AI- BASED DIABETES PREDICTION SYSTEM

(PHASE- 5)

1.INTRODUCTION

Diabetes is a chronic disease that directly affects the pancreas, and the body is incapable of producing insulin [22]. Insulin is mainly responsible for maintaining the blood glucose level. Many factors, such as excessive body weight, physical inactivity, high blood pressure, and abnormal cholesterol level, can cause a person get affected by diabetes [23].

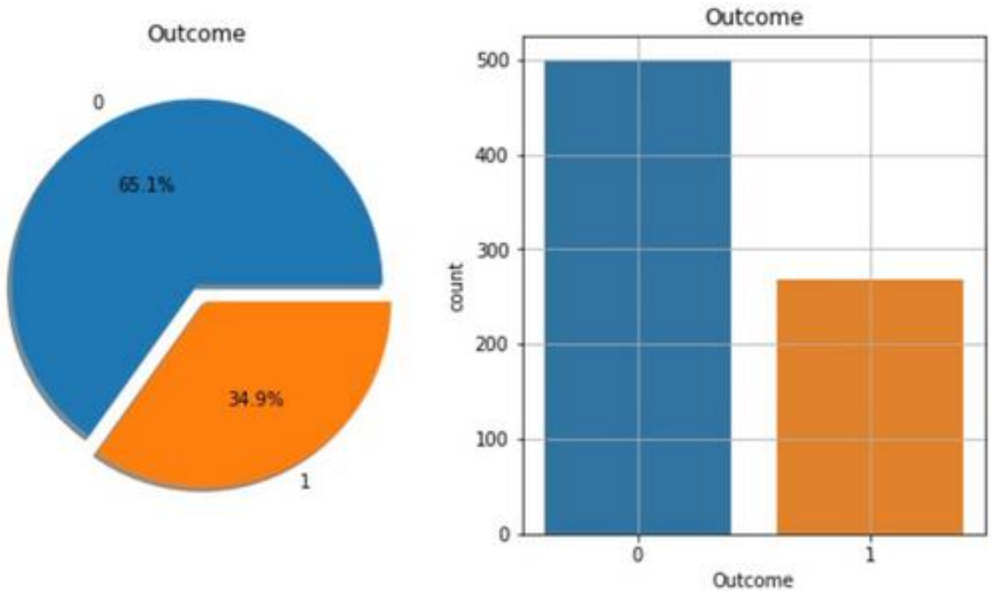
It can cause many complications, but an increase in urination is one of the most common ones [24]. It can damage the skin, nerves, and eyes, and if not treated early, diabetes can cause kidney failure and diabetic retinopathy ocular disease. According to IDF (International Diabetes Federation) statistics, 537 million people had diabetes around the world in 2021 [1]. In Bangladesh, approximately 7.10 million people had suffered from this disease, according to 2019 statistics [2].

2.DATASET

The Pima Indian dataset is an open-source dataset [6] that is publicly available for machine learning classification, which has been used in this work along with a private dataset. It contains 768 patients' data, and 268 of them have developed diabetes.

Figure 2 shows the ratio of people having diabetes in the Pima Indian dataset. Table 1 demonstrates the eight features of the open-source Piman Indian dataset.

RTML private dataset: A significant contribution of this work is to present a private dataset from Rownak Textile Mills Ltd, Dhaka, Bangladesh, referred to as RTML, to the scientific community. Following a brief explanation of the study to the female volunteers, they voluntarily agreed to participate in the study. This dataset comprises six features, that is, pregnancy, glucose, blood pressure, skin thickness, BMI, age, and outcome of diabetes from 203 female individuals aged between 18 and 77. In this work, blood glucose was measured by the GlucoLeader Enhance blood sugar meter.



