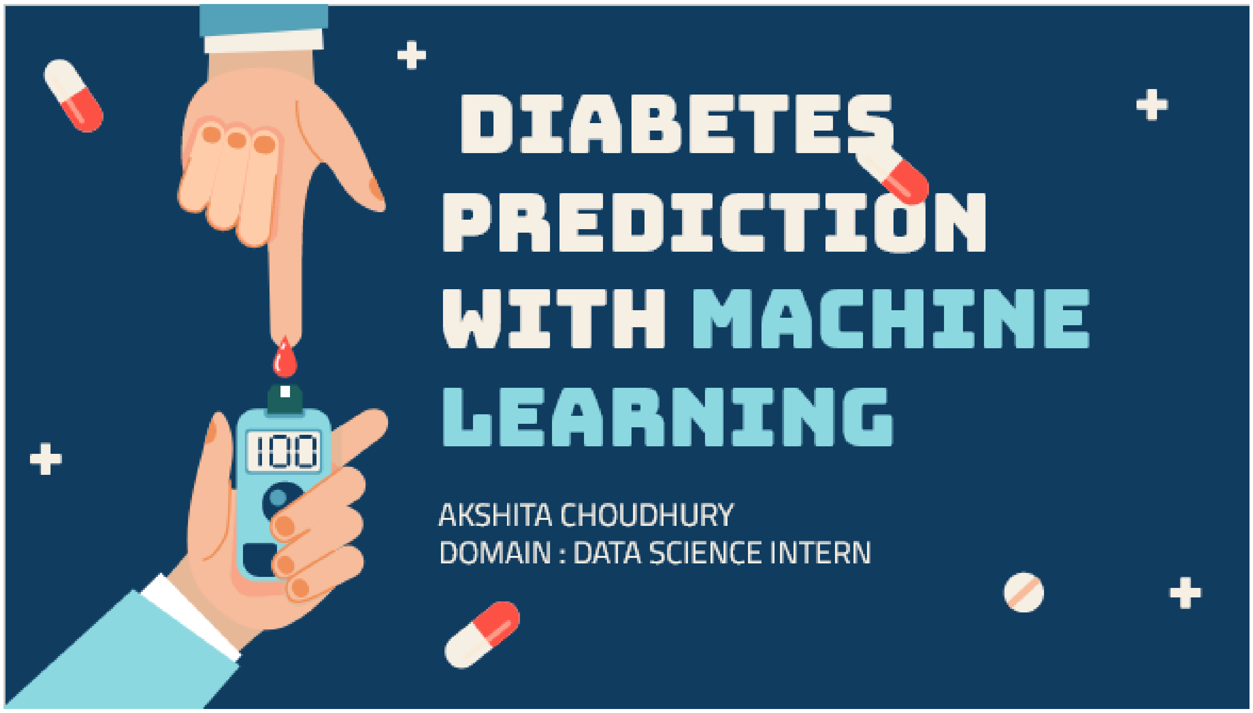
DIABETES PREDICTION

Using Machine Learning techniques


DIABETES PREDICTION WITH MACHINE LEARNING
AKSHITA CHOUDHURY
DOMAIN : DATA SCIENCE INTERN
100

AKSHITA CHOUDHURY

10.06.2023
DATA SCIENCE INTERN

TABLE OF CONTENTS

ABSTRACT

Diabetes is a chronic condition that poses a significant global healthcare challenge. The International Diabetes Federation reports that currently, 382 million people worldwide are living with diabetes, and this number is projected to double to 592 million by 2035. The disease is characterized by elevated blood glucose levels, leading to symptoms like frequent urination, increased thirst, and heightened hunger. Diabetes is a major contributor to various severe health complications, including blindness, kidney failure, amputations, heart failure, and stroke. Normally, when we consume food, our body converts it into glucose. The pancreas is responsible for releasing insulin, which acts as a key to unlock our cells and allow glucose to enter, providing energy. However, in diabetes, this process malfunctions. The disease manifests in different forms, with type 1 and type 2 diabetes being the most common. Additionally, gestational diabetes can occur during pregnancy, and there are other less prevalent forms as well. Machine learning is a rapidly advancing discipline within the field of data science that focuses on enabling machines to learn from experience. This project aims to create a system that can accurately predict the occurrence of diabetes at an early stage by leveraging the insights gained from various machine learning techniques. The algorithms employed include K-nearest neighbors, Naive Bayes, Random Forest, and Support Vector Machine. The accuracy of each model is evaluated, and the one demonstrating high accuracy is chosen as the final model for diabetes prediction.

