AI-Based Prediction Diabetes System Projected By Machine Learning Algorithm

PHASE 4 DOCUMENTATION

PROJECT: AI-Based Prediction Diabetes System

ABSTRACT:

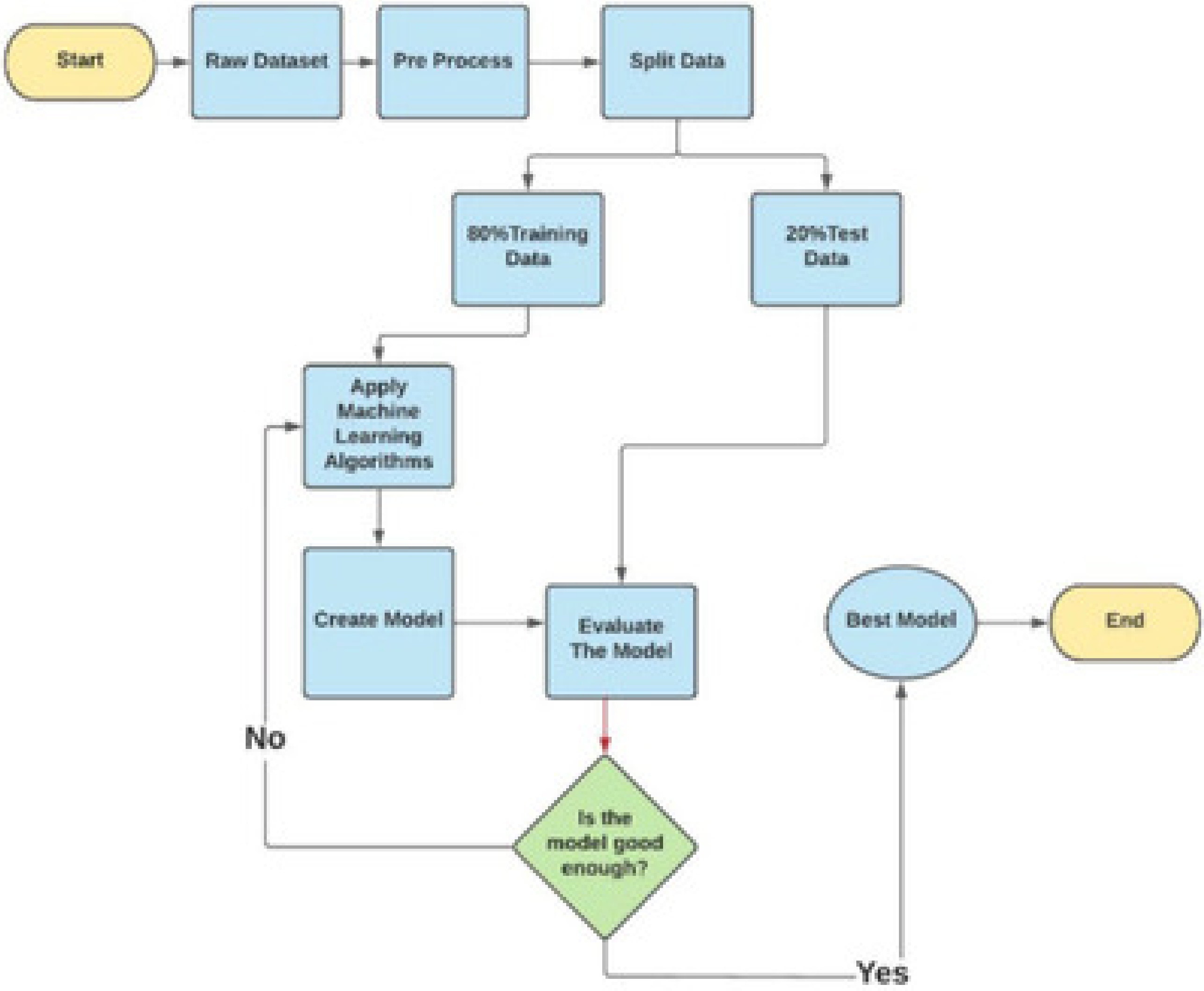
Globally, diabetes affects 537 million people, making it the deadliest and the most common non-communicable disease. Many factors can cause a person to get affected by diabetes, like excessive body weight, abnormal cholesterol level, family history, physical inactivity, bad food habit etc. Increased urination is one of the most common symptoms of this disease. People with diabetes for a long time can get several complications like heart disorder, kidney disease, nerve damage, diabetic retinopathy etc. But its risk can be reduced if it is predicted early. In this paper, an automatic diabetes prediction system has been developed using a private dataset of female patients in Bangladesh and various machine learning techniques. The authors used the Pima Indian diabetes dataset and collected additional samples from 203 individuals from a local textile factory in Bangladesh. Feature selection algorithm mutual information has been applied in this work. A semi-supervised model with extreme gradient boosting has been utilized to predict the insulin features of the private dataset. SMOTE and ADASYN approaches have been employed to manage the class imbalance problem. The authors used machine learning classification methods, that is, decision tree, SVM, Random Forest, Logistic Regression, KNN, and various ensemble techniques, to determine which algorithm produces the best prediction results. After training on and testing all the classification models, the proposed system provided the best result in the XGBoost classifier with the ADASYN approach with 81% accuracy, 0.81 F1 coefficient and AUC of 0.84. Furthermore, the domain adaptation method has been implemented to demonstrate the versatility of the proposed system. The explainable AI approach with LIME and SHAP frameworks is implemented to understand how the model predicts the final results. Finally, a website framework and an Android smartphone application have been developed to input various features and predict diabetes instantaneously.

