Universidade Presbiteriana Mackenzie

Tecnologia em Ciência de Dados - Projeto Aplicado 1

Franciele do Nascimento

João Victor Mendes Cunha

Leandro Rodrigues Dos Santos

Luiz Eduardo de Mei Salvador Coelho

Matheus Neves de Castro

Projeto Aplicado 1 - Programa Luz Para Todos

SÃO PAULO

2º Semestre de 2023

Autores:
Leandro Rodrigues Dos Santos / 23019689 / lers.138@gmail.com
Franciele do Nascimento / 23002042 / f.paterni@hotmail.com
João Victor Mendes Cunha / 23016094 / jvictormendesc@gmail.com
Luiz Eduardo de Mei Salvador Coelho / 23024585 / Luiz02coelho@gmail.com
Matheus Neves de Castro / 23011742 / 10923011742@mackenzista.com.br

Projeto Aplicado 1 - Programa Luz Para Todos e Mais Luz Para Amazônia

Professor: Everton Knihs

SÃO PAULO

2º Semestre de 2023

Índice

1. Introdução

A busca pelo desenvolvimento sustentável é um desafio complexo, especialmente quando se trata de regiões remotas e carentes de acesso a recursos básicos, como é o caso da Amazônia. O Ministério de Minas e Energia (MME), por meio do Programa Mais Luz para a Amazônia (MLA), visa proporcionar acesso à energia elétrica a essas regiões.

O Programa Mais Luz para a Amazônia (MLA) foi estabelecido para conceder acesso à energia elétrica às comunidades remotas dos estados da Amazônia Legal no Brasil. Seu objetivo principal é impulsionar o desenvolvimento socioeconômico dessas áreas, estimulando o aumento da renda familiar e a utilização sustentável dos recursos naturais. O processo de universalização teve início com o Decreto nº 4.873 em 2003, que criou o Programa Nacional de Universalização do Acesso à Energia Elétrica (Luz para Todos), e foi ampliado pelo Decreto nº 10.221 em 2020, que estabeleceu o Programa Mais Luz para a Amazônia. O programa ficou congelado durante o governo Bolsonaro e foi retomado recentemente em agosto de 2023.

No entanto, o sucesso e o impacto dessas iniciativas governamentais exigem uma análise profunda e embasada em dados. É nesse ponto que a metodologia de Descoberta de Conhecimento em Bases de Dados (KDD - Knowledge Discovery in Databases) emerge como uma ferramenta essencial. O processo de KDD, conduzido por um analista de dados e um especialista de domínio, oferece a capacidade de extrair insights valiosos a partir dos dados coletados no âmbito do programa MLA.

Este trabalho acadêmico objetiva fazer uma análise, do ponto de vista da metodologia KDD, do Dataset disponibilizado na página do programa Mais Luz para Todos e Mais Luz para a Amazônia. Utilizando esta base de dados, exploraremos como as conclusões aqui geradas podem nos mostrar o impacto do programa e como a tomada de decisões do poder público pode ser aprimorada com base em uma análise como esta.

2. Objetivos e Metas

O trabalho busca:

- Analisar os datasets dos programas "Mais Luz para Todos" e "Mais Luz para a Amazônia" usando a metodologia KDD.
- Explorar conclusões obtidas para entender o impacto dos programas.
- Utilizar análises para revelar insights sobre eficácia e áreas a serem melhoradas.
- Avaliar a utilidade do KDD na tomada de decisões governamentais para inclusão energética e desenvolvimento regional.
- Identificar oportunidades de otimização das políticas públicas com base nos resultados da análise.
- Exemplificar como análises de dados podem aprimorar iniciativas governamentais nos setores de energia e desenvolvimento.
- Avançar a compreensão do uso estratégico da análise de dados em políticas públicas, especialmente programas abrangentes.

3. DataSet e Metadados

Referências de aquisição do dataset:

Origem dos dados: Ministério de Minas e Energia (MME)

Link para o dataset: https://dados.gov.br/dados/organizacoes/visualizar/ministerio-de-minas-e-energia

CNPJ: 37.115.383/0001-53

Poder: Executivo **Esfera:** Federal

Localização: Esplanada dos Ministérios - Bloco U - Brasília/DF, CEP: 70.065-90

Site: mme.gov.br

Horário de atendimento ao público: Segunda à sexta-feira, das 8h às 18h

Telefone Geral: (61) 2032-5555 Protocolo: protocolo@mme.gov.br – Telefone: (61) 2032-5192

Dados públicos, sem limitação de uso

Descrição da origem:

O Ministério de Minas e Energia (MME) é uma instituição do Poder Executivo Federal responsável por formular e executar políticas públicas para a gestão sustentável dos recursos energéticos e minerais, contribuindo para o desenvolvimento socioeconômico do país. Foi criado em 1960 pela Lei nº 3.782 e tem passado por alterações legislativas ao longo dos anos. Sua missão é abranger uma ampla gama de temas relacionados à energia, mineração, recursos hídricos e outros aspectos ligados a essas áreas.

