**Faculdade de Informática e Administração Paulista**

Engenharia de Software

Vitor da Silva Sant’ana RM99586

Lucca Alexandre Machado RM99700

Victor Augusto Wittner RM 98667

GLOBAL SOLUTIONS 2

Data Science and Statistical Computing

ESPW

Bela Vista

2024

1. INTRODUÇÃO

O projeto aborda o tema “Emissão de carbono”, contendo dois *datasets*:

* ***data\_countries*** = versa sobre os países e suas emissões de carbono
* **data\_influents** = versa sobre influentes (causas) de diversas emissões de carbono

Sabe-se que estes fatores são muito importantes para a tomada de decisões, criação de projetos e produtos e para a adesão de procedimentos e protocolos em empresas, sejam elas produtoras ou de assessoria.

1. RESUMO

O projeto versa sobre a análise de dois datasets que armazenam dados sobre:

* Países e suas emissões de carbono
* Influentes principais na geração de CO2

Os datasets ajudam a obter informações sobre pontos que mais geram gás carbônico e seus influentes, como a quilometragem dirigida em um carro, uso de sacos plásticos de lixo, e os países que mais produzem (em quilotoneladas), bem como a sua metrica de toneladas *per capita*.

1. ESCOPO

Esse projeto explora o a emissão de carbono de diferentes formas, analisando correlações entre suas colunas, para identificar padrões que possam impactar na pegada de carbono.

**Colunas do *data\_countries***

1. **Country**: Nome do país ao qual os dados pertencem.
2. **Region**: Continente ou região geográfica do país.
3. **Date**: Ano de registro das emissões de carbono.
4. **Kilotons of Co2**: Emissões totais de dióxido de carbono, medidas em quilotoneladas.
5. **Metric Tons Per Capita**: Emissões de dióxido de carbono por pessoa em toneladas métricas

**Colunas do *data\_influents***

1. ***Body Type:*** *Tipo corporal da pessoa.*
2. ***Sex:*** *Sexo da pessoa.*
3. ***Diet:*** *Tipo de dieta da pessoa.*
4. ***How Often Shower:*** *Frequência de banho da pessoa.*
5. ***Heating Energy Source:*** *Fonte de energia usada para aquecimento.*
6. ***Transport:*** *Meio de transporte predominante.*
7. ***Vehicle Type:*** *Tipo de veículo utilizado.*
8. ***Social Activity:*** *Frequência de atividades sociais.*
9. ***Monthly Grocery Bill:*** *Valor mensal gasto em supermercado.*
10. ***Frequency of Traveling by Air:*** *Frequência de viagens de avião.*
11. ***Vehicle Monthly Distance Km:*** *Distância percorrida mensalmente com veículo, em*
12. *quilômetros.*
13. ***Waste Bag Size:*** *Tamanho das sacolas de lixo usadas.*
14. ***Waste Bag Weekly Count:*** *Quantidade de sacolas de lixo usadas por semana.*
15. ***How Long TV PC Daily Hour:*** *Tempo diário gasto assistindo TV ou usando computador, em horas.*
16. ***How Many New Clothes Monthly****: Quantidade de roupas novas adquiridas mensalmente.*
17. ***How Long Internet Daily Hour:*** *Tempo diário gasto na internet, em horas.*
18. ***Energy efficiency:*** *Indicação de eficiência energética.*
19. ***Recycling:*** *Tipos de materiais reciclados pela pessoa.*
20. ***Cooking\_With:*** *Equipamentos usados para cozinhar.*
21. ***CarbonEmission:*** *Emissão de carbono gerada pela pessoa.*
22. OBJETIVO

O objetivo do projeto é realizar uma análise no *dataset*, observando padrões e buscando conclusões sobre as correlações entre as colunas e as informações respondendo as seguintes perguntas:

1. **Qual é a relação entre o total de CO2 emitido e a quantidade de roupas compradas?**

* Correlação relativamente baixa (0.198), porém se enquadra em uma das mais altas de todos os influentes.

1. **Existe alguma correlação entre a emissão de CO2 e transporte utilizado pelas pessoas cotidianamente?**

* Sim! Os carros, ônibus e caminhões são os meis de transporte que, dentre os outros do dataset, mais geram pegada de carbono.

1. **Qual a influência do nível de desenvolvimento econômico (PIB) de um país nas suas emissões de CO2?**

* Seguindo as observações do dataset, quanto mais os países produzem (e aumentam o seu PIB), mais carbono geram, por isso que os Estados Unidos e a China lideram disparadamente o ranking de maiores geradores de CO2.

