UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

TRABALHO UNIDAD 1 – CIÊNCIA DE DADOS

LUIZ AFFONSO HENDERSON GUEDES DE OLIVEIRA

SELEÇÃO E ADEQUAÇÃO DE DADOS

RELATÓRIO DE EXECUÇÃO

DISCENTE[1]: ERNANE FERREIRA ROCHA JUNIOR
DISCENTE[2]: QUELITA MIRIAM NUNES FERRAZ

NATAL/RN, 14 DE ABRIL DE 2024

1. DA BASE DE DADOS ESCOLHIDA

A base de dados selecionada para este estudo é denominada como um **conjunto de dados de previsão de diabetes**. Este conjunto de dados compreende uma variedade de informações médicas e demográficas de pacientes, juntamente com seu status em relação à diabetes (positivo ou negativo). Entre as características incluídas estão idade, sexo, índice de massa corporal (IMC), presença de hipertensão, histórico de doenças cardíacas, hábitos de tabagismo, níveis de HbA1c e glicose no sangue.

A concepção inicial desta base de dados foi primariamente direcionada para o desenvolvimento de modelos de aprendizado de máquina visando prever a ocorrência de diabetes em pacientes com base em seus históricos médicos e informações demográficas. Este recurso revela-se de grande utilidade para iniciativas voltadas à melhoria da identificação precoce da doença, auxiliando os profissionais de saúde na identificação de pacientes em risco de desenvolver diabetes e na formulação de planos de tratamento personalizados.

Ademais, este conjunto de dados representa uma valiosa ferramenta para pesquisadores, assim como nós, interessados em explorar as relações entre diversos fatores médicos e demográficos e a probabilidade de desenvolvimento de diabetes, abrindo portas para investigações mais aprofundadas nesse campo crucial da saúde pública.

2. DAS CARACTERÍSTICAS E ATRIBUTOS EXISTENTES

Um script em Python foi desenvolvido para capturar informações sobre o arquivo. Para isso, utilizamos a biblioteca Pandas, que facilita a leitura e a exibição dos dados.

```
1  import pandas as pd
```

Foi criada uma função chamada *show_csv_info* para encapsular as operações. Podemos utilizá-la da seguinte maneira:

```
1  file_path = 'data/diabetes_prediction_dataset.csv'
2  show_csv_info(file_path)
```

A função e suas instruções estão detalhadas abaixo:

```
1   def show_csv_info(file_path):
2       """
```

```
    Display information about a CSV file.

    Args:
        file_path (str): The path to the CSV file.

    Returns:
        pandas.DataFrame: The DataFrame containing the CSV data.
    """
    file_size = os.path.getsize(file_path)
    print(f"File size: {file_size / (1024*1024):.2f} MB")

    data = pd.read_csv(file_path)

    num_rows, num_cols = data.shape
    print(f"Number of rows: {num_rows}")
    print(f"Number of columns: {num_cols}")

    print("\nAttributes and data types:")
    print(data.dtypes)

    print("\nDataFrame description:")
    print(data.describe())

    print("\nAdditional information:")
    print(""" .... """)

    return data
```

Inicialmente, exibimos o tamanho do arquivo a ser utilizado. Por exemplo, o arquivo CSV '*diabetes_prediction_dataset.csv*' possui um tamanho de 3.63 MB. Em seguida, apresentamos o número de linhas e colunas do arquivo, que totalizam 100.000 linhas e 9 colunas, respectivamente. Essas colunas representam os seguintes atributos, com seus respectivos tipos de dados:

COLUNA	TIPO	DESCRIÇÃO
gender	object	Refere-se ao sexo biológico do indivíduo, o qual pode ter impacto em sua susceptibilidade ao diabetes. Existem três categorias: masculino, feminino e outros.

age	float64	Um fator importante, pois o diabetes é mais comumente diagnosticado em adultos mais velhos. A idade varia de 0 a 80 anos em nosso conjunto de dados.
hypertension	int64	Condição médica na qual a pressão sanguínea nas artérias está persistentemente elevada. Possui valores 0 ou 1, onde 0 indica que o paciente não possui hipertensão, e 1 indica que o paciente possui hipertensão.
heart_disease	int64	Outra condição médica associada a um maior risco de desenvolver diabetes. Assim como a hipertensão, tem valores 0 ou 1, onde 0 indica a ausência de doença cardíaca e 1 indica a presença.
smoking_history	object	Considerado um fator de risco para o diabetes e pode exacerbar as complicações associadas ao diabetes. Em nosso conjunto de dados, há 5 categorias: nunca, ex-fumante, atual, nunca fumou e sem informações.
bmi	float64	Uma medida de gordura corporal com base no peso e na altura. Valores de IMC mais altos estão relacionados a um maior risco de diabetes. A faixa de IMC no conjunto de dados varia de 10,16 a 71,55.
HbA1c_level	float64	Uma medida do nível médio de açúcar no sangue de uma pessoa ao longo dos últimos 2-3 meses. Níveis mais altos indicam um maior risco de desenvolver diabetes.
blood_glucose_level	int64	Refere-se à quantidade de glicose no sangue em um determinado momento. Níveis elevados de glicose no sangue são um indicador-chave de diabetes.
diabetes	int64	A variável alvo sendo prevista, com valores 1 indicando a presença de diabetes e 0 indicando a ausência.