KEYWORDS:

AdaBoost, android Application, decision tree, diabetes, K-nearest neighbour, random forest, support vector machine

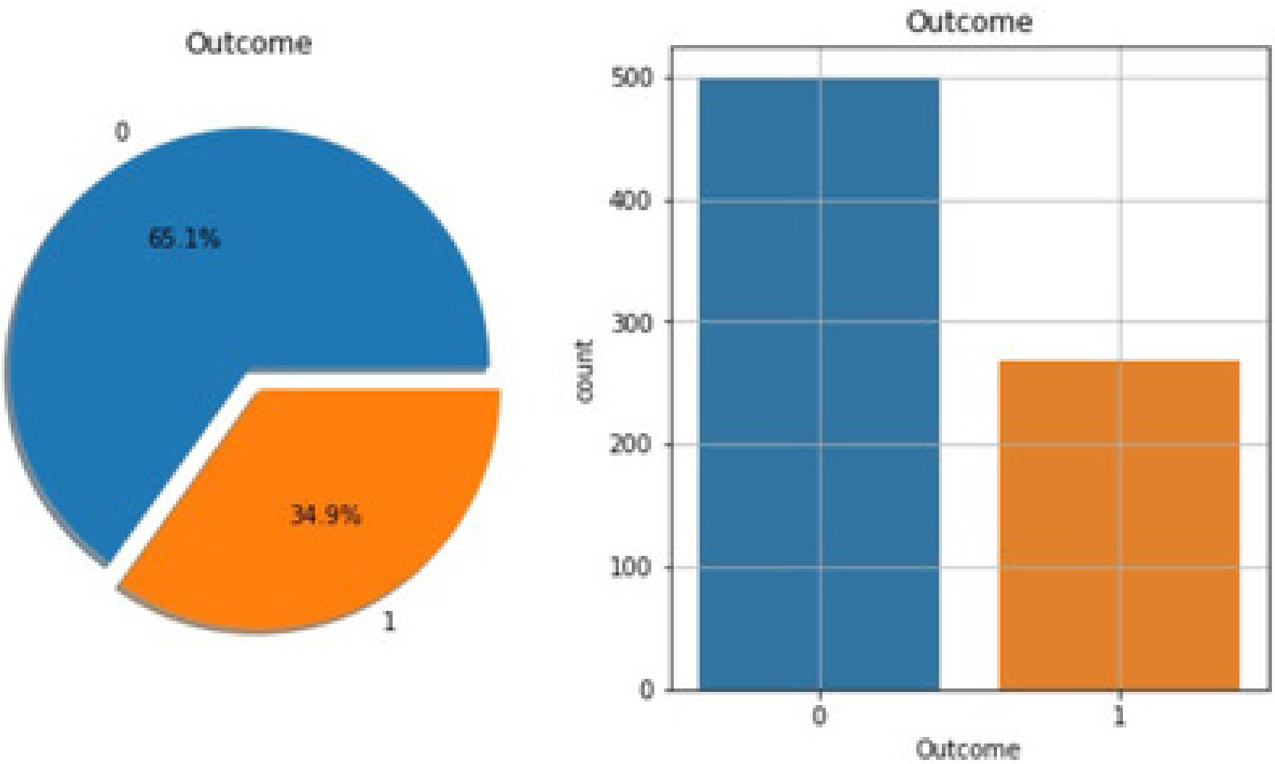
PROPOSED SYSTEM:

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.



