Eduardo Kurek
Vinícius Kurek

Projeto Final
Análise de Dados Aplicada

Relatório do projeto final solicitado pelo professor Eduardo Pena na disciplina de Ciência de Dados do Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Universidade Tecnológica Federal do Paraná – UTFPR

Departamento Acadêmico de Computação – DACOM

Bacharelado em Ciência da Computação – BCC

Campo Mourão

Novembro / 2025

Resumo

Este trabalho analisou a qualidade do ar em Nova Iorque utilizando o dataset *Air Quality NYC*, relacionando poluentes (NO2, PM2.5 e O3) com dados de saúde e tráfego de veículos. Foram testadas hipóteses sobre a previsibilidade da poluição, relação entre taxa de poluentes e estação do ano, e a relação entre PM2.5 e indicadores de mortalidade. Aplicaram-se modelos de Machine Learning com diferentes conjuntos de features, sendo a regressão linear para NO2 a mais eficaz ($R^2 = 0.542$, RMSE = 0.353). Os resultados mostraram uma possibilidade de previsão do poluente NO2 baseado nos anos anteriores, uma relação positiva entre a estação do ano e a emissão de poluentes, e que o poluente PM2.5 causa um alto risco à saúde.

Sumário

1 Definiçao do Problema

O dataset escolhido foi o Air Quality NYC, ele possui os índices de gases emitidos na cidade de NY e suas principais fontes. Também possui dados hospitalares causados por esses poluentes, tráfegos de caminhões e carros. O tamanho inicial da tabela é de 18.862 entradas.

Os dados foram coletados em diferentes regiões da cidade. A coluna 'Geo Type Name' diz qual a granularidade da informação. Alguns são agregados de outras regiões, 'Citywide' por exemplo, representa a média para a cidade como um todo, já a média dos 5 'Borough' é o equivalente à 'Citywide', e assim por diante. Utilizamos principalmente os dados da cidade (Citywide), mas para algumas análises utilizamos os dados dos distritos (Borough).

A Tabela 1 exibe um exemplo de entrada para o dataset.

Tabela 1 – Exemplo de uma entrada do dataset

Coluna	**Valor**
Unique ID	336867375
Indicator ID	375
Name	Nitrogen dioxide (NO2)
Measure	Mean
Measure Info	ppb
Geo Type Name	CD
Geo Join ID	407
Geo Place Name	Flushing and Whitestone (CD7)
Time Period	Winter 2014–15
Start_Date	2014-12-01T00:00:00.000Z
Data Value	23.97

Ha duas separações de dados para os indicadores, aqueles medidos sazonalmente (verão e inverno 2), e aqueles medidos anualmente 3.

Tabela 2 – Indicadores com variação sazonal (verão e inverno)

Indicator ID	**Name**
375	Nitrogen dioxide (NO2)
386	Ozonone (O3)
365	Fine particles (PM 2.5)

Tabela 3 – Indicadores medidos anualmente

Indicator ID	**Name**
647	Outdoor Air Toxics - Formaldehyde
646	Outdoor Air Toxics - Benzene
651	Cardiovascular hospitalizations due to PM2.5 (age 40+)
652	Cardiac and respiratory deaths due to Ozone

650	Respiratory hospitalizations due to PM2.5 (age 20+)
659	Asthma emergency departments visits due to Ozone
661	Asthma hospitalizations due to Ozone
657	Asthma emergency department visits due to PM2.5
639	Deaths due to PM2.5
653	Asthma emergency departments visits due to Ozone
655	Asthma hospitalizations due to Ozone
648	Asthma emergency department visits due to PM2.5
644	Annual vehicle miles traveled (cars)
645	Annual vehicle miles traveled (trucks)
643	Annual vehicle miles traveled
642	Boiler Emissions- Total NOx Emissions
641	Boiler Emissions- Total PM2.5 Emissions
640	Boiler Emissions- Total SO2 Emissions

