uma vantagem marginal em acurácia. A escolha final entre os dois pode depender de fatores adicionais, como complexidade do modelo e eficiência computacional, porém, do ponto de vista estritamente quantitativo, a Árvore de Decisão apresenta uma ligeiro desempenho superior.

A análise da importância dos recursos em um modelo de Floresta Aleatória revelou que a segurança (safety) foi identificada como o recurso mais crucial, seguida pela capacidade de passageiros (persons) e pelo preço de compra (buying).

Esses resultados sugerem que aspectos de segurança, praticidade e custo são considerações-chave para os consumidores ao avaliar carros. Em contraste, o tamanho do porta-malas (lug_boot) e o número de portas (doors) foram considerados os recursos menos influentes na decisão da classe do carro, indicando que características físicas menos impactam a escolha da classe.

Para os próximos passos no estudo deste modelo, sugere-se a aplicação de técnicas de balanceamento das classes (Oversampling, Undersampling e SMOTE), devido ao desbalanceamento observado, principalmente na classe 'unacceptable', que concentra 70% dos dados.

Esse desbalanceamento pode levar a resultados enviesados, onde o modelo pode ter dificuldade em aprender corretamente a classe minoritária, resultando em previsões menos precisas e confiáveis para essa classe. Portanto, é importante abordar esse desbalanceamento para garantir que o modelo seja capaz de generalizar bem para todas as classes de avaliação de carros.