Features of the Pima Indian Dataset

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

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. The blood pressure and skin thickness of the participants were obtained by OMRON HEM-7156T and digital LCD body fat caliper machines, respectively

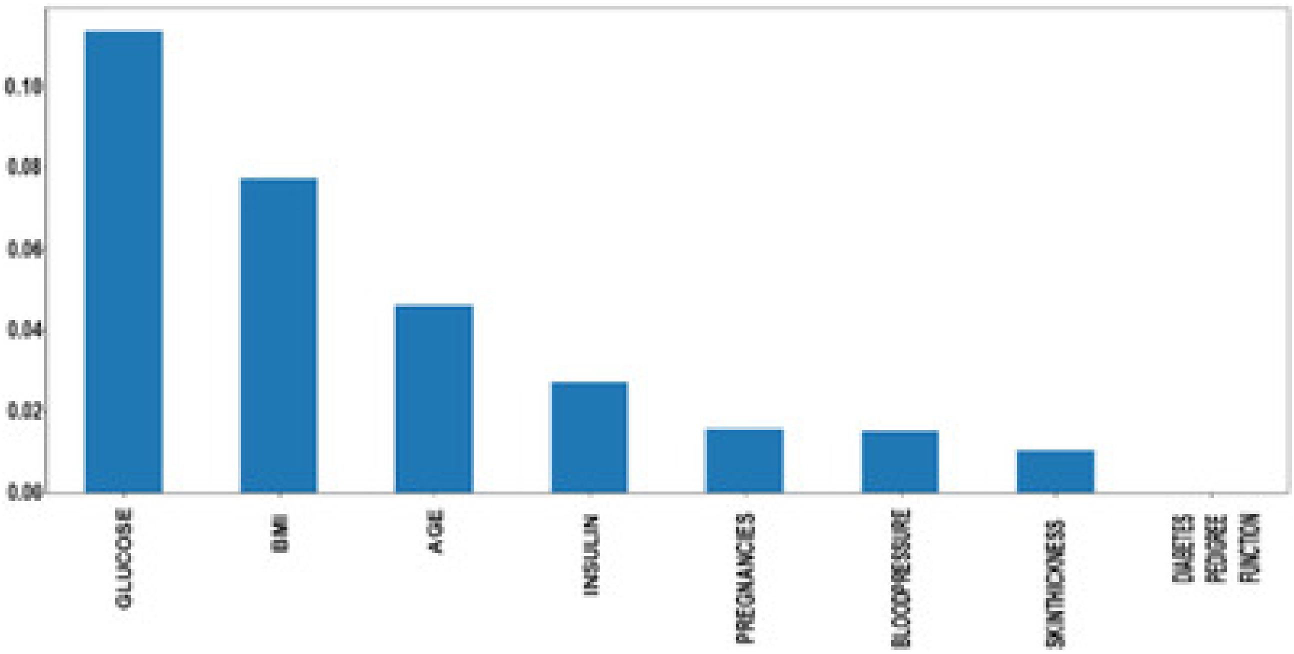
Features of the RTML private dataset

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/m2)	2.61	41.62	22.69
Age (years)	17	77	27.02

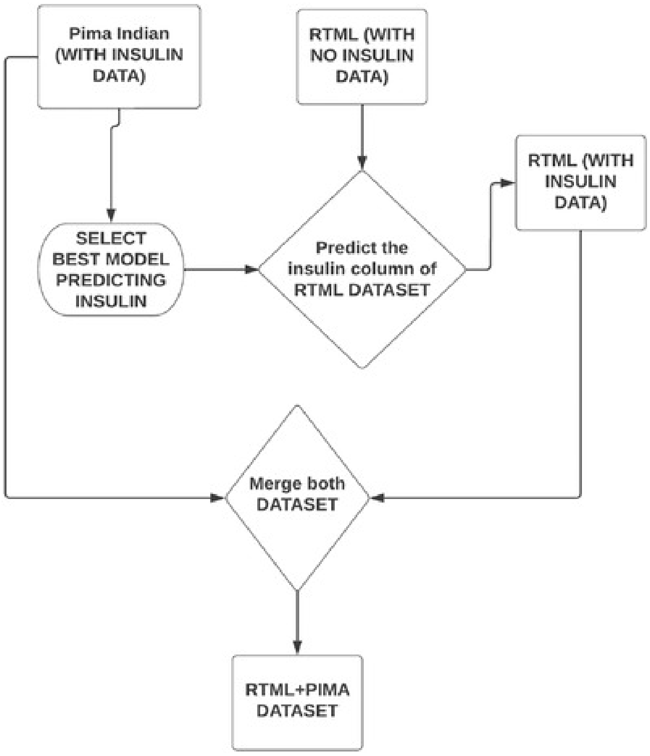
DATASET PREPROCESSING:

n 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. Before merging the collected dataset with the Pima Indian dataset, a model was created using the extreme gradient boosting technique (XGB regressor).


