Bahria University

Karachi Campus

COURSE: CSL-460

DATA MINING LAB

TERM: SPRING 2024, CLASS: BSE- 6(B)

SUBMITTED BY:

**Ahsan Sajjad**

**79309**

**(Name) (Enrollment No.)**

SUBMITTED TO:

Engr. HAMZA/Engr. HINA SHAKIR

SIGNED REMARKS: SCORE:

INDEX

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SNO | DATE | LAB NO | LAB BJECTIVE | REMARKS/SIGN |
| 01 | 14-02-24 | 01 | GUI IN PYTHON USING GOOGLE COLAB & Data Mining Libraries |  |
| 02 | 21-02-24 | 02 | Exploratory data analysis using python best EDA libraries |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Bahria University,

Karachi Campus

A logo with text on it

Description automatically generated

LAB EXPERIMENT 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. |
| **02** | 1. Calculate pairwise correlation coefficients between all numerical variables.  2. Create a correlation matrix and visualize it using a heatmap. |
| **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. Analyze the distribution of the target variable (diabetes or non-diabetes).  2. Visualize the target variable distribution using a histogram or a bar chart. |

Submitted On:

25-02-2024

(Date: DD/MM/YYYY)

**Task No. 01**: Data Profiling:

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

Identify the target variable and the predictor variables.

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

Identify missing values and their distribution across variables.

**Solution:**

import pandas as pd

import pandas as pd

import sweetviz as sv

from autoviz.AutoViz\_Class import AutoViz\_Class

import ydata\_profiling as ydp

diabetes\_df = pd.read\_csv('/content/diabetes.csv')

print("Dataset structure:")

print("Number of rows:", diabetes\_df.shape[0])

print("Number of columns:", diabetes\_df.shape[1])

print("\nData types:")

print(diabetes\_df.dtypes)

target\_variable = 'diabetes\_status'

predictor\_variables = [col for col in diabetes\_df.columns if col != target\_variable]

profile.to\_file("ydata\_profiling\_report.html")

from autoviz.AutoViz\_Class import AutoViz\_Class

# Instantiate the AutoViz class

AV = AutoViz\_Class()

# Load and visualize the dataset

df = AV.AutoViz('diabetes.csv')

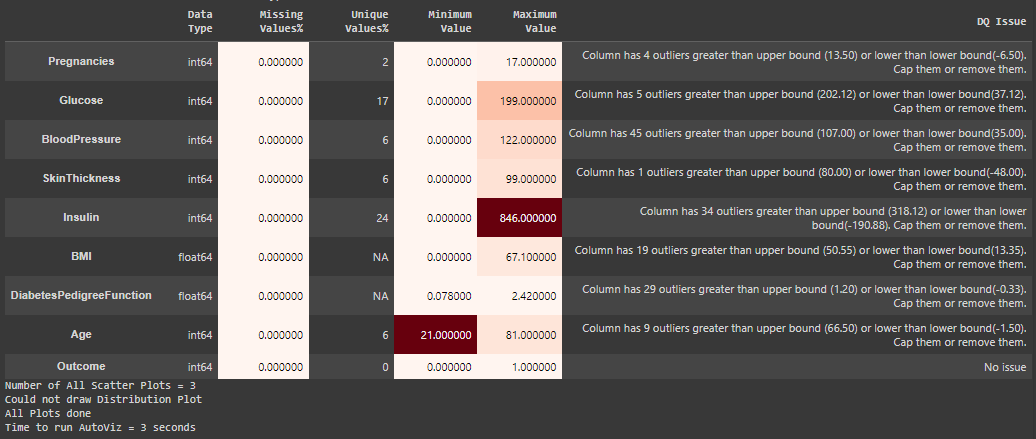
**Output:**

**A black screen with white text

Description automatically generated**

**A screen shot of a computer

Description automatically generated**

****

**A screenshot of a computer screen

Description automatically generated**

**A graph of blue rectangular shapes

Description automatically generated with medium confidence**

**Task No. 02**: **Feature Correlation:**

Calculate pairwise correlation coefficients between all numerical variables.

Create a correlation matrix and visualize it using a heatmap.