INTRODUCTION

Diabetes is a rapidly increasing disease that affects people of all age groups, including young individuals. To comprehend diabetes and its development, it is essential to understand the normal functioning of the body in the absence of diabetes. Our body derives sugar (glucose) from the food we consume, particularly carbohydrate-rich foods. Carbohydrates serve as the primary source of energy for our body, including individuals with diabetes. Examples of carbohydrate foods include bread, cereal, pasta, rice, fruits, dairy products, and starchy vegetables. Upon consumption, these foods are broken down into glucose, which circulates in the bloodstream. Some glucose is utilized by the brain to support cognitive function, while the rest is transported to our body's cells for energy production. Additionally, excess glucose is stored in the liver for future energy requirements.

In order for the body to utilize glucose for energy, insulin plays a crucial role. Insulin is a hormone produced by the beta cells in the pancreas. It functions like a key, binding to receptors on cells and allowing glucose to enter the cells from the bloodstream. However, if the pancreas fails to produce sufficient insulin (insulin deficiency) or if the body becomes resistant to the insulin it produces (insulin resistance), glucose accumulates in the bloodstream (hyperglycemia), leading to the development of diabetes. Diabetes Mellitus refers to elevated levels of sugar (glucose) in the bloodstream and urine.

There are three main types of diabetes:

1. Type 1 diabetes: In this type, the immune system is compromised, leading to the failure of cells in the pancreas to produce sufficient insulin. The exact causes of type 1 diabetes are not fully understood, and currently, there are no known methods of prevention.

2. Type 2 diabetes: This is the most common type of diabetes, accounting for approximately 90% of diagnosed cases. In type 2 diabetes, either the cells in the body produce a low quantity of insulin or the body is unable to utilize the insulin effectively. It is influenced by a combination of genetic factors and lifestyle choices.

3. Gestational diabetes: This type of diabetes occurs in pregnant women who experience a sudden rise in blood sugar levels. It typically develops during the later stages of pregnancy and may disappear after childbirth. However, there is a significant risk of developing type 1 or type 2 diabetes in the future for women who have experienced gestational diabetes during pregnancy.

Common symptoms of diabetes include:

- Frequent urination
- Increased thirst
- Fatigue or sleepiness
- Unintentional weight loss
- Blurred vision
- Mood swings
- Confusion and difficulty concentrating
- Frequent infections

The causes of diabetes can be attributed to various factors. Genetic factors play a significant role, with at least two mutant genes in chromosome 6 affecting the body's response to antigens. Additionally, viral infections have been linked to the development of both type 1 and type 2 diabetes. Studies have shown that infections caused by viruses such as rubella, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus can increase the risk of developing diabetes.

EXISTING METHODS

The early and accurate diagnosis of diabetes mellitus, especially in its initial stages, poses a challenge for medical professionals. However, artificial intelligence (AI) and machine learning techniques have been employed to assist in gaining preliminary knowledge about the disease and reducing the workload of healthcare providers. Numerous research studies have focused on predicting diabetes using machine learning and ensemble techniques, with many of them utilizing the well-known Pima Indian dataset. Here, we provide a brief overview of some of these articles.

In one study, a system was developed using the random forest algorithm for quick and accurate diabetes prediction. The authors employed conventional data preprocessing techniques and achieved a high accuracy level of 90%, outperforming other algorithms.

