Álgebra Linear - PCA

Fernanda Rafaela

Novembro, 2024

1 Dataset

O dataset é sobre preferencia de destino de férias, Praia ou Montanha, tendo diversas informações sobre estilo de vida das pessoas.
Link: https://www.kaggle.com/datasets/jahnavipaliwal/mountains-vs-beaches-preference?resource=download

2 Desenvolvimento

Para realizar a normalização dos dados e aplicar o algoritmo de PCA, serão usadas as seguintes bibliotecas no Python:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
```

Figure 1: Bibliotecas utilizadas para análise

Primeiro, importamos o arquivo CSV, transformamos as colunas de texto em valor numérico e normalizamos o dataset, calculando a média de cada coluna e subtraindo do total:

```python
df = pd.read_csv('mountains_vs_beaches_preferences.csv')
df['Preference'] = df.iloc[:, -1].map({0: 'Praia', 1: 'Montanha'})

text_columns = df.select_dtypes(include=['object']).columns
labels = df[text_columns[-1]] if len(text_columns) > 0 else None

label_encoder = LabelEncoder()
for col in text_columns:
    df[col] = label_encoder.fit_transform(df[col])

df_scaled = (df - df.mean()) / df.std()
```

Figure 2: Dataset

Em seguida, calculamos a matriz de covariância, transpondo-a, calculamos os autovetores e autovalores ordenando do maior para o menor, e calculamos a variancia explicada:

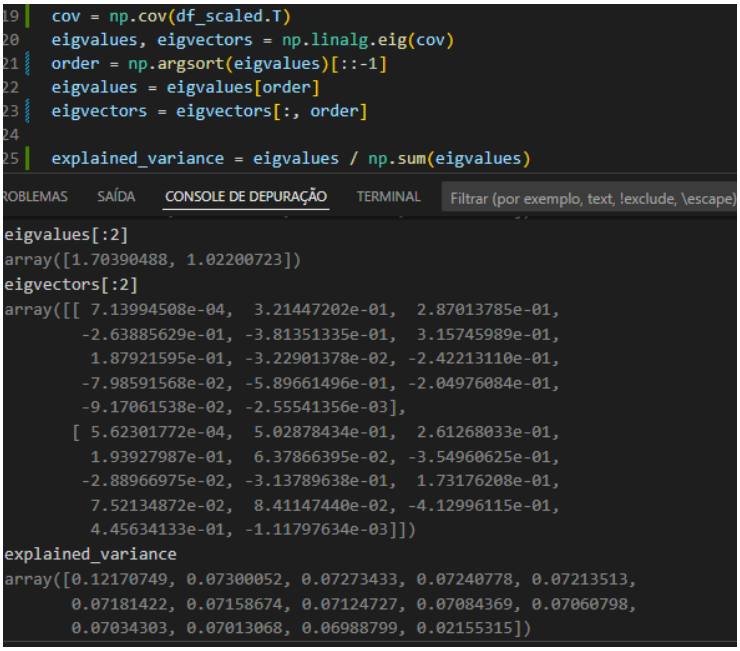

Figure 3: Algorítmo

Usando os 2 maiores autovalores calculamos o pca e colocamos a label que definimos durante o tratamento inicial do dataset:

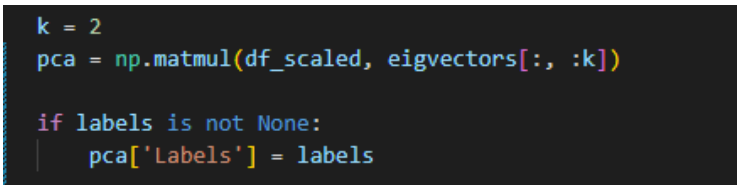

Figure 4: PCA Dataset

Por fim, o gráfico final usando os valores do PCA:

Figure 5: Gráfico PCA 2 dimensões

repo: https://github.com/xfeehrxx/pca