Escolhemos as seguintes perguntas e hipóteses para desenvolver neste trabalho:

- **P1**: É possível predizer a taxa de poluição emitida no ar nos próximos anos?
- **H1**: Haverá uma correlação negativa significativa entre o período do ano e a poluição do ar.
- **H2**: O período passado é um bom preditor para dizer se a poluição do ar vai subir ou diminuir no próximo ano
- **H3**: A concentração anual de PM2.5 apresenta uma relação maior com os indicadores de mortalidade do que com indicadores de morbidade

2 Metodologia

2.1 Limpeza de dados

Utilizamos algumas estratégias para a limpeza de dados, sendo elas:

- **Missing Values**: Não é todo indicador que possui dado de todos os anos (2005-2023). Dados sazonais possuem de 2008-2023, e indicadores anuais possuem de 2005-2019. Infeliz-

mente não temos o que fazer, alguns dados são medidas de um período maior (2012-2014), apenas distribuímos a mesma média para os anos 2012, 2013 e 2014;

- **Outliers**: Não identificamos nenhum outlier no dataset;
- **Inconsistências**: Não identificamos inconsistências no dataset;
- **Padronização**: Os indicadores sazonais terão a coluna inicial ‘Time Period‘ quebrada em duas: ‘Season‘ e ‘Year‘. Como o inverno começa em um ano e termina em outro, consideraremos o ano em que ele começa para manter a padronização. Os indicadores anuais possuem dois formatos: YYYY e YYYY-YYYY. Os anos que estão em um intervalo, separamos e copiamos a coluna ‘Data Value‘ para todos o intervalo.

2.2 Análise Exploratória

Realizamos algumas consultas SQL para explorar os dados, sendo elas:

2.2.1 Primeira Consulta

Essa consulta mostraria os anos em que a distribuição anual de PM2.5 em NYC está dentro dos limites sugeridos pela OMS (10 µg/m³ até 2020 e 5 µg/m³ a partir de 2021). Porém para mostrar os pontos num gráfico, não estamos filtrando fora os anos que não atendem a esses critérios por um motivo de visualização. Veja o resultado na Figura 1.