Other studies utilized the SVM algorithm and the Pima Indian Diabetes Dataset to analyze and predict diabetes. Different kernel functions were employed, and accuracies ranging from 0.69 to 0.82 were obtained. The SVM technique with the radial basis kernel function achieved the highest accuracy of 0.82.

Goyal et al. created a smart home health monitoring system for diabetes detection using the Pima Indian dataset. They employed conditional decision making, SVM, KNN, and decision tree algorithms. SVM demonstrated better performance with an accuracy of 75%.

Some explored the prediction of diabetes using ensemble methods and the Pima Indian dataset, considering the area under the ROC curve (AUC) as their accuracy measure. The proposed ensemble classifier achieved an AUC value of 0.95.

Jackins et al. [17] proposed a multi-disease prediction system, including diabetes, using machine learning techniques and the Pima Indian dataset. They found that Naive Bayes performed better than the random forest technique with a 0.43% increase in accuracy.

Mounika et al. [19] predicted diabetes probabilities using machine learning techniques and the public Pima Indian dataset. Kumari et al. [21] employed a soft voting classifier-based ensemble approach for diabetes prediction, achieving an overall highest accuracy and F1 score of 0.791 and 0.716, respectively.

Prabhu and Selvabharathi [3] used the Pima Indian diabetes dataset to predict diabetes using the deep belief network model. They performed data preprocessing, constructed the network model, and fine-tuned the test dataset. Their implementation and simulation were carried out using MATLAB, and an F1 score of 0.808 was achieved, outperforming other classification methods.

Some studies utilized custom datasets or a combination of different datasets. For example, in [14], a type 2 diabetes early prediction system was proposed using machine learning approaches with a private dataset from a Korean hospital. Synthetic oversampling, SMOTE, and undersampling techniques were employed to address data imbalance issues. The random forest and SVM classifiers achieved the highest F1 score of 74%.

Olisah et al. [15] implemented diabetes mellitus forecasting using advanced feature selection and machine learning models. They utilized the Pima Indian and LMCH Iraqi databases, applied a polynomial regression-based preprocessing technique, and performed hyperparameter tuning. The optimized deep neural network (DNN) technique achieved accuracies of 0.972 and 0.973 for the Pima and LMCH datasets, respectively.

In summary, these studies showcase the application of machine learning techniques and the utilization of various datasets to predict diabetes, providing valuable insights into

PROPOSED METHOD WITH ARCHITECTURE

The project runs on a Python script that demonstrates a machine learning approach to predict diabetes using the Pima Indian diabetes dataset combined with RTML dataset . Here's a breakdown of the code's steps and architecture:

1. Data Collection:

- The code reads two CSV files (`diabetes.csv` and `Dataset2.csv`) containing the Pima Indian dataset and the RTML dataset
- The data from both files are concatenated into a single dataframe (`dataset`).

2. Data Preprocessing:

- Zero values in specific columns (e.g., Glucose, BloodPressure, BMI, SkinThickness, Insulin) are replaced with appropriate measures (median or mean) to handle missing or inconsistent data.
- Descriptive statistics (mean, count, etc.) of the dataset are calculated.
- A heatmap is created to visualize the correlation between different features.

3. Exploratory Data Analysis (EDA):

- Histograms and distribution plots are created to analyze the distribution of specific features (e.g., Glucose, Insulin, BMI, Age, BloodPressure, Pregnancies) among individuals with diabetes (Outcome = 1).

4. Data Splitting:

- The dataset is split into training and testing sets using the `train_test_split` function from scikit-learn.
- The size of the training and testing sets is printed for reference.

5. Feature Scaling:

- The independent features in the training and testing sets are standardized using the `StandardScaler` from scikit-learn.

6. Model Building:

- The script uses the K-nearest neighbors (KNN) algorithm to build a classification model.
- Grid search is performed using `GridSearchCV` to find the best value for the hyperparameter `n_neighbors` of the KNN classifier.
- The model is trained with the best parameters and evaluated on the test set.
- Confusion matrix and accuracy metrics are computed and displayed.