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

MACHINE LEARNING CLASSIFIERS:

In this work, various machine learning and ensemble techniques have been employed to implement the automatic diabetes prediction system, briefly discussed below. GridSearchCV framework has been employed in this research to find the optimal values of different hyperparameters for all the machine learning models to prevent overfitting.

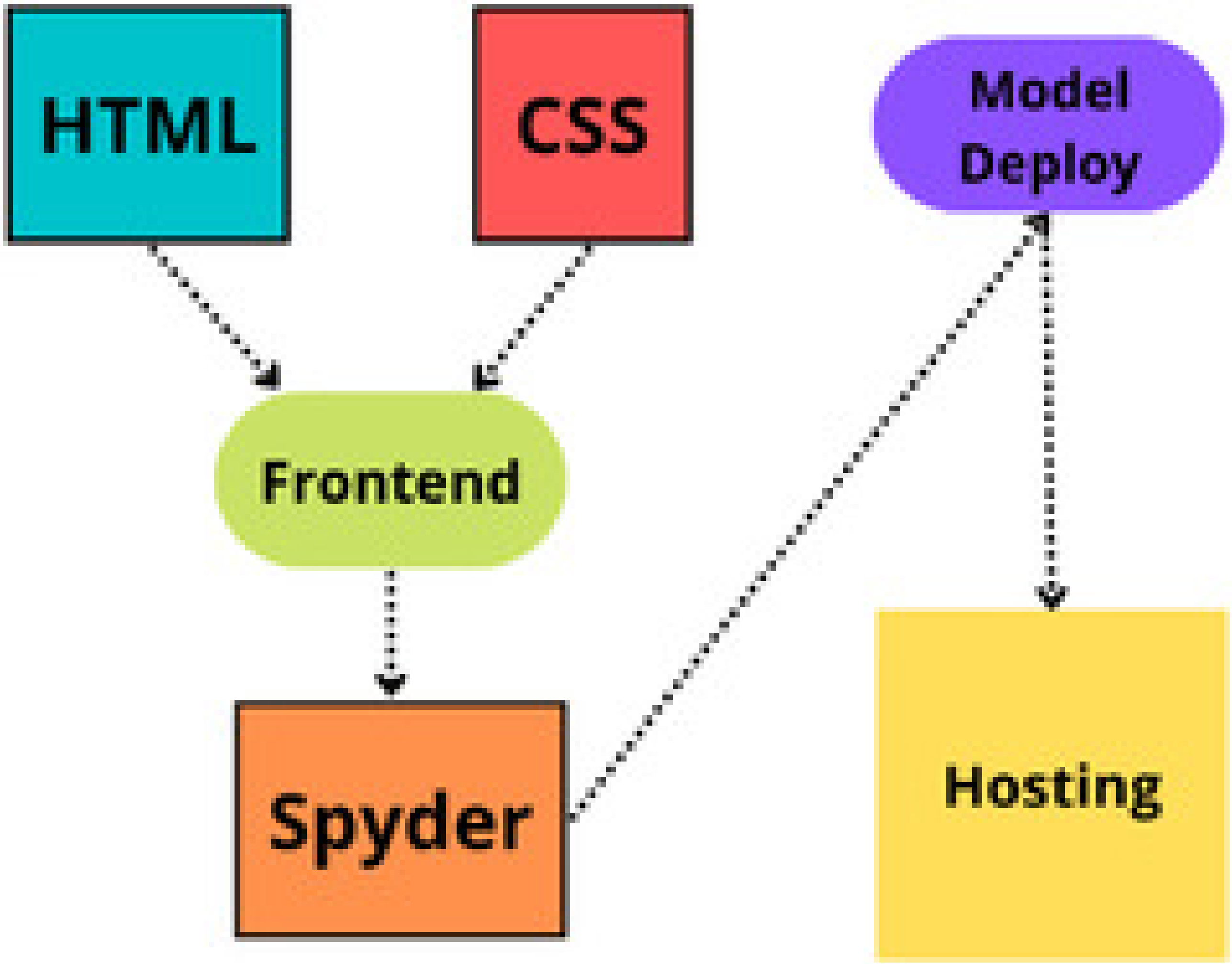
Decision tree: A decision tree represents the learning function provided by a set of rules. The decision tree learning technique performs a method for approximating discrete-valued target functions. Gini or entropy [7] are used to determine information gain, and each node is chosen based on these coefficients, which are expressed as

KNN classifier: A discrete-valued function can be approximated by *K* number of nearest classifiers [8]. To categorize, it creates a plane with the available training points and calculates the distance between the query and trained points. It determines the *K* number of neighbours (depending on the dataset) and classifies them using majority voting. In our research, we used *K* = 5 for the binary classification.