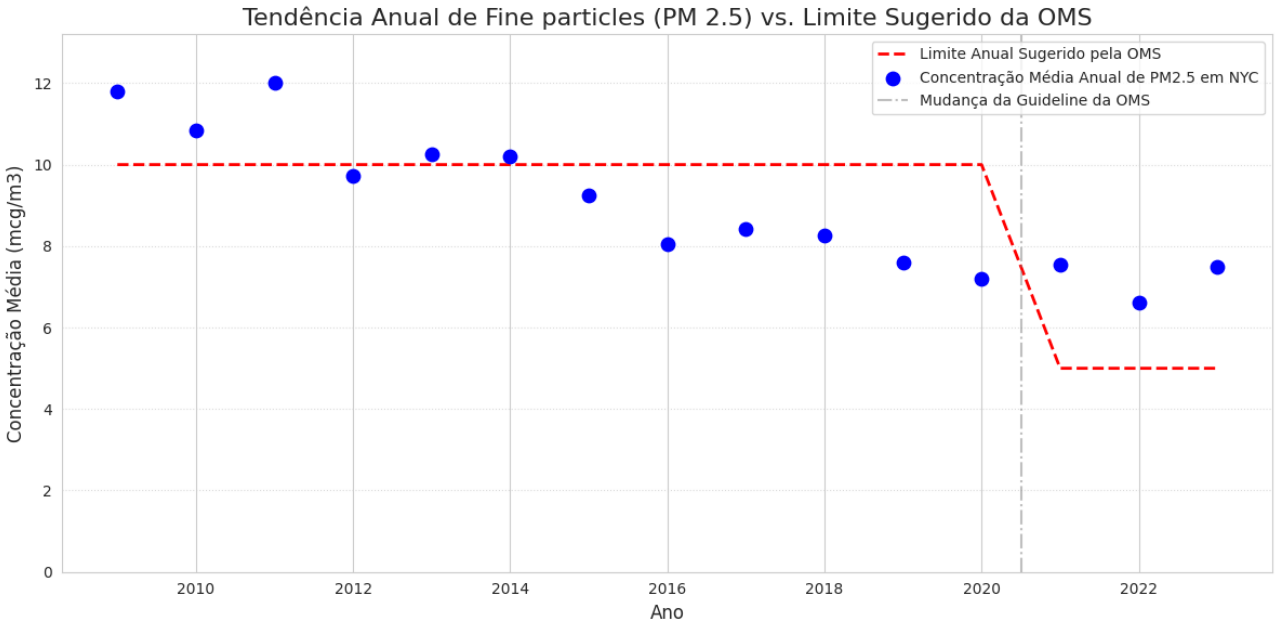

Figura 1 – Resultados da consulta 1

2.2.2 Segunda Consulta

Distribuição de Dióxido de Nitrogênio (NO2) ao longo dos anos por estação utilizando a medida de partes por bilhão (ppb) 2.

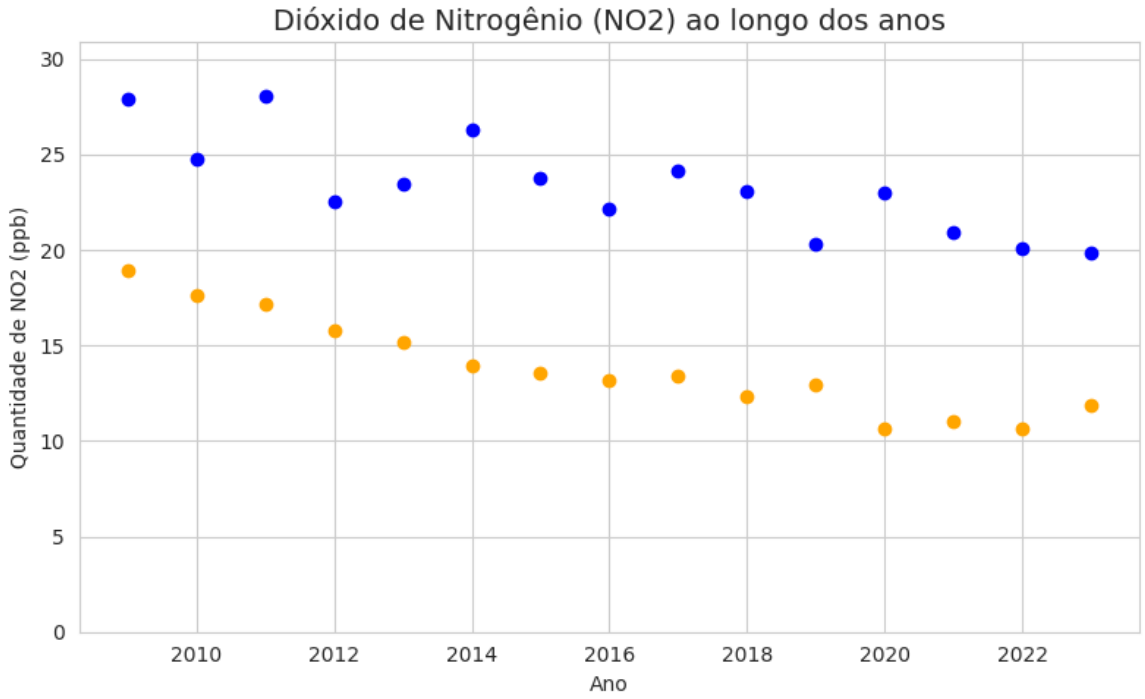

Figura 2 – Resultados da consulta 2. Pontos em azul são do inverno, já pontos em laranja são referentes ao verão