3. Dataset preprocessing

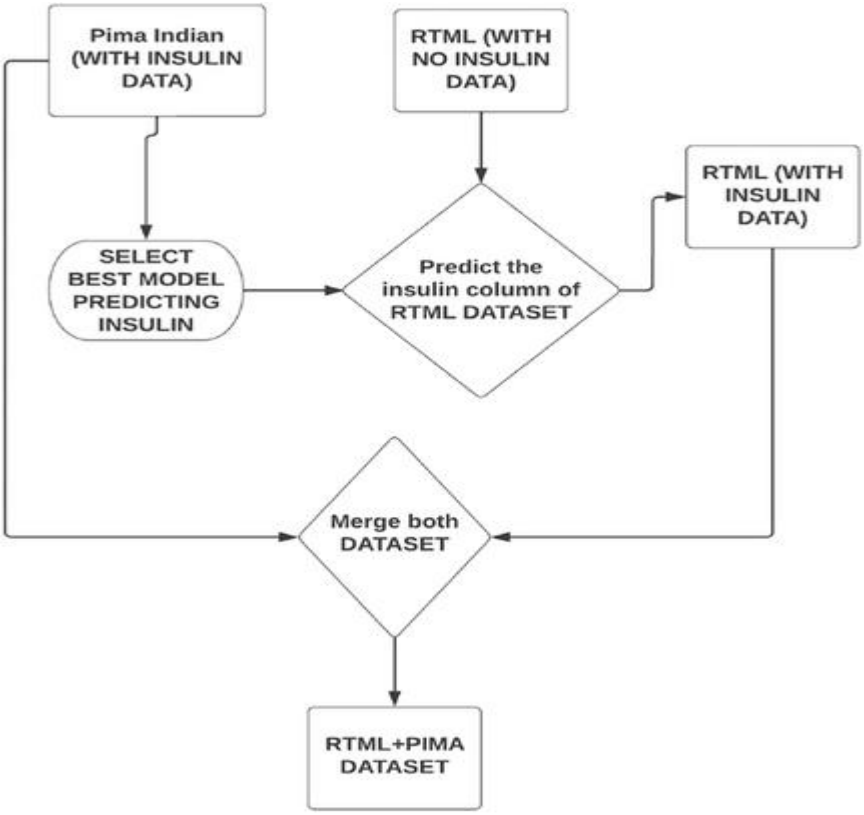
In the merged dataset, we discovered a few exceptional zero values. For example, skin thickness and Body Mass Index (BMI) cannot be zero. The zero value has been replaced by its corresponding mean value.

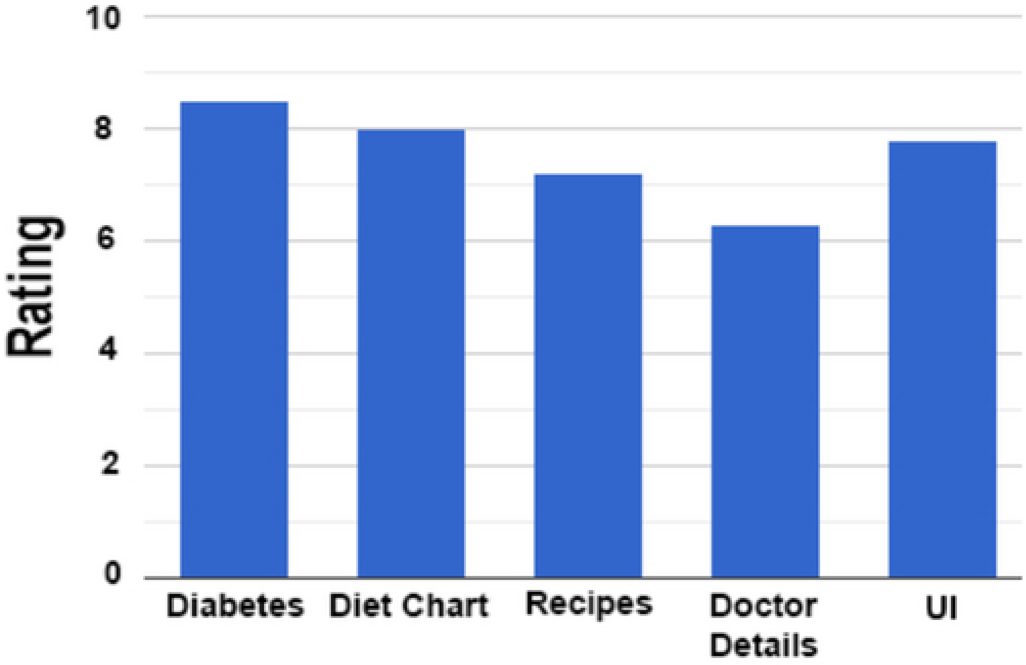
The training and test dataset has been separated using the holdout validation technique, where 80% is the training data and 20% is the test data.

Mutual Information: Mutual information attempts to measure the interdependence of variables. It produces information gain, and its higher values indicate greater dependency [8].

Semi-supervised learning: A combined dataset has been used in this work by incorporating the open-source Pima Indian and private RTML datasets.

According to Table 2, the RTML dataset does not contain the insulin feature, which is predicted using a semi-supervised approach.


