Diabetes Prediction System

Phase 4 project submission

Project name: Diabetes Prediction System

Phase 4 :Development part 2

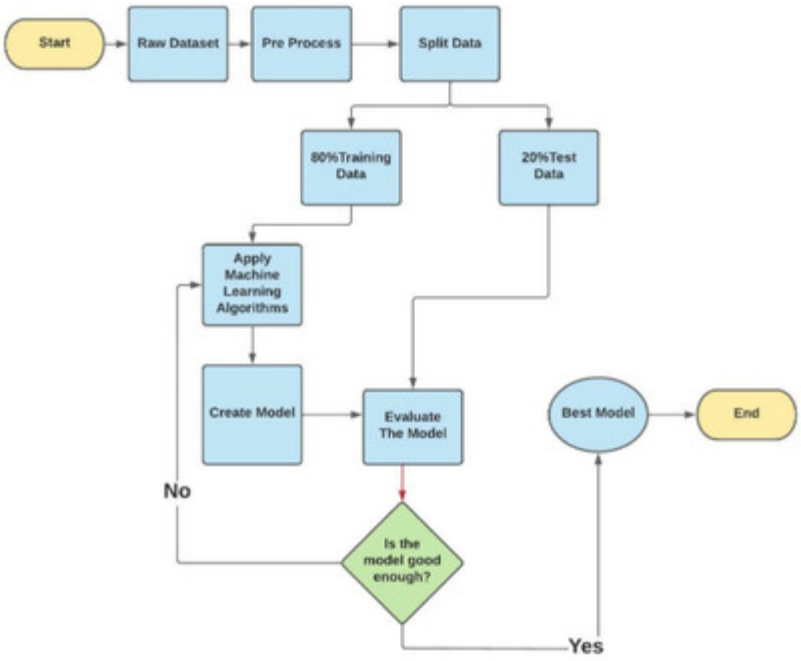
Table of contents

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].

PROPOSED SYSTEM

This section describes the working procedures and implementation of various machine learning techniques to design the proposed automatic diabetes prediction system. Figure 1 shows the different stages of this research work. First, the dataset was collected and preprocessed to remove the necessary discrepancies from the dataset, for example, replacing null instances with mean values, dealing with imbalanced class issues etc. Then the dataset was separated into the training set and test set using the holdout validation technique. Next, different classification algorithms were applied to find the best classification algorithm for this dataset. Finally, the best-performed prediction model is deployed into the proposed website and smartphone application framework.



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.

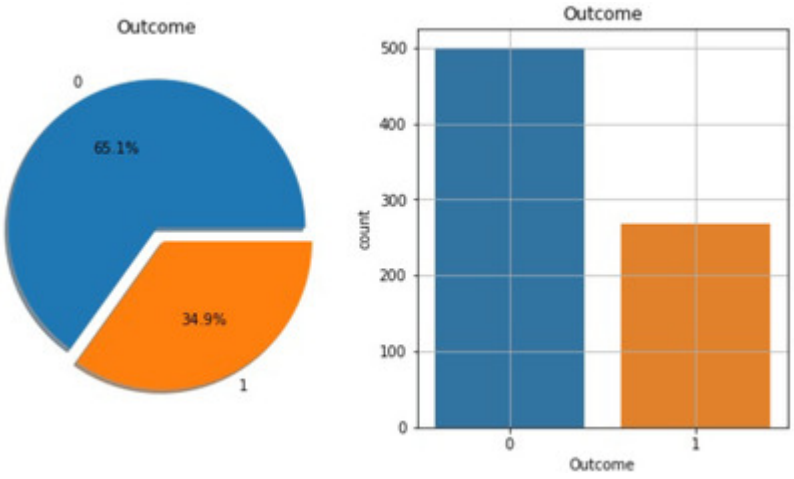


TABLE 1

Features of the Pima Indian Dataset

Pregnancies	Skin thickness	Diabetes pedigree function
Glucose	Insulin	Age
Blood pressure	BMI	

TABLE 2

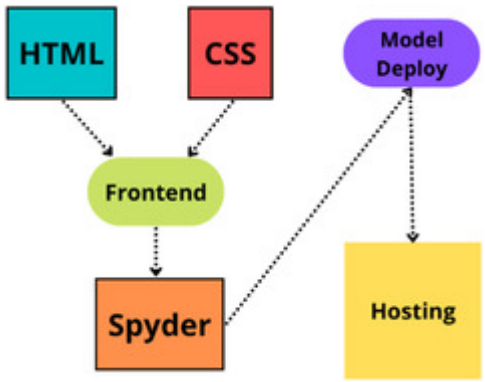
Features of the RTML private dataset

Features	Minimum	Maximum	Average
Pregnancies	0	8	1.61
Glucose (mg/dL)	52.2	274	109.39
Blood pressure (mm Hg)	5.9	115	71.09
Skin thickness (mm)	2.9	23.3	10.78
BMI (kg/m^2)	2.61	41.62	22.69
Age (years)	17	77	27.02

DEPLOYMENT OF THE PREDICTION SYSTEM

The proposed machine learning-based diabetes prediction system has been deployed into a website and smartphone application framework to work instantaneously on real data.