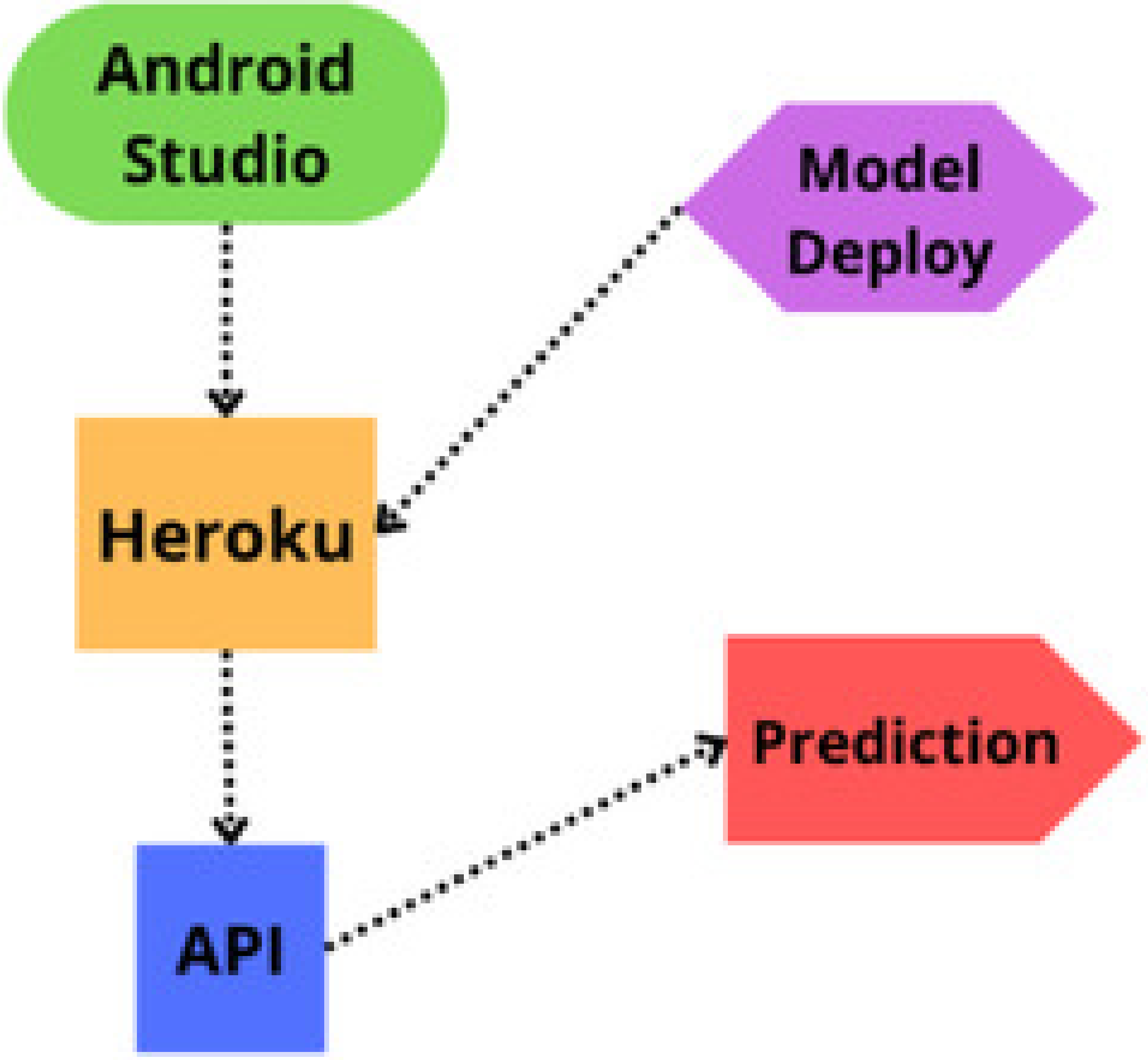
DEVELOPMENT OF THE PREDICTION SYSTEM:

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.