Pima Indian (WITH INSULIN DATA)
RTML (WITH NO INSULIN DATA)
RTML (WITH INSULIN DATA)
SELECT BEST MODEL PREDICTING INSULIN
Predict the insulin column of RTML DATASET
Merge both DATASET
RTML+PIMA DATASET


Rating
10
8
6
4
2
0
Diabetes
Diet Chart
Recipes
Doctor Details
UI

4.PROGRAM

INPUT

```
#Installation of required libraries
import numpy as np
import pandas as pd
import statsmodels.api as sm
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import scale, StandardScaler
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_sco
re
from sklearn.metrics import confusion_matrix, accuracy_score, mean_squared_error,
r2_score, roc_auc_score, roc_curve, classification_report
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
from lightgbm import LGBMClassifier
from sklearn.model_selection import KFold
import warnings
warnings.simplefilter(action = "ignore")
```

INPUT

```
#Reading the dataset
df = pd.read_csv("../input/pima-indians-diabetes-database/diabetes.csv")
```

INPUT

```
# The first 5 observation units of the data set were accessed.
df.head()
```

INPUT

```
# The size of the data set was examined. It consists of 768 observation units and
9 variables.
df.shape
```

OUTPUT

```
(768, 9)
```

INPUT

```
# The distribution of the Outcome variable was examined.
df["Outcome"].value_counts()*100/len(df)
```

OUTPUT

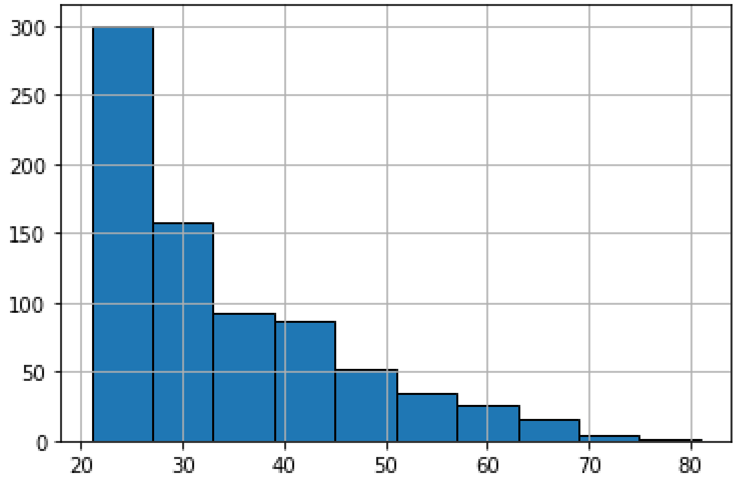
```
0    65.104167
1    34.895833
Name: Outcome, dtype: float64
```

OUTPUT

```
# The classes of the outcome variable were examined.
df.Outcome.value_counts()
0    500
1    268
Name: Outcome, dtype: int64
```

INPUT

```
# The histagram of the Age variable was reached.
df["Age"].hist(edgecolor = "black")
```



INPUT

```
print("Max Age: " + str(df["Age"].max()) + " Min Age: " + str(df["Age"].min()))
```