7. Additional Models:

- The script also builds and evaluates models using Support Vector Machine (SVM), Naive Bayes, and Random Forest Classifier algorithms.

- Confusion matrices and accuracy metrics are computed and displayed for each model.

8. Model Saving:

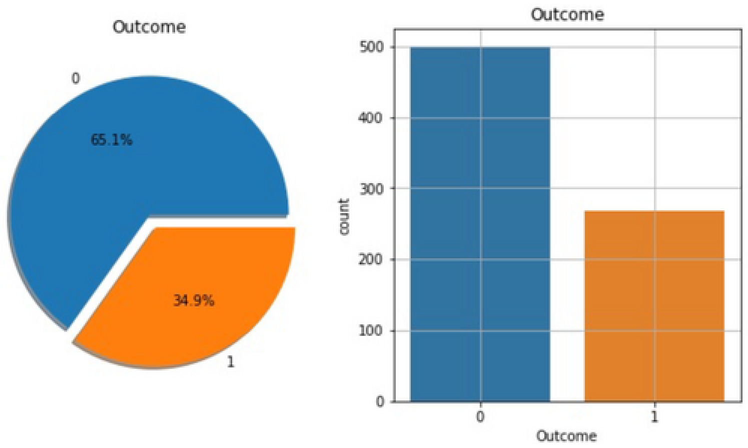
- The trained SVM classifier and the StandardScaler object are saved using the `pickle` module for future use.

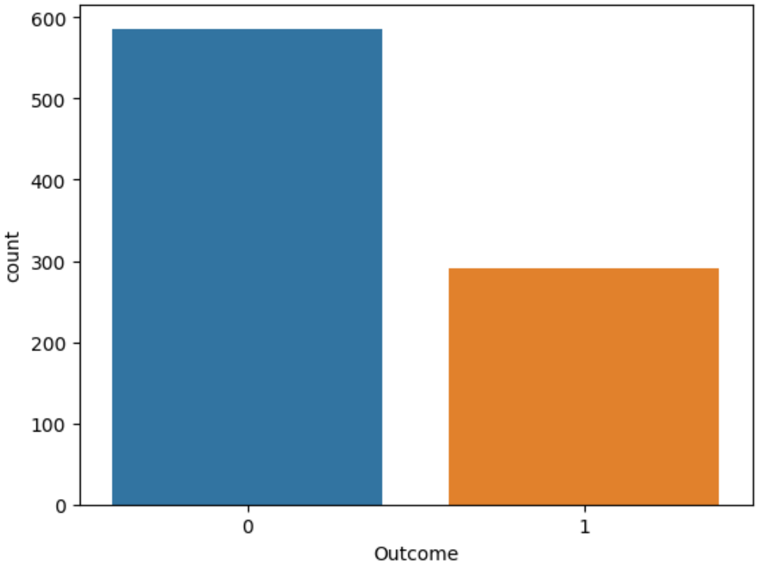
Overall, this code provides a pipeline for data preprocessing, exploratory data analysis, model training, evaluation, and saving. It explores multiple classification algorithms to predict the presence or absence of diabetes based on various features from the Pima Indian diabetes dataset.

METHODOLOGY

In this section, we will explore various classifiers commonly used in machine learning for predicting diabetes. We will also outline our proposed methodology aimed at improving accuracy. This paper employs four different methods, which will be described in detail. The outcome of these methods will be accuracy metrics, which will allow us to evaluate the performance of the machine learning models. Once the models are developed, they can be utilized for making predictions.

The dataset used in this study is the diabetes dataset, obtained from the source https://github.com/tansin-nabil/Diabetes-Prediction-Using-Machine-Learning/blob/main/RTML%20with%20Insulin.csv and https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database . The objective of this analysis is to predict, based on these measurements, whether a patient is diabetic or not. 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.

