**Bahria University,**

**Karachi Campus**

**A logo with text on it

Description automatically generated**

**COURSE:**

**DATA MINING**

**Term: Spring 2024**

**Class: BSE- 6(B)**

**Submitted By:**

**NAME: AIMAN ZIA SATTI**

**ENROLLNMENT: 02-131212-028**

**Submitted To:**

**DR. HINA SHAKIR / ENGR.HAMZA**

**Signed Remarks:**

**INDEX**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SNO** | **DATE** | **LAB NO** | **LAB OBJECTIVE** | **SIGN** |
| **01** | **14-2-24** | **01** | Gui In Python Using Google Colab & Data Mining Libraries |  |
| **02** | **21-2-24** | **02** | Exploratory Data Analysis Using Python Best EDA Libraries |  |
| **03** |  |  |  |  |
| **04** |  |  |  |  |
| **05** |  |  |  |  |
| **06** |  |  |  |  |
| **07** |  |  |  |  |
| **08** |  |  |  |  |
| **09** |  |  |  |  |
| **10** |  |  |  |  |
| **11** |  |  |  |  |
| **12** |  |  |  |  |
| **13** |  |  |  |  |
| **14** |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**Bahria University,**

**Karachi Campus**

**A logo with text on it

Description automatically generated**

**LAB NO. 02**

**LIST OF TASKS**

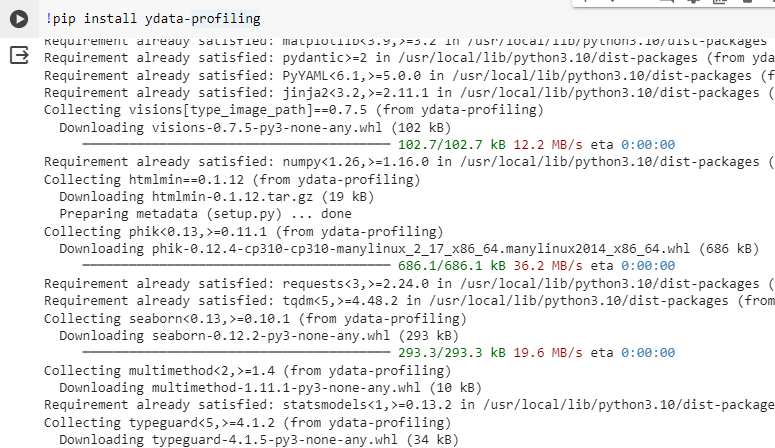
|  |  |
| --- | --- |
| TASK NO | OBJECTIVE |
| 01 | * 1. Perform a basic data profiling to understand the structure of the dataset, including the number of rows, columns, and data types.   2. Identify the target variable and the predictor variables.   3. Compute summary statistics (mean, median, standard deviation, etc.) for each numerical variable.   4. Identify missing values and their distribution across variables. |
| 02 | * 1. Perform a basic data profiling to understand the structure of the dataset, including the number of rows, columns, and data types.   2. Identify the target variable and the predictor variables.   3. Compute summary statistics (mean, median, standard deviation, etc.) for each numerical variable.   4. Identify missing values and their distribution across variables. |
| 03 | * 1. Identify potential outliers in the dataset using appropriate techniques, such as box plots, scatter plots.   2. Visualize the distribution of each numerical variable to identify any extreme values.   3. Discuss the potential impact of outliers on the analysis and modeling process. |
| 04 | * 1. Identify potential outliers in the dataset using appropriate techniques, such as box plots, scatter plots.   2. Visualize the distribution of each numerical variable to identify any extreme values.   3. Discuss the potential impact of outliers on the analysis and modeling process. |

**Submitted On:**

***22nd Feb, 2024***

**(Date: DD/MM/YY)**

**TASK # 1:**



A screenshot of a computer

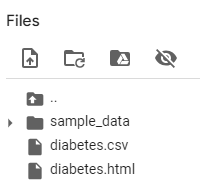
Description automatically generated

A screenshot of a computer

Description automatically generated

A green and white text

Description automatically generated



Perform a basic data profiling to understand the structure of the dataset, including the number of rows, columns, and data types.

A screenshot of a computer

Description automatically generated



A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

Compute summary statistics (mean, median, standard deviation, etc.) for each numerical variable.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Identify missing values and their distribution across variables.

A graph with blue lines

Description automatically generated with medium confidence

**TASK # 2:**

Calculate pairwise correlation coefficients between all numerical variables.



A screenshot of a computer

Description automatically generated

Create a correlation matrix and visualize it using a heatmap.