2.2.3 Terceira Consulta

Taxa de variação de mortes devido ao PM 2.5 ao longo dos anos 3

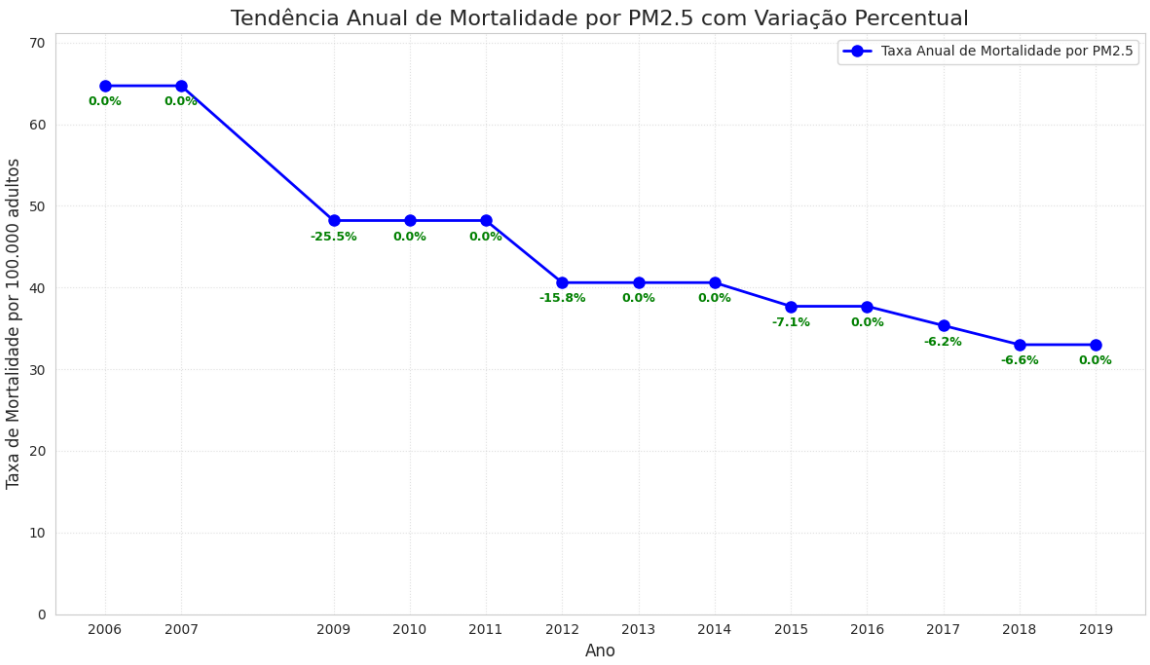

Figura 3 – Resultados da consulta 3

2.3 Aplicação das hipóteses

2.3.1 H1

Definição da hipótese: Haverá uma correlação negativa significativa entre o período do ano e a poluição do ar; Hipótese aplicada à PM2.5 e Dióxido de Nitrogênio (NO2).

Resultado: Hipótese totalmente falha, a hipótese se mostrou tendo uma correlação positiva entre o verão e o inverno pata o PM2.5 quanto para o NO2.

Tabela 4 – Correlação entre poluentes e estação do ano

Poluente	**Coeficiente r**	**P-valor**	**Significância ($P < 0.05$)**
PM2.5	0.695	0.00406	Sim
NO2	0.765	0.00090	Sim

2.3.2 H2

Definição da hipótese: O período passado é um bom preditor para dizer se a poluição do ar vai subir ou diminuir no próximo ano; Hipótese aplicada na maior resolução de dados possível para o conjunto enquanto fazendo uso de lag=1, isso significa que temos resultados para o verão/inverno de PM2.5, verão/inverno de NO2 e verão para O3.

Resultado: Há a possibilidade de utilizar regressão para fazer a correlação para todos indicadores exceto para o ozônio, o qual não mostrou correlação nenhuma com o mesmo período anterior.