1. **O tempo de uso da internet influencia diretamente a quantidade de CO2 emitido?**

* Sim, porém em baixíssima quantidade, considerado quase irrelevante.

1. **Existe uma tendência de aumento das emissões de CO2 ao longo do tempo?**

* Sim, seguindo toda a inovação e adesão da tecnologia até nas coisas mais simples, é possível afirmar que, sem correta abordagem, o nível de produção de CO2 tende a subir.

1. **Qual a relação entre o nível de urbanização e as emissões de CO2?**

* Com o aumento da urbanização, produção, industrialização e adaptação das pessoas a produtos mais tecnológicos e facilitadores, as emissões de CO2 aumentam, devido ao fato de ser difícil encontrar equilíbrio entre a praticidade e o cuidado com o meio ambiente.

1. **Os países com maior emissão de CO2 por capita estão mais ou menos propensos a implementar políticas ambientais?**

* Para os países que mais emitem CO2, a conscientização é um assunto que não agrada totalmente, pois irá limitar a produção a um nível ou a uma resposta ao meio ambiente. Entretanto, como temos organizações como a ONU, esquivar-se desses movimentos de cuidado é um perigo aos maiores produtores, pois indicam falta de interesse no meio ambiente, o que pode afetar diretamente a sua economia e a sua imagem por conta dos valores mundiais.

1. **Há uma relação entre o valor de compra em supermercados e as emissões de CO2?**

* Sim! Esta é uma das variáveis que também mostrou correlação entre o restante do dataset, mesmo que baixa (0.08).

1. **Quais são os 5 países que mais emitem CO2?**

* China, Estados Unidos, Canada, Brasil e Australia.

1. **Os tamanhos e quantidade de sacolas de lixo utilizadas por semana influenciam na emissão de CO2?**

Sim! Este é o segundo fator do dataset que mais influencia na emissão de CO2, com correlação de 0.15.

1. ANÁLISE EXPLORATÓRIA

O dataset foi analisado, buscando informações que anulem a qualidade dos dados, afim de limpar e tratá-los para que os cálculos sejam feitos com dados bons.

De início, foi verificado se existia algum valor nulo no *dataset*, usando a função ***isnull().any(),*** e o resultado foi que, de todas as colunas dos dois datasets, apenas a coluna ‘Vehicle Type’ do *dataset* de influents continha um ou mais valores nulos, sabendo disso, foi executada uma query que troca todos os valores nulos por ‘Not informed’.

Logo após isso, foram gerados gráficos de violinos, que são boxplots mais aprofundados, pois apresentam a densidade dos dados também, para todas as colunas do *dataset.*

Após isso, foram extraídas dos datasets as suas colunas numéricas, que nos interessam nesse momento para gerar gráficos de violino, pois gráficos não indicam palavras diferentes.

Os gráficos do dataset de influentes está corretamente avaliado, a distribuição segue a média e os valores máximos.

Entretanto, o dataset de países trouxe dúvida nos gráficos gerados, por o valor máximo não é acompanhado pela distribuição dos dados, o que gera dúvida e possibilidade de existência de algum *outlier,* que é um valor extremo quando comparado à distribuição do conjunto.

Gráfico

Descrição gerada automaticamente Gráfico, Histograma

Descrição gerada automaticamente

Sendo assim, foram calculadas as medianas das colunas e as mesmas foram comparadas às suas médias, para avaliar se são muito diferentes.

Tela de celular com aplicativo aberto

Descrição gerada automaticamente

Quanto à coluna Metric Tons Per Capita, a mesma aparenta estar nos conformes e não contar com outliers.

A coluna Kilotons of Co2 apresenta valores muito discrepantes, necessitando de avaliação, pois pode representar um outlier, mas não é certo, pois a unidade de medida, os dados, e o tipo de dado podem influenciar no resultado.

Sendo assim, foi feita uma query para retornar os registros com os valores de quilotoneladas de co2 por país mais altas, o que resultou em registros com valores parecidos, porém poucos.  
 O que diminui a probabilidade de existência de valores errôneos é de que os países que têm os números altos são superprodutores (Estados Unidos e China), têm muitas indústrias e movimentam muito capital e serviços, o que aumenta significantemente a produção de CO2. Então, a ideia de outliers foi amenizada.

Foi usada agora a função ***sns.kdeplot()*** para plotar histogramas de linhas contínuas,

função da biblioteca Seaborn e, colocando-o na mesma linha, comparamos com o ***violinplot,*** para

verificar se os dados condizem nos dois gráficos.