Descrição do dataset:

O dataset contém informações relacionadas aos programas "Luz para Todos" e "Mais Luz para a Amazônia", que fazem parte do processo de Universalização do Acesso e Uso da Energia Elétrica. O programa "Luz para Todos" foi instituído pelo Decreto nº 4.873 de 2003, enquanto o programa "Mais Luz para a Amazônia" foi instituído pelo Decreto nº 10.221 de 2020. Esses programas visam expandir o acesso à energia elétrica em domicílios, especialmente em áreas rurais e na região amazônica.

Os dados incluem informações como a quantidade de domicílios atendidos pelos programas e a quantidade de recursos aplicados. Esses dados são fornecidos no formato CSV e são provenientes do Sistema de Controle do Acesso a Energia Elétrica (SCAEE).

4. Repositório

   O link para o repositório do trabalho é:

   https://github.com/MatNev/Projeto_Aplicado_I

5. Cronograma geral

Cronograma Geral do Projeto Aplicado 1				
	Tarefa	Responsável	Status	Concluído em:
Etapa 1 : Kick-Off	Criar grupo	Leandro	Completo	15/08
	Definição da temática	Leandro	Completo	21/08
	Escolha do dataset	Leandro	Completo	21/08
	Criar Github	Matheus Castro	Completo	21/08
	Fazer sumário	Leandro	Completo	28/08
	Introdução	João Cunha	Completo	28/08
	Objetivos e metas	Leandro	Completo	28/08
	Estruturar cronograma geral	João Cunha/Matheus Castro	Completo	25/08
	Dataset e Metadados	Leandro	Completo	28/08
ETAPA 2 - *Definição do produto*	- Elaboração da proposta de solução analítica.	Leandro	Não iniciado	
	- Análise exploratória de dados.	Luis Coelho e Matheus Castro	Não iniciado	
ETAPA 3 - *Storytelling*	- Como apresentar resultados analíticos?	João Cunha	Não iniciado	
	- Data Storytelling.	Franciele	Não iniciado	
ETAPA 4 - *Encerramento*	- Orientação para ajustes finais.	Todos do grupo	Não iniciado	
	- Apresentação dos resultados.	Todos do grupo	Não iniciado	

6. Elaboração da proposta de solução analítica.

Analisando os conjuntos de dados relativos aos programas "Mais Luz para Todos" e "Mais Luz para a Amazônia" por meio da aplicação da metodologia KDD (Knowledge Discovery in Databases), nosso objetivo principal é explorar conclusões derivadas desses dados, com vistas a compreender o impacto desses programas nas regiões em que foram implementados.

Como parte do processo de preparação dos dados, realizamos a conversão do formato de arquivo original, que era CSV, para o formato .xlsx, e procedemos com a correção de palavras que continham erros ortográficos ou acentuação não reconhecida.

O conjunto de dados que será analisado é um arquivo Excel (.xlsx) de origem governamental, datado de julho de 2023, emitido pelo Ministério de Minas e Energia do Governo Federal do Brasil.

Essa abordagem visa aplicar rigor acadêmico ao processo de análise de dados, garantindo que as informações extraídas sejam sólidas e fundamentadas, contribuindo assim para um entendimento mais abrangente e embasado sobre o impacto e eficácia dos programas mencionados.

A seguir uma tabela com o conteúdo dos dados:

Nome do Campo	Tipo de Dado	Descrição
Lpt1Programa	Alfanumérico	Programa em que foi celebrado o contrato
Lpt1QtdDomicilios	Numérico (PK)	Quantidade de domicílios atendidos pelo programa
Lpt1Mes	Alfanumérico	Mês do atendimento
Lpt1Ano	Numérico	Ano do atendimento
Lpt1Município	Alfanumérico	Município onde se localiza o domicílio atendido
Lpt1Estado	Alfanumérico	Estado onde se localiza o município

7. Análise exploratória de dados.

Temos para análise o número de domicílios atendidos pelos programas e os anos que o programa está em atividade.

Número de exemplares (linhas) e dimensões (colunas) : linhas : 158608 e colunas 6
Tipos de dados: numéricos e alfanuméricos
Valores perdidos ou incorretos: nenhum

Scripts da Análise Exploratória em Python

Coluna Ano:

```
import pandas as pd

excel = pd.read_excel (r"C:\Users\dante\Downloads\Projeto Aplicado\Aula 02\Folder\
Domiciliosatendidos.xlsx")

# Carregue os dados do arquivo Excel em um DataFrame do pandas
df = pd.read_excel (r"C:\Users\dante\Downloads\Projeto Aplicado\Aula 02\Folder\
Domiciliosatendidos.xlsx")

# Calcule as medidas de posição (média, moda, mediana)
media = df["Ano"].mean()
moda = df["Ano"].mode().iloc[0]
mediana = df["Ano"].median()

# Calcule as medidas de dispersão (variância, desvio padrão, amplitude)
variancia = df["Ano"].var()
desvio_padrao = df["Ano"].std()
amplitude = df["Ano"].max() - df["Ano"].min()

print("Média:", media)
print("Moda:", moda)
print("Mediana:", mediana)
print("Variância:", variancia)
print("Desvio Padrão:", desvio_padrao)
print("Amplitude:", amplitude)
```