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

**Solution:**

import dtale

import pandas as pd

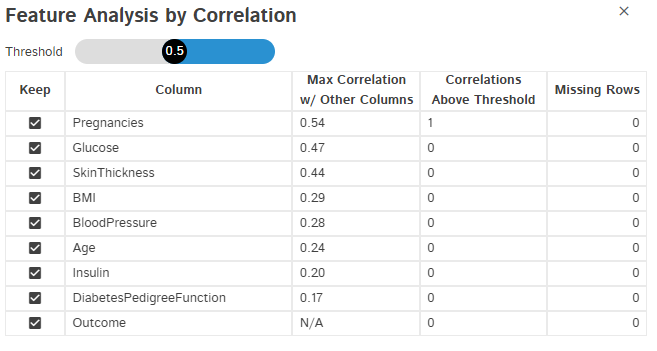
import dtale.app as dtale\_app

dtale\_app.USE\_COLAB = True

df = pd.read\_csv('/content/diabetes.csv')

dtale.show

**Output:**

****

**A screenshot of a computer

Description automatically generated**

**A graph with blue lines

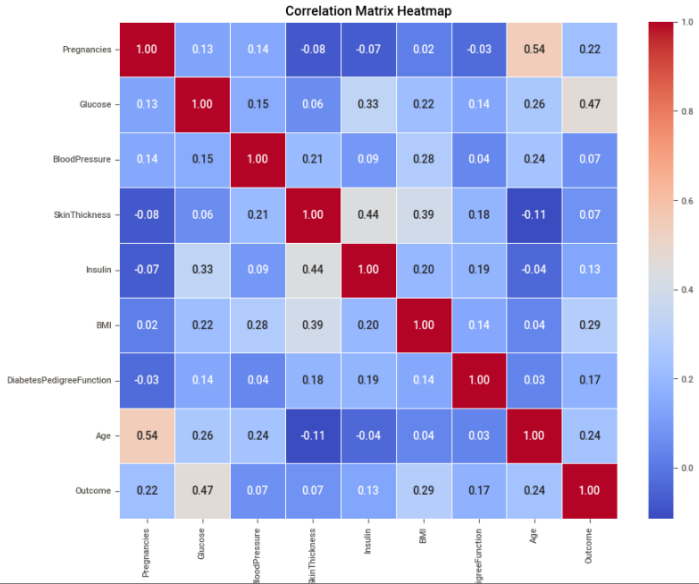
Description automatically generated**

**Heatmap**

**A screenshot of a computer screen

Description automatically generatedA screenshot of a chart

Description automatically generated**

****

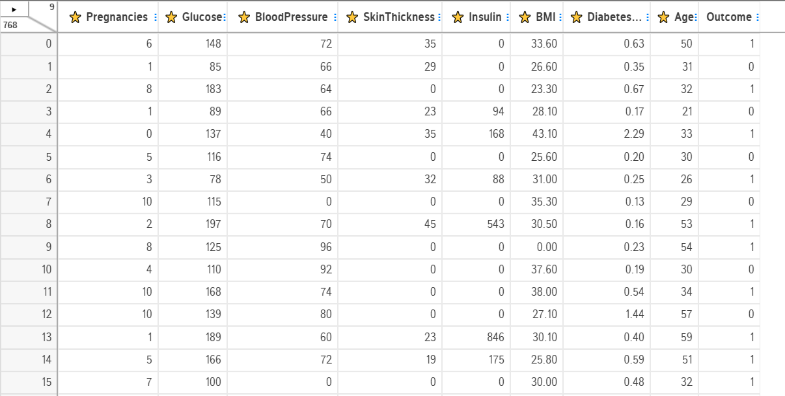
**Task No. 03**: **Outlier Detection:**

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

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

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

**Solution: Output:**

****import dtale

import pandas as pd

import dtale.app as dtale\_app

dtale\_app.USE\_COLAB = True

df = pd.read\_csv('/content/diabetes.csv')

dtale.show(df)

**A screen shot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

Outliers can affect the final outcome majorly. A model will take account of these values too that are wrong eventually the outcome will not be what we require.

**Task No. 04**: **Target Variable Analysis:**

Analyze the distribution of the target variable (diabetes or non-diabetes).

Visualize the target variable distribution using a histogram or a bar chart.

Identify any potential imbalance in the target variable and discuss its impact on model performance.

**Solution:**

import pandas as pd

import sweetviz as sv

df = pd.read\_csv('/content/diabetes.csv')

report = sv.analyze(df, target\_feat="Pregnancies")

report.show\_html('diabetes\_sweetviz\_pregnancies\_report.html')

**A screenshot of a computer

Description automatically generatedOutput:**

**A graph with a red line

Description automatically generated**

**A white paper with blue numbers and black text

Description automatically generated**