Android smartphone application: To demonstrate the automatic diabetes forecasting system in real time, we also designed an Android smartphone application to test its performance. Android Studio is used for the frontend part of this application. We employed Java as the necessary coding language. After that, the model has been implemented in Android Studio using the pickle package. While developing the API, we used Heroku to host our model on the corresponding hosting server

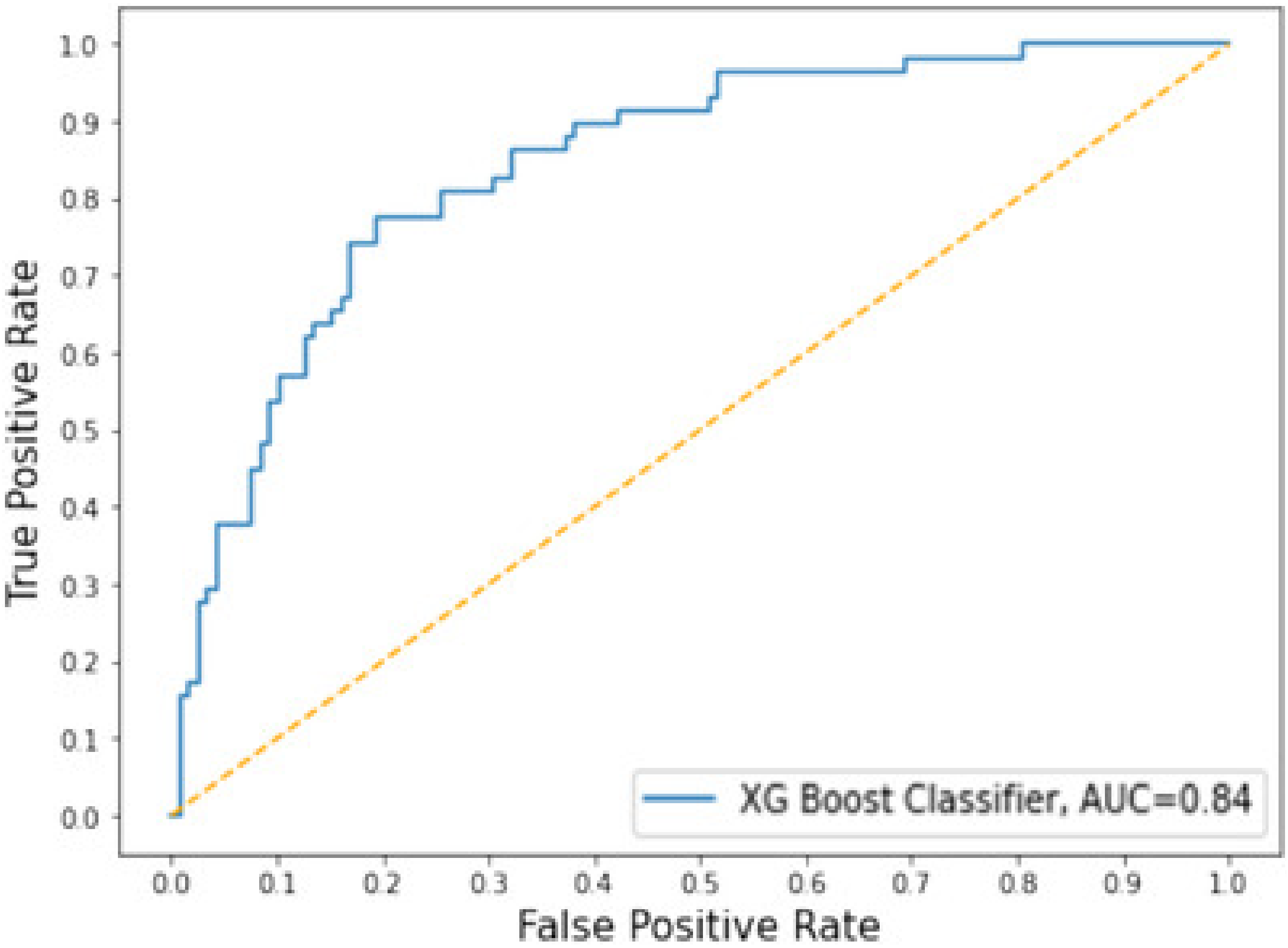


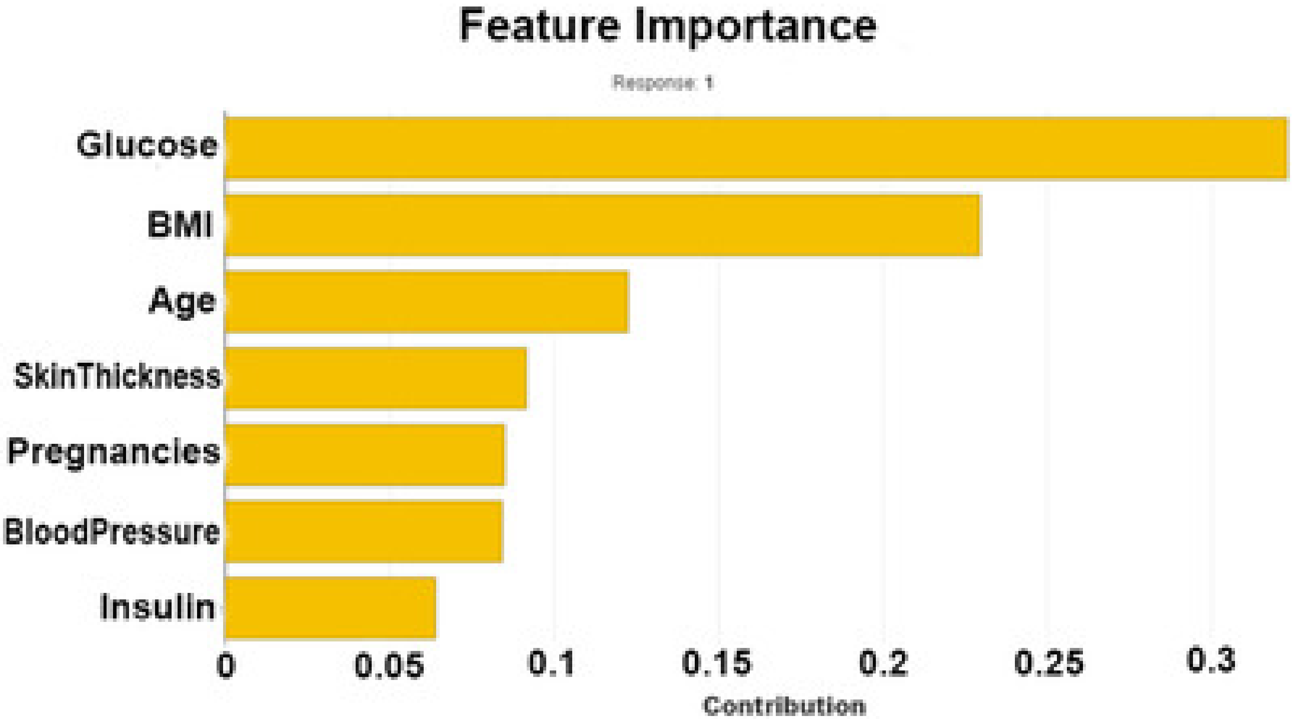
RESULTS AND DISCUSSION:

This section presents the results and discussion of the proposed automatic diabetes prediction system. First, the performance of various machine learning techniques is