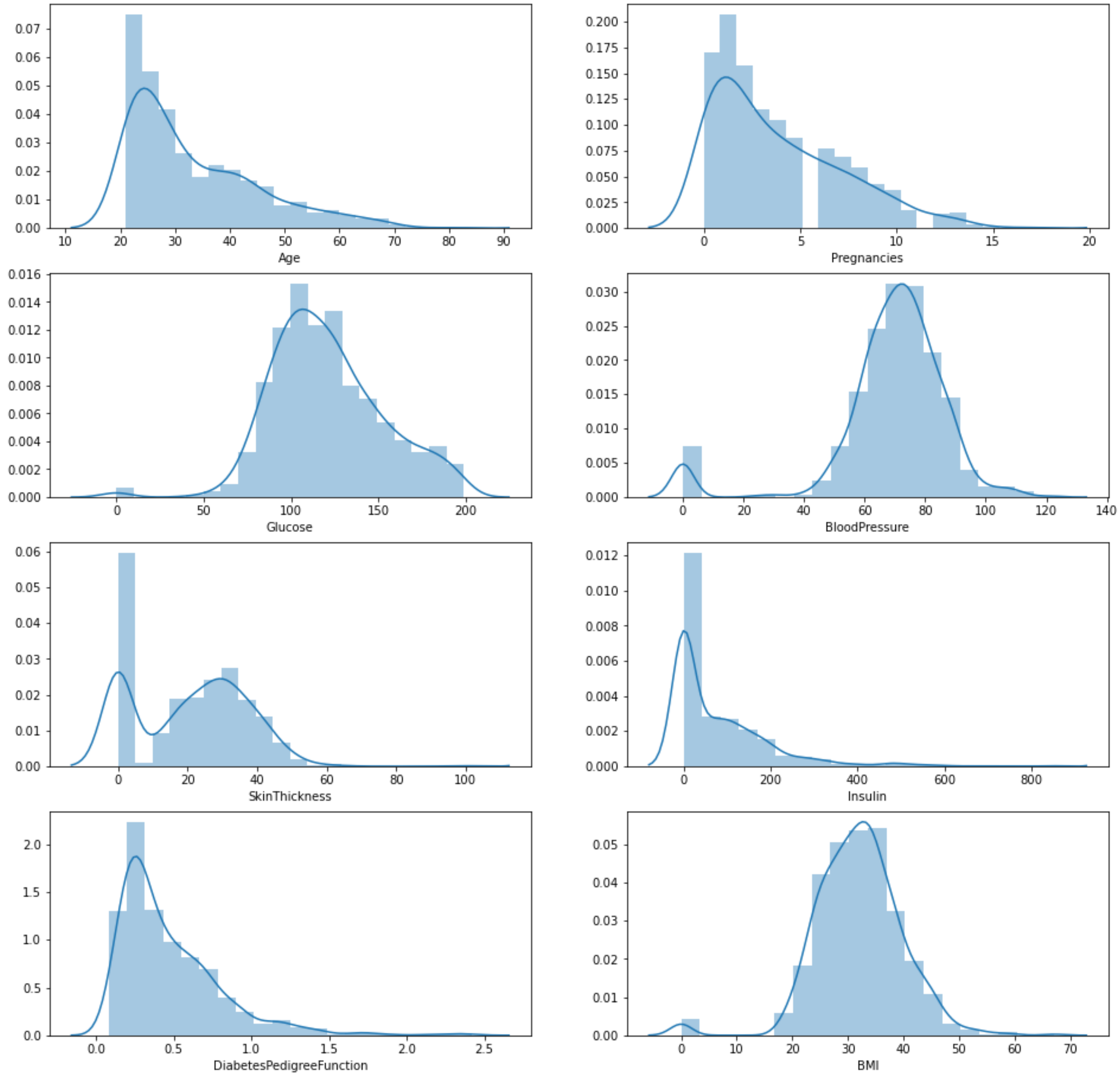
```
Max Age: 81 Min Age: 21
```

INPUT

```
# Histogram and density graphs of all variables were accessed.
fig, ax = plt.subplots(4,2, figsize=(16,16))
sns.distplot(df.Age, bins = 20, ax=ax[0,0])
sns.distplot(df.Pregnancies, bins = 20, ax=ax[0,1])
sns.distplot(df.Glucose, bins = 20, ax=ax[1,0])
sns.distplot(df.BloodPressure, bins = 20, ax=ax[1,1])
sns.distplot(df.SkinThickness, bins = 20, ax=ax[2,0])
sns.distplot(df.Insulin, bins = 20, ax=ax[2,1])
sns.distplot(df.DiabetesPedigreeFunction, bins = 20, ax=ax[3,0])
sns.distplot(df.BMI, bins = 20, ax=ax[3,1])
```

OUTPUT

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f77b83d5950>
```



5.CONCLUSION

Diabetes can be a reason for reducing life expectancy and quality. Predicting this chronic disorder earlier can reduce the risk and complications of many diseases in the long run. In this paper, an automatic diabetes prediction system using various machine learning approaches has been proposed.

The open-source Pima Indian and a private dataset of female Bangladeshi patients have been used in this work. SMOTE and ADASYN preprocessing techniques have been applied to handle the issue of imbalanced class problems.

This research paper reported different performance metrics, that is, precision, recall, accuracy, F1 score, and AUC for various machine learning and ensemble techniques. The XGBoost classifier achieved the best performance with 81% accuracy and an F1 score and AUC of 0.81 and 0.84, respectively, with the ADASYN approach.

Next, the domain adaptation technique has been applied to demonstrate the versatility of the proposed prediction system. Finally, the best-performed XGBoost framework has been deployed into a website and smartphone application to predict diabetes instantly. There are some future scopes of this work, for example, we recommend getting additional private data with a larger cohort of patients to get better results. Another extension of this work is combining machine learning models with fuzzy logic techniques and applying optimization approaches.