Web application: We have used HTML and CSS for the frontend part of the proposed website. After that, we finalized the machine learning model XGBoost with ADASYN, as it provided the best performance. The model deployment has been done with Spyder, a Python environment platform that works with Anaconda. Figure 5 shows the illustration of the website application development process.



PROGRAM

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier, BaggingClassifier,
GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix, classification_report

import warnings
warnings.filterwarnings('ignore')
data = pd.read_csv("/kaggle/input/pima-indians-diabetes-database/diabetes.csv")
data.head()
print("Shape of data is", data.shape)
print("="*50)

data.info()
data.isnull().sum()
data.duplicated().sum()
data.hist(figsize=(15, 10))
plt.show()
corr = data.corr()
```

```
corr

plt.figure(figsize=(8,5))

sns.heatmap(corr, cmap='YlGnBu', annot=True, vmin=-1, vmax=1)

plt.title('Relation between features and Diabetes')

plt.show()
```

Importing Libraries

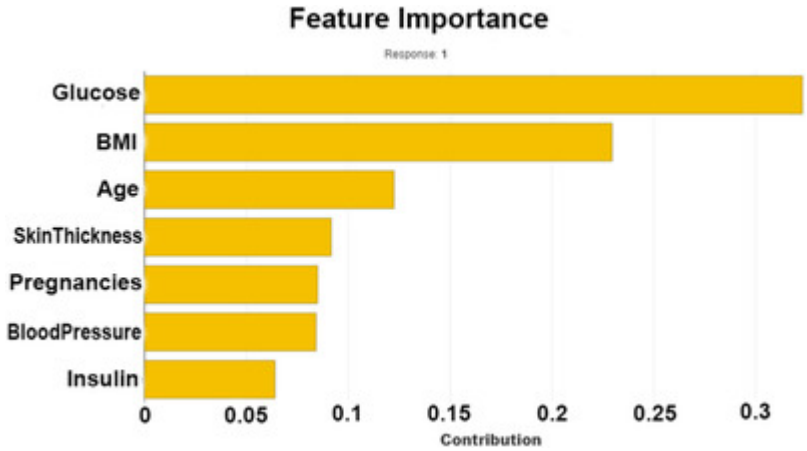
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

sns.set()

from mlxtend.plotting import plot_decision_regions
import missingno as msno
from pandas.plotting import scatter_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import classification_report
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
```

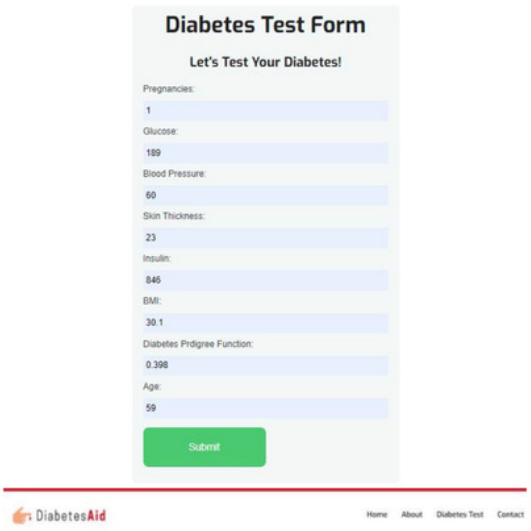
Here we will be reading the dataset which is in the CSV format

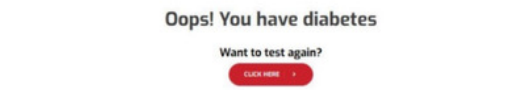
```
diabetes_df = pd.read_csv('diabetes.csv')
diabetes_df.head()
```

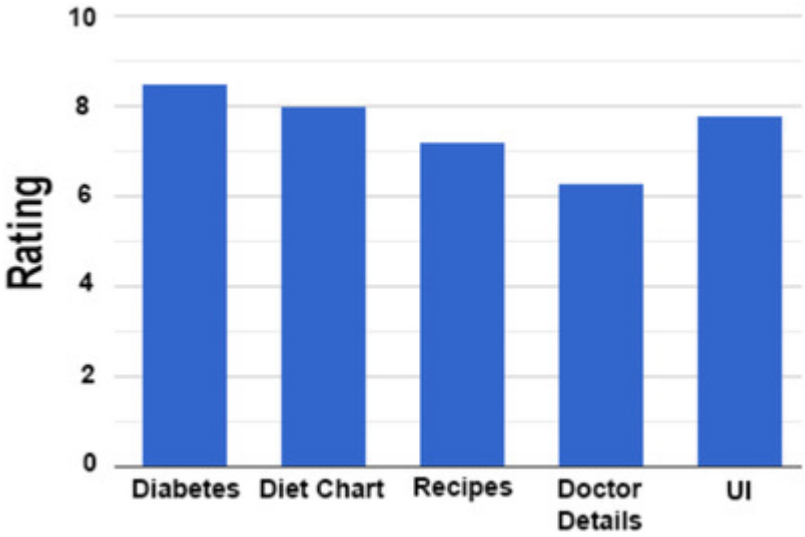
Output:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1









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.