Mineração de Dados

Atividade do dia **05/10/2024**

Nesta atividade vocês devem seguir o passo a passo a seguir encontrando uma base de dados no formato CSV e executar os passos abaixo com os códigos sugeridos em cada etapa. **Observe que os códigos são exemplos que devem ser alterados de acordo com a base de dados escolhida**.

Esta atividade deve ser salva em um repositório do Github e o link enviado na atividade correspondente.

Guia de Atividade: Análise de Dados com Google Colab (Foco no KDD)

Objetivo:

Realizar uma análise de dados em quatro etapas principais: busca por uma base de dados, limpeza e estruturação dos dados, visualização e conclusões a partir dos dados observados.

Passos da Atividade

1. Como Encontrar uma Base de Dados

Antes de iniciar a análise de dados, é importante selecionar um conjunto de dados que seja relevante para o estudo.

Fontes sugeridas:

- Kaggle - Base de dados de várias áreas.
- UCI Machine Learning Repository - Repositório de datasets clássicos.
- Google Dataset Search - Ferramenta de busca para bases de dados.

Atividade: Escolha um dataset em formato CSV que contenha pelo menos uma variável numérica e uma variável categórica para realizar a análise.

2. Limpeza e Estruturação dos Dados

Agora que o dataset foi baixado, vamos importar e realizar uma limpeza nos dados, preparando-os para a análise.

Passos:

1. Importe o dataset para o Google Colab e visualize as primeiras linhas:

```
import pandas as pd

# Suba o arquivo CSV no Colab
from google.colab import files
uploaded = files.upload()

# Leia o dataset
df = pd.read_csv('nome_do_arquivo.csv')

# Exiba as primeiras linhas do dataframe
df.head()
```

2. **Exploração dos dados:**
    - Verifique a presença de valores nulos:

```
df.isnull().sum()
```

3. **Tratamento dos dados faltantes:**
    - Para valores numéricos, substitua por média/mediana:

```
df['coluna_numérica'].fillna(df['coluna_numérica'].mean(), inplace=True)
```

    - Para valores categóricos, substitua por um valor padrão:

```
df['coluna_categórica'].fillna('Desconhecido', inplace=True)
```

4. **Conversão de Tipos de Dados:** Assegure-se de que os tipos de dados estão corretos (numérico vs categórico):

```
df['coluna_categórica'] = df['coluna_categórica'].astype('category')
df['coluna_numérica'] = pd.to_numeric(df['coluna_numérica'], errors='coerce')
```

5. **Exploração Estatística dos Dados:**
    - Estatísticas descritivas para variáveis numéricas:

```
df.describe()
```

    - Frequência de valores para variáveis categóricas:

```
df['coluna_categórica'].value_counts()
```

3. Criação de Gráficos (Visualização dos Dados)

Agora que os dados estão limpos e estruturados, vamos criar gráficos para entender a distribuição e os padrões nas variáveis numéricas e categóricas.

Gráficos para variáveis numéricas:

1. **Histograma** - Visualizar a distribuição de uma variável numérica:

```
import matplotlib.pyplot as plt
import seaborn as sns

sns.histplot(df['coluna_numérica'], bins=10, kde=True)
plt.title('Distribuição da Variável Numérica')
plt.show()
```

2. **Boxplot** - Detectar outliers e a dispersão dos dados:

```
sns.boxplot(x=df['coluna_numérica'])
plt.title('Boxplot da Variável Numérica')
plt.show()
```

Gráficos para variáveis categóricas:

1. **Gráfico de Barras** - Visualizar a contagem de categorias:

```
sns.countplot(x='coluna_categórica', data=df)
plt.title('Distribuição de Categorias')
plt.show()
```

2. **Gráfico de Pizza** - Proporções de categorias:

```
df['coluna_categórica'].value_counts().plot.pie(autopct='%1.1f%%')
plt.title('Proporção das Categorias')
plt.show()
```

Gráficos relacionais:

1. **Gráfico de Dispersão** - Verificar correlação entre duas variáveis numéricas:

```
sns.scatterplot(x='coluna_numérica_1', y='coluna_numérica_2', data=df)
plt.title('Correlação entre Variáveis Numéricas')
plt.show()
```

4. Conclusão dos Dados Observados

Após a exploração dos dados e a criação de gráficos, é hora de analisar os insights obtidos.

Atividade: Com base nos gráficos e análises realizadas, responda às seguintes perguntas:

1. Quais padrões você observou nas variáveis numéricas? (exemplo: distribuição normal, outliers)
2. Como estão distribuídas as variáveis categóricas? Alguma categoria se destaca em termos de frequência?
3. Você identificou correlações entre variáveis numéricas? Qual pode ser a relação entre essas variáveis?
4. Com base nos dados, quais são as principais conclusões que você pode tirar? O que essas conclusões indicam para o contexto do dataset escolhido?