Tabela 5 – Correlação entre poluentes por estação e ano

Poluente	**Estação**	**Coeficiente r**	**P-valor**	**Conclusão**
PM2.5	Verão	0.682	0.00724	Significativa ($\mathbf{P < 0.05}$)
PM2.5	Inverno	0.630	0.01583	Significativa ($\mathbf{P < 0.05}$)
NO2	Verão	0.934	0.00000	Altamente significativa ($\mathbf{P \ll 0.001}$)
NO2	Inverno	0.730	0.00702	Significativa ($\mathbf{P < 0.05}$)
O3	Anual	0.085	0.77211	Não significativa

2.3.3 H3

Definição da hipótese: A concentração anual de PM2.5 apresenta uma relação maior com os indicadores de mortalidade do que com indicadores de morbidade;

Resultado: Podemos analisar que a taxa de mortalidade é consideravelmente maior do que morbidades para o PM2.5.

Tabela 6 – Correlação entre indicadores de saúde e PM2.5

Indicador de Saúde	**Coeficiente r**	**P-valor**	**Significância ($P < 0.05$)**
Mortalidade	0.944	0.00001	Altamente significativa
Emergências de Asma	0.927	0.00004	Altamente significativa
Hospitaliz. Respiratórias	0.850	0.00093	Altamente significativa
Hospitaliz. Cardiovasculares	0.634	0.03611	Significativa

2.4 Modelagem com Machine Learning

2.4.1 Feature Engineering

Para feature engineering, analisamos dados do NO2, PM2.5 e 03, agrupando os valores por ano, de forma a ignorar as estações. Pensamos em realizar a regressão com dois tipos de parâmetros:

- LAGs de cada um dos poluentes, para identificar se é possível predizer o poluente com base em dados antigos.
- LAGs de outros poluentes, para identificar se é possivel predizer um poluente baseado em outro.

Logo, separamos o dataset em dois, um que continha apenas os LAGs (1, 2, 3) e YEAR baseados no próprio poluente. E outro que possuia os LAGs (1, 2, 3) e YEAR de todos os poluentes. Chamamos o primeiro dataset de simples e o segundo de complexo.

2.4.2 Seleção de Features

Para selecionar as melhores features para o conjunto Citywide, ou seja, aquelas que podem predizer melhor o futuro, optamos pelo seletor `SelectKBest`, usando o algoritmo *f_regression* do `scikit-learn`. Além disso ao lidarmos com o conjunto dos Boroughs (distritos) utilizamos o RFECV com uma simplificação do classificador a ser utilizado, por exemplo, ao selecionar features para o Linear Regression, também foi treinado o RFECV com o mesmo.

2.4.3 Treino

Optamos por testar as regressões com os dataframes separados de duas formas diferentes, quando nos referimos à citywide, estamos considerando entradas no dataframe as quais tiveram a coluna "GeoPlaceName"filtradas pela string "Citywide", o que resulta em dados considerando a média da cidade toda. Já "Borough"são as denominações para os 5 distritos os quais Nova Iorque é divido: Brooklyn, Queens, Staten Island, Bronx e Manhattan; Estamos imaginando que tal divisão pelo chamado "Boro"deveria resultar em dados mais precisos.

Para ambos os tipos de modelos resolvemos dividir seus respectivos datasets da seguinte forma: Dataset_treino=df[df['Year'] $<$ 2020] e Dataset_teste=df[df['Year'] $>=$ 2023]

Para o treino dos baselines utilizamos a regressão linear, já para o treino dos modelos em si foi utilizado a regressão linear, a random forest e o gradient_boosting

2.4.4 Treino do Modelo Citywide

Para o modelo simples, foi feita a seleção entre as 4 features eenquanto 3 delas foram aplicadas no conjunto chamado features_simples, que contém as seguintes features: ['Lag1',

'Lag2', 'Lag3', 'Year'], Note que os lags aqui são referentes ao próprio identificador, algo que será diferente para o modelo complexo.