discussed. Next, the implemented website framework and Android smartphone application are demonstrated. We used precision, recall, F1 score, AUC, and classification accuracy to evaluate various ML models. Equations of these metrics are expressed as

Performance metrics of various classifiers with SMOTE technique in the merged dataset

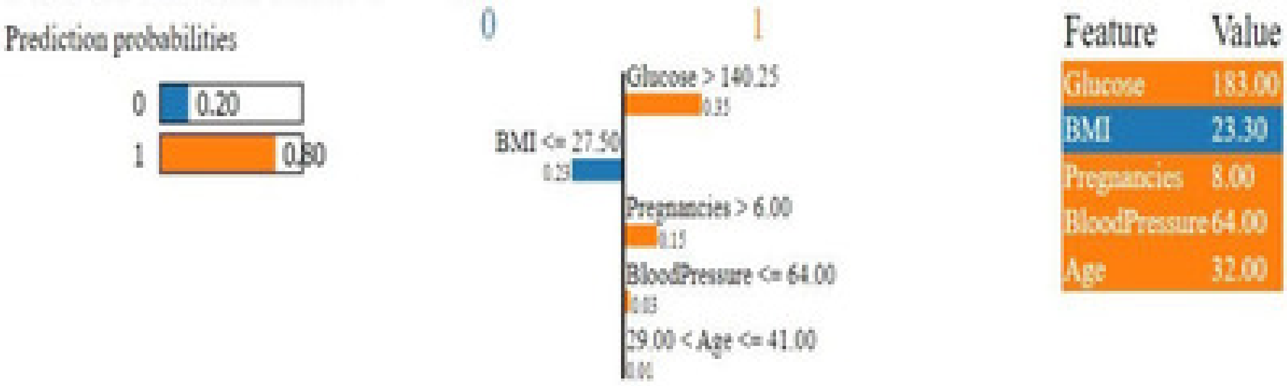
Classifier	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**AUC**
Logistic regression	0.78	0.77	0.77	77%	0.88
KNN	0.78	0.76	0.76	76%	0.85
Random forest	0.78	0.78	0.78	78%	0.87
Decision tree	0.75	0.73	0.73	73%	0.75
Bagging	**0.80**	**0.79**	**0.79**	**79%**	**0.87**
Adaboost	0.79	0.78	0.78	78%	0.85
XGboost	0.78	0.78	0.78	78%	0.84
Voting	0.79	0.79	0.79	79%	0.86
SVM	0.78	0.75	0.76	75%	0.87