Table 1: Atributos e seus tipos presentes no .csv

Além disso, fornecemos uma breve descrição do DataFrame montado com o Pandas, que inclui estatísticas importantes, como média, desvio padrão, valores mínimos e máximos para cada coluna.

```
count  100000.000000  100000.00000  100000.000000  100000.000000  100000.000000       100000.000000  100000.000000
mean       41.885856       0.07485       0.039420      27.320767       5.527507          138.058060       0.085000
std        22.516840       0.26315       0.194593       6.636783       1.070672           40.708136       0.278883
min         0.080000       0.00000       0.000000      10.010000       3.500000           80.000000       0.000000
25%        24.000000       0.00000       0.000000      23.630000       4.800000          100.000000       0.000000
50%        43.000000       0.00000       0.000000      27.320000       5.800000          140.000000       0.000000
```

```
75%        60.000000       0.00000       0.000000      29.580000       6.200000         159.000000       0.000000
max        80.000000       1.00000       1.000000      95.690000       9.000000         300.000000       1.000000
```

Como informação adicional, é apresentada uma descrição fornecida pelo próprio fornecedor dos dados, o Kaggle:

```
"The Diabetes prediction dataset is a collection of medical and demographic data from patients,
along with their diabetes status (positive or negative). The data includes features such as age,
gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and
blood glucose level. This dataset can be used to build machine learning models to predict diabetes
in patients based on their medical history and demographic information. This can be useful for
healthcare professionals in identifying patients who may be at risk of developing diabetes and in
developing personalized treatment plans. Additionally, the dataset can be used by researchers to
explore the relationships between various medical and demographic factors and the likelihood of
developing diabetes."
```

Esta descrição oferece um contexto valioso sobre os dados, destacando seu potencial para desenvolvimento de modelos de aprendizado de máquina e sua relevância para profissionais de saúde e pesquisadores na compreensão da diabetes.

3. DAS ANÁLISES PROPOSTAS

- **Análise Geral (Ignorando Gênero):**
  - Realizar todas as análises anteriores considerando o conjunto de dados como um todo, sem levar em conta o gênero dos pacientes.
- **Agrupamento por Gênero:**
  - Inicialmente, agruparemos a quantidade de mulheres/homens com diabetes no conjunto de dados e, a partir daí, expandiremos nossas análises.
    * Quantas dessas mulheres/homens possuem histórico de fumante?
    * Quantas delas têm mais de X anos?
    * Quantas possuem hipertensão?
    * ...
- **Interpretação de Métricas:**
  - Vamos compreender o significado (e de que forma foram estimadas e seu impacto na definição que trás a consequencia principal: ter ou não diabetes) das seguintes métricas: BMI (Índice de Massa Corporal), HbA1c_level (Nível de Hemoglobina A1c), heart_disease (Doença Cardíaca) e o que constitui um nível alto de blood_glucose_level (Nível Elevado de Glicose no Sangue).
  - Identificar quantas pessoas (mulheres/homens) apresentam essas métricas elevadas ou baixas demais.
- **Implementação de Algoritmos de Agrupamento:**
  - Considerar a aplicação de algoritmos de agrupamento, como o Kmeans, para identificar padrões nos dados que possam não ser imediatamente visíveis.

- **Outras Ideias:**
  - Explorar correlações entre as diferentes variáveis do conjunto de dados.
  - Analisar a progressão da diabetes ao longo do tempo para entender melhor a trajetória dos pacientes.
  - Investigar a relação entre o tratamento prescrito e os resultados obtidos pelos pacientes.
  - Considerar a inclusão de dados demográficos adicionais, como etnia e localização geográfica, para uma análise mais abrangente.
  - Avaliar a eficácia de diferentes abordagens de tratamento com base nos resultados obtidos pelos pacientes.

Essas são algumas sugestões para expandir e aprofundar a análise dos dados sobre diabetes.

4. Apresentação:

- Video Lendo/explicando os pontos acima.

5. Referências

- Kaggle. **Conjunto de dados de previsão de diabetes**, 2024. Disponível em: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset. Acesso em: 06 de abril de 2024.

6. Anexos

- Repositório da diciplina/projeto no Github.
- Vídeo - Explicando base de dados escolida e análises a serem realizadas.