1. ESTATÍSTICA

Como início, foram obtidas as métricas de variação de dados dos datasets:

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

E os valores obtidos foram:

Texto

Descrição gerada automaticamente

Em resumo, todos os datasets apresentam variação moderada/alta, explica por causa de alguns fatores:

* Quantidade de dados – como são muitos dados avaliados, é provável que se tenha uma variação maior.
* Campo avaliado – o material que estão inseridos os dados abrem espaço para muita variação (geração de carbono)
* Países – países como a China e os Estados Unidos causam um impacto coletivo no dataset, por mais que sejam apenas dois, a força que eles têm para alterar a média, o desvio padrão e a variância é muito alta.

Analisando a correlação entre as variáveis dos datasets:

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

O dataset de países não apresentou correlação negativa ou positiva significativa.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

O dataset de influentes apresentou correlação positiva ou negativa em alguns pontos:

* Distância percorrida de carro em quilômetros X emissão de carbono
* Aquisição de roupas novas por mês X emissão de carbono
* Valor mensal gasto em supermercados X emissão de carbono
* Quantidade de sacolas de lixo utilizadas por semana X emissão de carbono

(Distância em quilômetris dirigida mensalmente X emissão de carbono)

Plotando um gráfico de calor para visualizar e entender melhor:

Gráfico

Descrição gerada automaticamente com confiança média

Dataset de países

Como o dataset só tem duas colunas numéricas e as duas foram analisadas, resultando no que foi dito anteriormente, as variáveis só têm correlação entre si mesmas.

Gráfico

Descrição gerada automaticamente

Dataset de influentes

Para avaliar melhor as correlações, foram plotados gráficos de dispersão (Scatter), facilitando a visualização:

Plotando gráfico de dispersão para as correlações mais altas:

Km’s rodados de carro por mês | Emissão de carbono

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

O Gráfico mostra concentração maior e quase total na parte central-inferior do gráfico, pois os dados seguem uma linha, que sobe um pouco.

Os dados fora dessa média são poucos, apresentando correlação maior e variância um pouco menor.

Roupas novas por mês | Emissão de carbono

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

O gráfico apresenta uma correlação consideravelmente boa, porém com foco na parte central-inferior do gráfico.

Os dados fora dessa média estão em quantidade maior, o que indica variância maior dos dados.

Valor gasto em supermercados | Emissão de carbono

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

O gráfico apresenta uma correlação mediana, principalmente por conta da parte inferior total que está um pouco mais dispersa que os outros dois, com concentração na parte central-inferior do gráfico.

Os dados fora da região onde está concentrada a maior parte do gráfico é grande, o que indica dispersão e variância altos.

Quantidade de sacolas de lixo utilizadas por semana | Emissão de carbono

Gráfico

Descrição gerada automaticamente

Este gráfico é diferente, ele não apresenta valores do conjunto dos reais, apenas inteiros, então os pontos serão distribuidos assim, em fileiras.

O gráfico apresenta concentração na parte central-inferior dos dados, entre 0 e 5000 sacolas de lixo.

Os dados fora dessa média não são muitos, o que indica uma dispersão e variância menores.

Como o gráfico apresenta valores inteiros e únicos, a correlação se torna um pouco mais acídua, podendo alterar para menor ou maior mais facilmente, uma vez que “não tem meio termo”, a quantidade contém valores absolutos.

1. TESTE DE HIPÓTESE

Os testes de hipótese servem para avaliar e calcular uma correlação, para averiguar se as variáveis se influenciam de fato ou não.

Fatores usados no teste de hipótese:

* **Valor p**: É a chance de que os resultados que foram apresentados acontecessem por acaso. Se for baixo, significa que o resultado não aconteceu por sorte, foi estatisticamente influenciado.
* **Graus de liberdade**: É o número de opções que temos para fazer os cálculos.
* **Chi²:** Serve para ver se as contagens que foram observadas são diferentes do que era esperado.
* **Frequências esperadas**: São os números que deveriam aparecer se tudo estivesse normal.

Nos testes de hipótese, os resultados obtidos foram:

kms por mês X emissão de carbono

|  |  |
| --- | --- |
| Chi-square | 16606018.02 |
| Valor p | 0.0 |
| Graus de liberdade | 14039016 |
| Hipótese nula | Rejeitada |

quantidade de roupas novas por mês x emissão de carbono

|  |  |
| --- | --- |
| Chi-square | 176303.01 |
| Valor p | 0.06386 |
| Graus de liberdade | 175400 |
| Hipótese nula | Não rejeitada |

valor mensal gasto em supermercados X emissão de carbono

|  |  |
| --- | --- |
| Chi-square | 875094.80 |
| Valor p | 0.11268 |
| Graus de liberdade | 873492 |
| Hipótese nula | Não rejeitada |

quantidade de sacolas de lixo utilizadas por semana X emissão de carbono

|  |  |
| --- | --- |
| Chi-square | 21482.88 |
| Valor p | 0.01750 |
| Graus de liberdade | 21048 |
| Hipótese nula | Rejeitada |

Como previsto, os modelos que tinham a maior correlação foram confirmados.