True Positive Rate
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
False Positive Rate
XG Boost Classifier, AUC=0.84


Feature Importance
Response: 1
Glucose
BMI
Age
SkinThickness
Pregnancies
BloodPressure
Insulin
0
0.05
0.1
0.15
0.2
0.25
0.3
Contribution

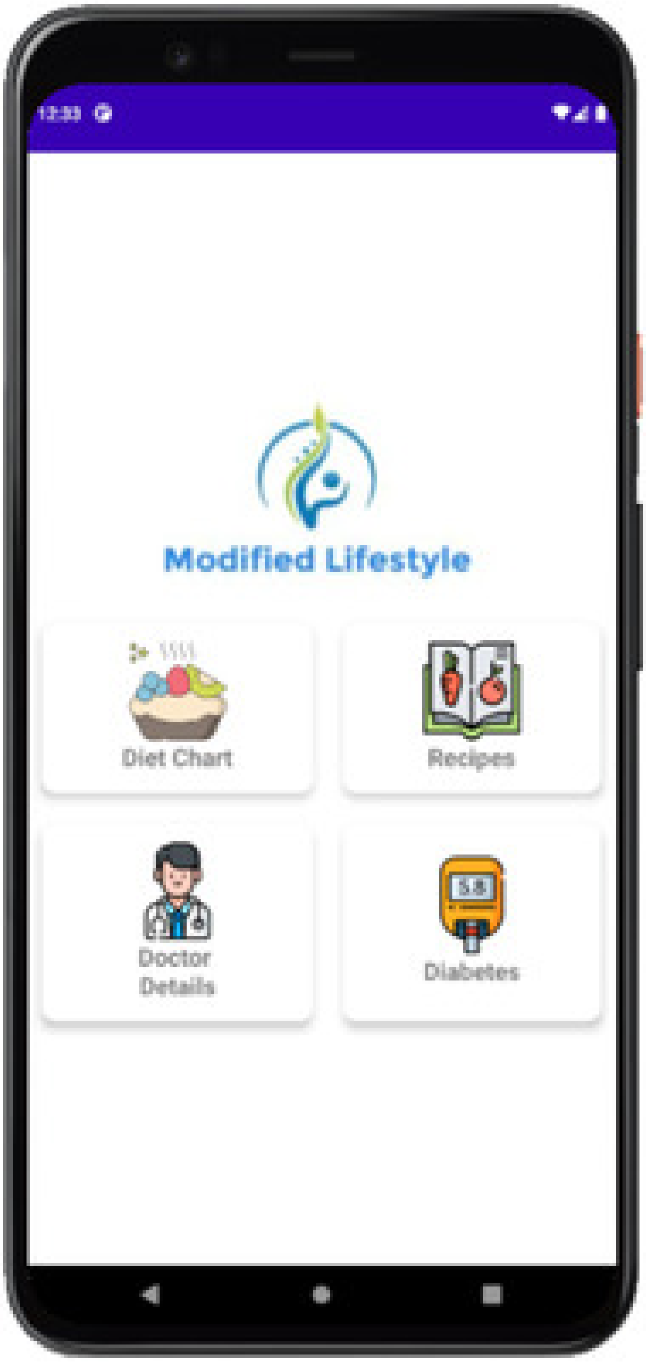
AI techniques with SHAP and LIME frameworks are implemented to understand how the model predicts the decision

an interpretation of the XGBoost model implemented by the LIME explainable AI method. According to this figure, the model predicts diabetes correctly for this specific person with 80% confidence



Feature	Value
Glucose	183.00
BMI	23.30
Pregnancies	8.00
BloodPressure	64.00
Age	32.00

displays the home screen of the proposed Android mobile application created using the best classification algorithm XGBoost. Finally, a survey was conducted in which users rated the application's various features. Android application's survey results. Sixteen volunteers reviewed the application in total, and all of them were female. The participants rated each feature on a scale of 1 to 10, and their average was calculated. According to this figure, the diabetes prediction and daily diet chart features of the application achieved the highest ratings of 8.40 and 8, respectively.



CONCLUSIONS:

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.