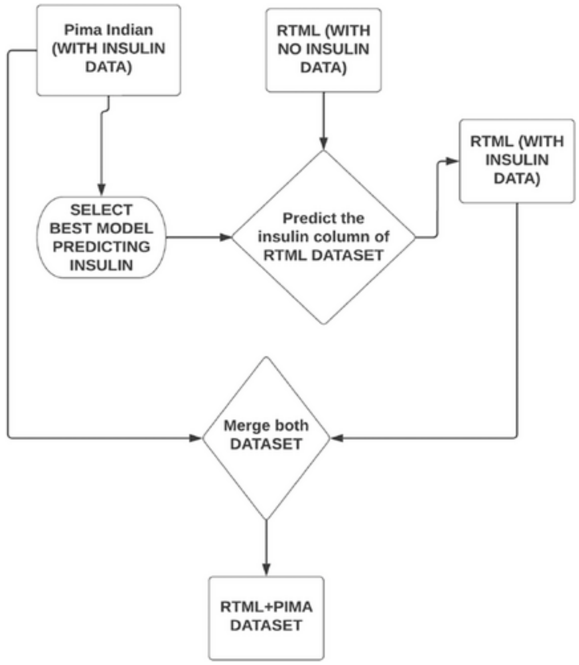




After merging with RTML dataset with insulin and dropping the ‘Diabetes Pedigree Function’ from the PIMA indian dataset.

Features of the dataset used in this project:

1. Pregnancies
2. Skin Thickness
3. Glucose
4. Insulin
5. Age
6. Blood Pressure
7. BMI


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

Dependencies used:

1. Pandas - 2.0.1
2. Pip - 32.1.2
3. Seaborn - 23.1.2
4. Matplotlib - 0.12.2
5. Numpy - 1.24.3
6. Sklearn - 0.0.post4

IMPLEMENTATION

The K-nearest neighbors (KNN) classifier approximates a discrete-valued function by considering K nearest neighbors. It constructs a plane using the training points and calculates the distance between the query point and the trained points. By determining the K nearest neighbors based on the dataset, it performs classification using majority voting. For our research, we utilized K = 5 for binary classification.

Random forest is an ensemble learning model that combines the predictions of multiple decision trees. It averages the predictions to make a final decision [7]. In our research, we applied random forest with 400 estimators, a minimum of 5 samples per leaf, and used the 'Gini' impurity metric. We also employed hyperparameter tuning to optimize the performance of the random forest model.

Support vector machine (SVM) is a supervised classification algorithm that determines the best hyperplane for classification [11]. In our study, we experimented with various SVM kernels using the training set. Ultimately, we found that an SVM with a linear kernel and parameter values of C = 10 and

= 1 yielded the best results for our dataset.

Naive Bayes classification: Naive Bayes is a discrete-valued function that approximates a given function by considering the probability of each feature independently. To classify data, it calculates the likelihood of each class based on the available training data. In our research, we used Naive Bayes for binary classification and employed the principle of maximum likelihood estimation to determine the class probabilities.

CONCLUSIONS

In evaluating the performance of the Random Forest classification model, it is important to analyze the accuracy of its predictions. In the given code snippet, the accuracy of a model is calculated by comparing the predicted labels (`y_pred`) with the actual labels (`y_test`).

Upon examining the results, it is observed that out of the total number of samples, there are 313 correct predictions and 39 incorrect predictions. The number of correct predictions represents the instances where the predicted labels match the actual labels, while the number of incorrect predictions indicates the instances where the predicted labels differ from the actual labels.

To quantify the accuracy of the Random Forest model, the ratio of correct predictions to the total number of predictions is calculated. In this case, the accuracy is determined to be approximately 0.8892, or 88.92%. This means that the model correctly classifies the data in nearly 89% of the cases.

Measuring accuracy is a common way to assess the performance of a classification model, as it provides an overall understanding of how well the model predicts the target variable. However, it is important to consider other evaluation metrics as well, such as precision, recall, and F1-score, depending on the specific requirements and characteristics of the problem at hand. These metrics can provide a more comprehensive analysis of the model's performance, especially in situations where the classes are imbalanced or have varying costs associated with different types of errors.