AI BASED DIABETES PREDICTION SYSTEM

TEAM LEADER

712221121010 – MATHESWARAN V

Phase 3 Document Submission

Project Title: AI Based Diabetes Prediction System

Phase 3: Development Part 1

Topic: Start building AI Based Diabetes Prediction System by loading and preprocessing the dataset

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.

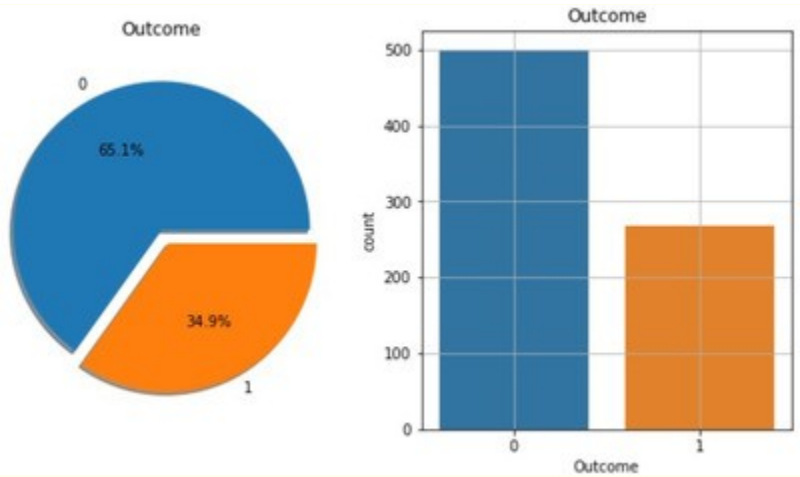


FIGURE 2
Percentage of people having diabetes in the Pima Indian dataset

TABLE 1

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. Table 2 illustrates distinct features of the private RTML dataset with their minimum, maximum, and average values.

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

2.2. 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].

Figure 3 shows the mutual information of various features, that is, the importance of each attribute of this dataset. For example, according to this figure, the diabetes pedigree function seems less important according to this mutual information technique.

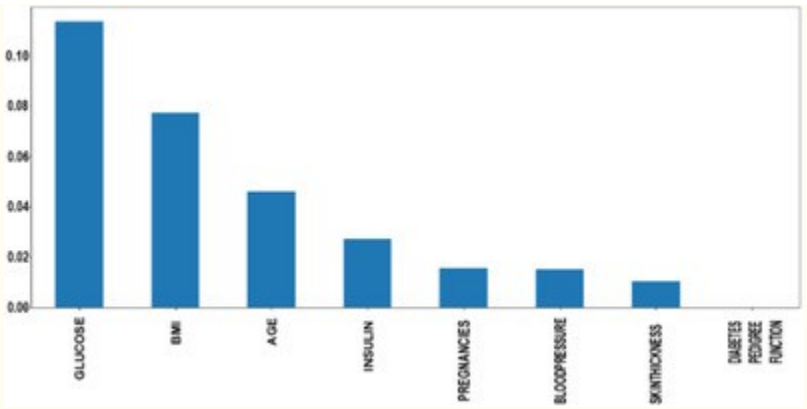


FIGURE 3
Feature importance hierarchy

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). Various regression and ensemble learning techniques have been successfully used in many works to predict missing values [25, 26]. An extensive investigation has been performed while choosing the best-performed regressor technique to predict the insulin feature of the RTML dataset from the Pima Indian dataset. As the actual value of the insulin was not available in the RTML dataset, the Pima Indian dataset was initially used to select the best regression model. First, the Pima Indian dataset was divided into an 8:2 ratio and three supervised regression models, extreme gradient boosting technique (XGB), support vector regression (SVR), and Gaussian process regression (GPR), have been employed to predict the selected outcome, that is, insulin of the validation samples of the Pima Indian dataset. Next, we computed the root mean square error (RMSE) of various regression frameworks as

$$\mathrm{RMSE}=\sqrt{\sum_{i=1}^{N}\left(\mathrm{Predicted}_i-\mathrm{Actual}_i\right)^2/N}$$

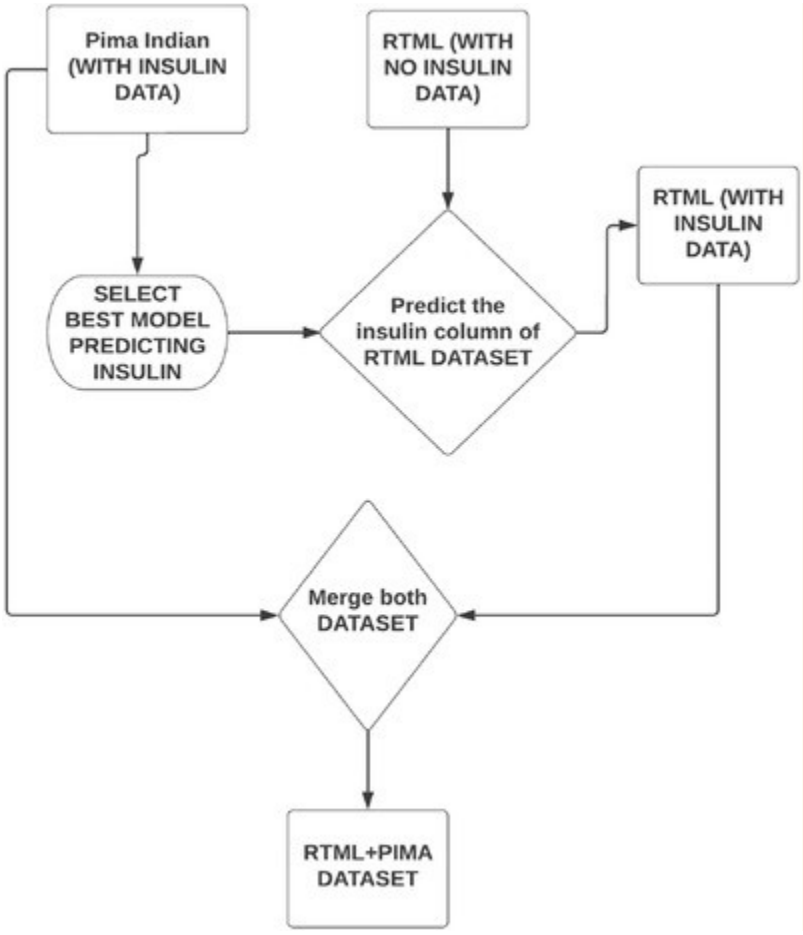
where *N* denotes the total number of validation samples of the Pima Indian dataset.