Coluna Domicílios:

```
import pandas as pd

excel = pd.read_excel (r"C:\Users\dante\Downloads\Projeto Aplicado\Aula 02\Folder\
Domiciliosatendidos.xlsx")

# Carregue os dados do arquivo Excel em um DataFrame do pandas
df = pd.read_excel (r"C:\Users\dante\Downloads\Projeto Aplicado\Aula 02\Folder\
Domiciliosatendidos.xlsx")

# Calcule as medidas de posição (média, moda, mediana)
media = df["Domicilios"].mean()
moda = df["Domicilios"].mode().iloc[0]
mediana = df["Domicilios"].median()

# Calcule as medidas de dispersão (variância, desvio padrão, amplitude)
variancia = df["Domicilios"].var()
desvio_padrao = df["Domicilios"].std()
amplitude = df["Domicilios"].max() - df["Domicilios"].min()

print("Média:", media)
print("Moda:", moda)
print("Mediana:", mediana)
print("Variância:", variancia)
print("Desvio Padrão:", desvio_padrao)
print("Amplitude:", amplitude)
```

Resultados obtidos :

Coluna Ano

Média: 2009.4980171114769
Moda: 2006
Mediana: 2009.0
Variância: 18.172888329455127
Desvio Padrão: 4.262967080503335
Amplitude: 19

Média (Mean): A média é de aproximadamente 2009.50. Essa medida sugere uma tendência central dos dados em torno desse ano.

Moda (Mode): A moda é 2006, o valor que ocorre com mais frequência no conjunto de dados. Isso sugere que o ano 2006 é o mais comum nos dados, indicando possivelmente um pico de ocorrência desse ano.

Mediana (Median): A mediana é 2009, o valor que divide o conjunto de dados ao meio quando ordenados. Isso sugere que metade dos valores estão abaixo de 2009 e metade estão acima. A mediana é útil para entender a centralização dos dados e não é afetada por valores extremos.

Variância (Variance): A variância é de aproximadamente 18.17 anos

Desvio Padrão (Standard Deviation): O desvio padrão é de aproximadamente 4.26. Ele é a raiz quadrada da variância e fornece uma medida da dispersão dos valores. Um desvio padrão baixo indica que os valores estão agrupados em torno da média.

Amplitude (Range): A amplitude é 19, a diferença entre o maior e o menor valor no conjunto de dados. Isso representa a variação total dos anos. Uma amplitude relativamente pequena sugere que os dados estão concentrados em um intervalo estreito.

Em resumo, os resultados indicam que os dados estão agrupados em torno da média, com uma variação relativamente baixa em relação à média. O ano de 2006 é a moda, indicando que é o ano mais comum nos dados, enquanto a mediana representa o ano do meio. A baixa variância e desvio padrão sugerem que os valores estão próximos da média, com pouca dispersão. A amplitude pequena também confirma que os dados estão concentrados em um intervalo estreito de anos. Isso pode indicar que há uma tendência ou padrão nos dados que se concentra em torno do ano de 2009.

Coluna Domicílios

Média: 23.059132869249396
Moda: 1.0
Mediana: 6.0
Variância: 3295.845826383289
Desvio Padrão: 57.40945763881844
Amplitude: 2841.0

1. Média (Mean): A média é de aproximadamente 23.06. Isso indica que, em média, os valores no conjunto de dados estão próximos desse valor. Essa medida sugere uma tendência central dos dados em torno desse ponto.

Moda (Mode): A moda é 1.0, o valor que ocorre com mais frequência no conjunto de dados. Isso pode ser indicativo de que existe uma categoria ou condição dominante no conjunto de dados.

Mediana (Median): A mediana é 6.0, o valor que divide o conjunto de dados ao meio quando ordenados. Isso sugere que metade dos valores estão abaixo de 6.0 e metade estão acima. A mediana é útil para entender a centralização dos dados e não é afetada por valores extremos.

Variância (Variance): A variância é de aproximadamente 3295.85. Isso indica que os valores no conjunto de dados estão relativamente dispersos em relação à média. Valores mais altos de variância representam maior dispersão, o que pode indicar uma grande variedade de valores no conjunto de dados.

Desvio Padrão (Standard Deviation): O desvio padrão é de aproximadamente 57.41. Ele é a raiz quadrada da variância e fornece uma medida da dispersão dos valores. Um desvio padrão maior indica maior variabilidade nos dados em relação à média.

Amplitude (Range): A amplitude é 2841.0, a diferença entre o maior e o menor valor no conjunto de dados. Isso representa a variação total dos valores. Uma amplitude tão grande sugere uma grande variação nos dados, com valores extremamente diferentes entre si.

Em resumo, os resultados indicam que os dados possuem uma ampla dispersão em relação à média, com uma variância e desvio padrão consideráveis. A presença de uma moda clara em 1.0 sugere uma categoria ou condição que é dominante no conjunto de dados. A mediana e a amplitude oferecem informações adicionais sobre a centralização e a variação dos dados, respectivamente. Essas medidas são essenciais para entender a distribuição e a tendência dos valores no conjunto de dados, o que pode ser valioso para análises posteriores e tomada de decisões.