SURYA GROUP OF INSTITUTIONS

VIKRAVANDI-605 652

NAAN MUDHALVAN-PROJECT

COURSE

IBM-(ARTIFICIAL INTELLIGENGE)

PROJECT TITLE

AI-BASED DIABETES PREDICTION SYSTEM

PHASE 5

NAME:DHIVYA SRI G

REG NO:422221104010

DEP:CSE

YEAR&SEM:3&5

TEAM-02

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

The novelty of this work is to implement an automatic diabetes prediction website and Android application for a private dataset of female Bangladeshi patients using machine learning and ensemble techniques.

Phases devolpment

Phase 1:

- **Collect the simple data for AI based diabetes prediction system.**
- **Collect data set and design processing.**
- **Perform test and train dataset.**

Phase 2:

- **To design to innovation to solve the problem.**
- **Collect the simple data for diabetes prediction system.**
- **Collect data set and design processing.**
- **Perform test and train dataset.**

Phase 3:

- **In this section being building your project by loading and preprocessing the dataset.**
- **Collect the simple data for diabetes prediction system.**
- **Collect data set and design processing.**
- **Perform test and train dataset.**

Phase 4:

- **In this section being building your project by performing different activities like features engineering,model training,evaluation ,etc as per the instruction in the project.**
- **Collect the simple data for diabetes prediction system.**
- **Collect data set and design processing.**
- **Perform test and train dataset.**

INTRODUCTION

Diabetes is a chronic disease that directly affects the pancreas, and the body is incapable of producing insulin. 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. It can cause many complications, but an increase in

urination is one of the most common ones]. 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 Bangladesh, approximately 7.10 million people had suffered from this disease

Early and accurate diagnosis of diabetes mellitus, especially during its initial development, is challenging for medical professionals. Artificial intelligence and machine learning techniques, providing a reference, can help them gain preliminary knowledge about this disease and reduce their workload accordingly. Significant numbers of research have been performed to predict diabetes automatically using machine learning and ensemble techniques. Most of these works employed the open-source Pima Indian dataset [6]. Some of these articles on automatic diabetes prediction employing the Pima Indian dataset are briefly discussed in the following paragraphs. For instance, Kumar et al. [4] used the random forest algorithm to design a system that can predict diabetes quickly and accurately. The dataset used in this work was collected from the UCI learning repository. First, the authors used conventional data preprocessing techniques, including data cleaning, integration, and reduction.

The accuracy level was 90% using the random forest algorithm, which is much higher when compared to other algorithms. In a recent paper [5], Mohan and Jain used the SVM algorithm to analyze and predict diabetes with the help of the Pima Indian Diabetes Dataset. This work used four types of kernels, linear, polynomial, RBF, and sigmoid, to predict diabetes in the machine learning platform.

system based on accuracy, precision, recall, and F1 score, utilize more custom data to merge with the existing dataset, and apply an explainable AI technique.

This paper implements diabetes mellitus prediction through machine learning. The significant contribution of this work is as follows:

- A significant contribution of this work is to present a unique dataset of diabetes mellitus containing 203 samples. This private dataset has been obtained from female employees of Rownak Textile Mills Ltd, Dhaka, Bangladesh, referred to as the 'RTML dataset' in this paper. We have collected six features from 203 individuals, that is, pregnancy, glucose, blood pressure, skin thickness, BMI, age, and final outcome of diabetes.
- Another contribution of this work is to keep similarities with the feature of the Pima Indian dataset. The missing insulin feature of the RTML dataset was predicted using a semi-supervised technique.
- SMOTE and ADASYN techniques are implemented to minimize the class imbalance issue. Hyperparameter tuning has also been performed in this work.
- Explainable AI technique with SHAP and LIME libraries is implemented to understand how the model predicts the decision. This approach helps to interpret what features play the most crucial role in terms of prediction.
- A website and an Android application have been developed with the finalized best-performed model of this research work to make instantaneous predictions with real-time data.

The novelty of this work is to implement an automatic diabetes prediction website and Android application for a private dataset of female Bangladeshi patients using machine learning and ensemble techniques. An external file that holds a picture, illustration, etc.

Exploratory Data Analysis

	count	**75%**

Glucose	**768.0**	**140.25000**
BloodPressure	**768.0**	**80.00000**
SkinThickness	**768.0**	**32.00000**
Insulin	**768.0**	**127.25000**

PROPOSED SYSTEM

This section describes the working procedures and implementation of various machine learning techniques to design the proposed automatic diabetes prediction system. 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.

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	

Open in a separate window

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

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

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.

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

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

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

TABLE 3

RMSE of various regression models on the Pima Indian dataset

Regression model	RMSE

Regression model	RMSE
XGB	0.36
SVR	0.45
GPR	0.43

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

$$Precision=TPTP+FP$$

$$Recall=TPTP+FN$$

$$F1score=2\times Recall\times PrecisionRecall+Precision$$

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.