According to Table 3, the XGB technique exhibits the lowest RMSE of insulin on the Pima Indian dataset. Therefore, this model has been used to predict the missing insulin column of the collected RTML dataset from the Pima Indian dataset. The working steps of predicting insulin in the RTML dataset have been illustrated in Figure 4.

TABLE 3

RMSE of various regression models on the Pima Indian dataset

Regression model	RMSE
XGB	0.36
SVR	0.45
GPR	0.43



Merged dataset: After the semi-supervised approach, we predicted the insulin feature and merged the RTML dataset with the Pima Indian dataset. The merged dataset contained 877 data with all the features, excluding the diabetes pedigree function, as it was the least important feature according to mutual information.

SMOTE and ADASYN for class imbalance: The merged dataset used in this work comprises the imbalance problem with 302 and 669 diabetes and non-diabetes samples, respectively. To take care of this problem, the SMOTE and ADASYN techniques have been applied to the training dataset, leaving the testing data unaffected. Adaptive Synthetic Sampling, known as ADASYN, is a synthetic data generation technique with the characteristics of not duplicating minority samples and generating more data for 'harder to learn' examples [13]. As a result, the minority class will be sampled to the same extent as the majority class.

Min–Max normalization: In this research, we used the min–max normalization technique. The data has been scaled to the same range using the following equation:

$$X_{scaled} = X - X_{min} X_{max} - X_{min}$$

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

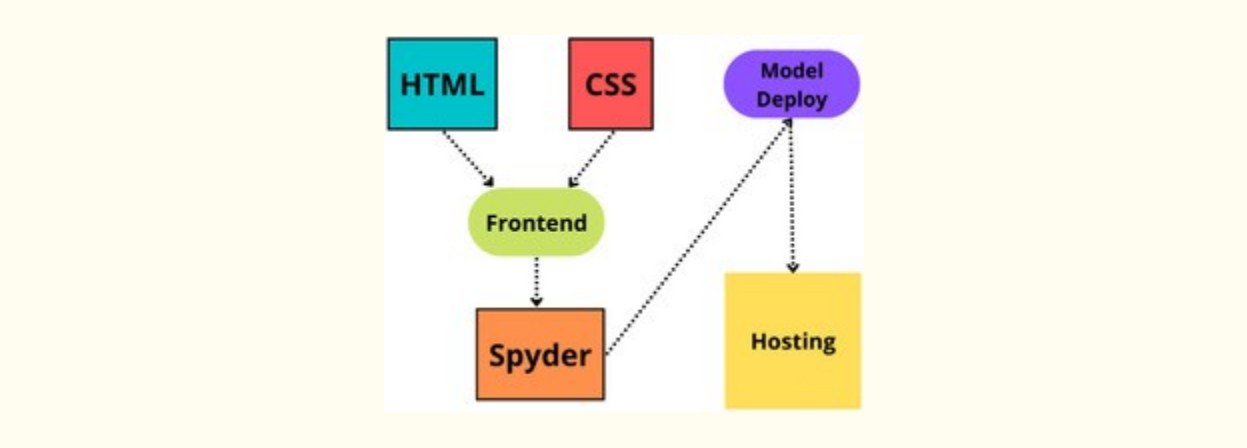
$$\mathrm{Gini}_i = 1 - \sum_{k=1}^{n} \left(p_{i,k}\right)^2$$

$$\mathrm{Entropy} = \sum_{i=1}^{n} -p_i \log_2 p_i$$

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.



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

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall}+\text{Precision}}$$

where TP denotes the model is predicting positive, and the result is also positive. FP indicates the positive prediction of the model, but the result is negative. TN expresses the model is predicting negative, and the result is also negative. FN indicates the model predicts negative, but the result is positive. In this work, the holdout validation approach with a stratified 8:2 train-test split has been used for all the machine learning models.

Table 4 compares different performance metrics of various classifiers for the merged dataset with SMOTE synthetic oversampling technique. According to this table, the bagging classifier achieved the best overall performance with 79% accuracy and 0.79 and 0.87 F1 score and AUC, respectively.

TABLE 4

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