A screenshot of a computer

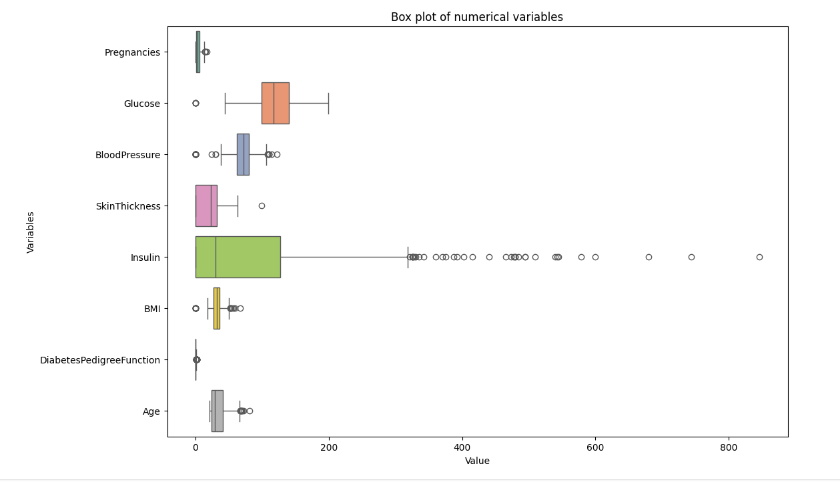
Description automatically generated

Identify highly correlated variables and discuss their potential impact on model performance.



**TASK 3:**

Identify potential outliers in the dataset using appropriate techniques, such as box plots



Visualize the distribution of each numerical variable to identify any extreme values.

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

numerical\_columns = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age']

for column in numerical\_columns:

    plt.figure(figsize=(8, 6))

    sns.histplot(df1[column], kde=True, bins=20, color='skyblue')

    plt.title(f'Distribution of {column}')

    plt.xlabel(column)

    plt.ylabel('Frequency')

    plt.show()

plt.figure(figsize=(12, 8))

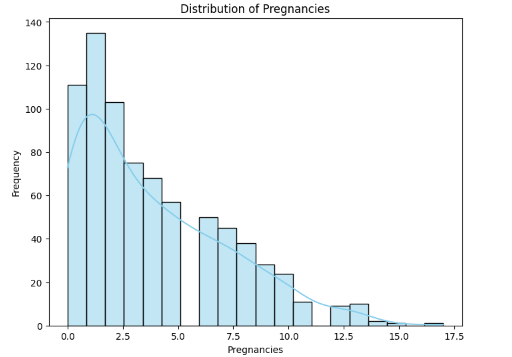
sns.boxplot(data=df1[numerical\_columns], orient='h', palette='Set2')

plt.title('Box plot of numerical variables')

plt.xlabel('Value')

plt.ylabel('Variables')

plt.show()



A diagram of a distribution of glucose

Description automatically generated

A diagram of a distribution of blood pressure

Description automatically generated

A graph of skin thickness

Description automatically generated

A graph of a distribution of insulin

Description automatically generated

A diagram of a distribution of bmi

Description automatically generated

A graph of a normal distribution of diabetes

Description automatically generated

A graph of a distribution of age

Description automatically generated

**Discuss the potential impact of outliers on the analysis and modeling process.**

Outliers can skew statistical measures such as the mean and standard deviation, potentially leading to biased estimates.

They can influence the results of certain statistical tests and machine learning algorithms, leading to incorrect conclusions or poorly performing models.

Depending on the context, outliers may represent valid data points or errors in measurement. It's crucial to understand the nature of outliers before deciding whether to remove or transform them.

Outliers may also indicate interesting or important phenomena in the data that warrant further investigation.

**TASK 4:**

* 1. Analyze the distribution of the target variable (diabetes or non-diabetes).
  2. Visualize the target variable distribution using a histogram or a bar chart.
  3. Identify any potential imbalance in the target variable and discuss its impact on model performance.

import matplotlib.pyplot as plt

import seaborn as sns

plt.figure(figsize=(6, 4))

sns.countplot(data=df1, x='Outcome', palette='Set2')

plt.title('Distribution of Outcome (Diabetes vs Non-Diabetes)')

plt.xlabel('Outcome')

plt.ylabel('Count')

plt.show()

outcome\_counts = df1['Outcome'].value\_counts()

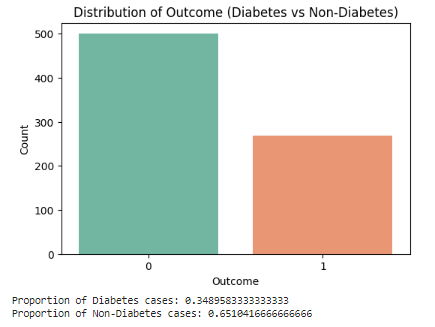
total\_samples = len(df1)

proportion\_diabetes = outcome\_counts[1] / total\_samples

proportion\_non\_diabetes = outcome\_counts[0] / total\_samples

print("Proportion of Diabetes cases:", proportion\_diabetes)

print("Proportion of Non-Diabetes cases:", proportion\_non\_diabetes)



**Imbalanced classes** can lead to **biased models**, where the model may become overly sensitive to the majority class and perform poorly on the minority class.

In the case of medical diagnosis like diabetes detection, misclassification of positive cases (diabetes) can have significant consequences, potentially leading to missed diagnoses or delayed treatments.

Algorithms trained on imbalanced data may prioritize accuracy over other performance metrics, resulting in poor sensitivity or specificity, which are crucial for medical applications.

Techniques such as resampling (e.g., oversampling minority class, undersampling majority class), using different evaluation metrics (e.g., precision, recall, F1-score), or employing algorithms designed to handle imbalanced data (e.g., ensemble methods, cost-sensitive learning) can help mitigate the impact of class imbalance and improve model performance.