Para o modelo complexo, fizemos a seleção entre 10 features, chamadas features complexas: ['Lag1_NO2', 'Lag2_NO2', 'Lag3_NO2', 'Lag1_PM25', 'Lag2_PM25', 'Lag3_PM25', 'Lag1_O3', 'Lag2_O3', 'Lag3_O3', 'Year'] Perceba que aqui, além de procurarmos por relações entre os lags do próprio indicador também estamos lidando com valores de outros indicadores, por isso temos que fazer a identificação dos lags. Como baseline, utilizamos o modelo de Regressão Linear(RL) com o TOP 1, escolhido com o SelectKBest, do conjunto de features simples.

Resultados: Ao executar o modelo Baseline, tivemos resultados do R^2 negativos para PM2.5 e O3. Apenas NO2 mostrou um valor positivo, de 0.18. Sendo assim, isso já nos deu uma dica de que ao menos o NO2 seria possível prever algo utilizando modelos de Machine Learning, e ao executar os testes com múltiplos parâmetros tivemos resultados ainda mais animadores para o NO2. Porém, para PM2.5 e 03 os resultados continuaram ruins, e por isso mostraremos apenas os resultados de NO2 na Tabela 7. O único modelo que teve bons resultados foi a regressão linear, os outros dois modelos utilizados não mostraram bons resultados.

Tabela 7 – Resultados de Regressão Linear para NO2 (Citywide)

Conjunto	R^2	**RMSE**	**Features**
Simples	0.542	0.353	Lag1(-0.81), Lag2(-0.87), Year(-2.16)
Complexo	0.488	0.373	Lag1_PM25(-0.17), Lag1_NO2(-0.35), Year(-1.27)

Observe que para o conjunto Citywide, tentar relacionar NO2 com Lags de outros indicadores foi um problema, como pode ser visto no R^2 de features complexas, que acabou por selecionar o Lag1 de PM2.5 para tentar prever o modelo de NO2 e isso reduziu a precisão do mesmo. Portanto as melhores features para tal foram as selecionadas no conjunto de features simples.

2.4.5 Treino do Modelo Borough

Aqui temos apenas um conjunto de features, que seria equivalente às features simples do modelo Citywide: ['Lag1', 'Lag2', 'Lag3', 'Year']. Algoritmo utilizado para seleção de features foi o RFECV do scikit-learn. Como baseline, utilizamos o modelo de Regressão Linear(RL) com o TOP 1 escolhido através do RFECV do conjunto de features.

Para testar também utilizamos o algoritmo RFECV e infelizmente não tivemos nenhum resultado bom em nenhum dos Boros/Indicadores, nem nos treinos baselines nem nos testes aplicados. Sendo assim nem apresentaremos aqui nessa seção, pois todos foram negativos.

3 Discussão dos resultados

Os resultados obtidos pelas hipóteses nos mostram que há uma grande relação entre a emissão de poluentes e a estação do ano, sendo que no inverno são emitidos mais poluentes. Também vimos que o período passado é um bom preditor para previsão apenas no poluente NO2, nos outros poluentes não houve relação. Já no PM2.5, notamos que no geral, enfermidades causadas por esse poluente são mais graves, e ocasionam em mais mortes.

Já para os modelos de Machine Learning, podemos concluir que é possível predizer apenas o poluente NO2, com as features LAG1, LAG2, e YEAR. O modelo com a regressão linear se mostrou o mais eficaz, random forest e gradient_boost não tiveram bons resultados. O R^2 foi de 0.542, já RMSE foi de 0.353.

4 Trabalhos Futuros

Para trabalhos futuros, pode-se analisar a relação entre os poluentes e a saúde da população, trabalhando em um dataset que contenha dados mais distribuídos e mais quantitativos, pois o dataset escolhido não possuía dados suficientes de hospitalidades para realizar uma análise precisa.

Também indicamos o uso de bases de dados de outros países, para se fazer uma comparação entre as leis relacionadas à redução de poluentes daquele governo. Analisando se estão sendo eficientes e tentar identificar quais leis demonstraram melhores resultados ao longo dos anos.