Hipótese nula rejeitada significa que as variáveis influenciam uma na outra.

Hipótese nula não rejeitada informa não correlação.

Sacolas de lixo utilizadas por mês e quantidade de quilômetros rodados de carro são as variáveis que mais influenciam na emissão de carbono, ou seja, que mais necessitam de atenção!

1. MACHINE LEARNING

Foram usados alguns métodos da biblioteca scikit-learn, e assim treinados os dados, usando quatro tipos de classificadores:

* Regressão Linear
* Regressão Logística
* Random Forest
* Árvore de Decisão

Regressão Linear

Para a regressão linear não foi possível fazer um gráfico de matriz de confusão, pois o mesmo informa apenas sobre modelos de classificação, e não de valores contínuos como o da regressão linear.

Gráfico, Gráfico de linhas, Gráfico de dispersão

Descrição gerada automaticamente

A linha azul retrata uma ocorrência normal dos dados, e o modelo de regressão são os pontos em vermelho.

O gráfico induz grande concentração perto da linha de previsão, em casos normais, o que indica acerto do modelo de aprendizado de máquina.

Concentração na parte esquerda-inferior do gráfico.

Grande dispersão dos dados em relação à apresentação média de uniformização.

Considerando:

* Erro absoluto médio: 601
* Erro absoluto quadrático: 585830

Regressão Logística

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

A linha azul retrata uma ocorrência normal dos dados, e o modelo de regressão são os pontos em vermelho.

O gráfico induz grande concentração ao redor, porém em pontos absolutos, da linha de previsão, em casos normais, o que indica certo acerto do modelo de aprendizado de máquina.

Concentração na parte esquerda-inferior do gráfico.

Baixa dispersão dos dados em relação à apresentação média de uniformização.

Sabendo que, os dados estão dispersos em pontos pois os valores não tinham alto grau de liberdade, cabiam em valores absolutos.

Considerando:

* Erro absoluto médio: 772
* Erro absoluto quadrático: 1005049

Random Forest

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

A linha azul retrata uma ocorrência normal dos dados, e o modelo de Random Forest são os pontos em vermelho.

O gráfico induz grande concentração ao redor, da linha de previsão, em casos normais, o que indica certo acerto do modelo de aprendizado de máquina.

Concentração na parte esquerda-inferior do gráfico.

Grande dispersão dos dados em relação à apresentação média de uniformização.

Valores muito dispersos ao redor da média e em locais com valores muito diferentes.

Considerando:

* Erro absoluto médio: 856
* Erro absoluto quadrático: 1246346

Árvore de decisão

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

A linha azul retrata uma ocorrência normal dos dados, e o modelo de regressão são os pontos em vermelho.

O gráfico induz concentração ao redor da linha de previsão, porém em baixa quantidade.

Concentração na parte esquerda-inferior do gráfico.

Muito alta dispersão dos dados em relação à apresentação média de uniformização.

Dados muito dispersos e em locais muito diferentes da média.

Considerando:

* Erro absoluto médio: 863
* Erro absoluto quadrático: 1287226

1. HALLWAY TESTING

**Sugestões de melhoria (criadas por pessoas de fora do trabalho):**

* Maior variedade de gráficos
* Melhorar eficácia dos modelos de machine learning
* Encontrar novos classificadores na literatur estatística que se encaixem melhor no projeto

1. CONCLUSÃO

Com a finalização do projeto, foi possível confirmar que a emissão de carbono está, na maior parte das vezes, tomando início em atos cotidianos, onde é necessário redobrar a atenção.

Andar de carro e utilizar sacolas de lixo são os principais fatores domiciliares que aumentam a emissão e pegada de carbono.

Com isso, teoria confirmada por testes de hipótese e machine learning, faz-se importante tomar atenção com os atos, criando medidas alternativas sustentáveis.

1. BIBLIOGRAFIA

KAGGLE. Acesso em 19/11/2024. Disponível em:

https://www.kaggle.com/datasets/ravindrasinghrana/carbon-co2-emissions

KAGGLE. Acesso em 19/11/2024. Disponível em:

https://www.kaggle.com/datasets/dumanmesut/individual-carbon-footprint-calculation

MEDIUM. Acesso em 20/11/2024. Disponível em:

https://medium.com/@jvsavietto6/machine-learning-métricas-validação-cruzada-bias-e-variância-380513d97c95