SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA

SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERNG

SESSION: July-Dec 2022

PROJECT REPORT

On

DIABETES DETECTION USING ML

SUBJECT: INTRODUCTION TO DATA SCIENCE [BTIBM505]

BRANCH: B.TECH-CSE
SECTION: E(BDA)
III YEAR/ 5th SEM

Submitted By:

Viral Parikh
[20100BTBDAI07224]
Divyansh Sharma
[20100BTBDAI077206]
Yash Purohit
[20100BTBDAI07225]

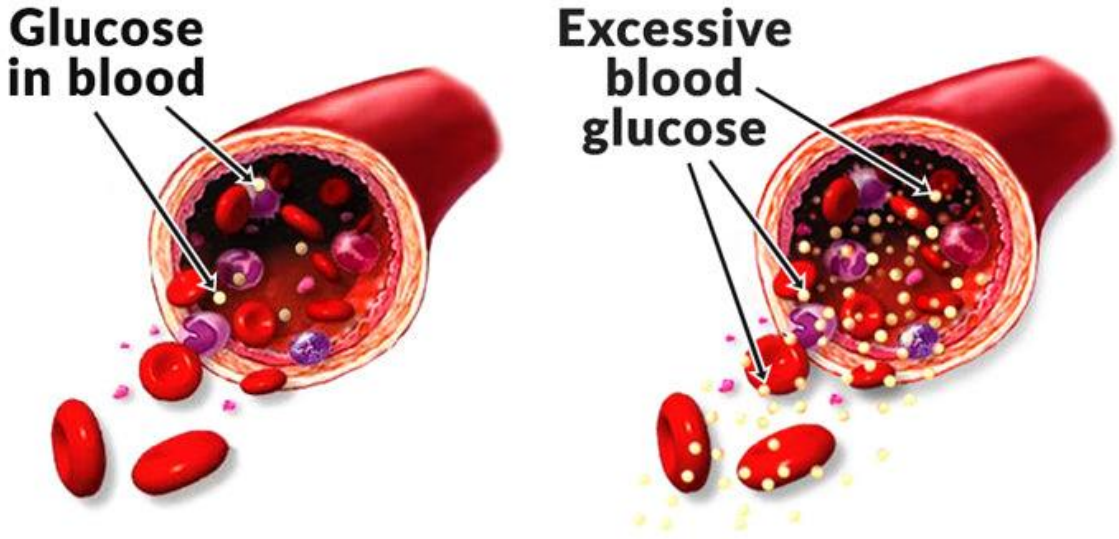
Submitted To:

Mr. Om Kant Sharma

DIABETES DETECTION

Diabetes mellitus is a chronic, lifelong disease caused by excessively high blood sugar levels. Classification strategies are widely used in the medical field to classify data into different classes based on certain constraints against an individual classifier. Diabetes is a disease that affects the body's ability to produce the hormone insulin, thereby making carbohydrate metabolism abnormal and raising blood sugar level. As reported by the World Health Organization report in 2019 reported 463 million are with diabetes, 1.5 million deaths, as the report indicates that is not difficult to guess how much diabetes is very serious and chronic.

Many researchers conduct experiments to diagnose diseases using different machine learning approach classification algorithms such as logistic regression, KNN, SVM, Random Forest classifier because researchers have proven demonstrated that machine learning algorithms are more effective.



INTRODUCTION

Diabetes is one of deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain. Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in the upcoming years. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D appears to be the most common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is thought to be due to autoimmunological destruction of the Langerhans islets hosting pancreatic-β cells.

T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other forms of DM, classified on the basis of insulin secretion profile and/or onset, include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the Young), neonatal, mitochondrial, and pregnancy diabetes. The symptoms of DM include polyuria, polydipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose =7.0 mmol/L).

Diabetes Detection Kit

PROBLEM

Diabetes detection using machine learning is done by classification.
Health condition diagnosis is an essential and critical aspect for healthcare professionals. Classification of a diabetes type is one of the most complex phenomena for healthcare professionals and comprises several tests. However, analyzing multiple factors at the time of diagnosis can sometimes lead to inaccurate results. Therefore, interpretation and classification of diabetes are a very challenging task. Recent technological advances, especially machine learning techniques, are incredibly beneficial for the healthcare industry. Numerous techniques have been presented in the literature for diabetes classification.

The proposed diabetes classification and prediction system has exploited different machine learning algorithms. First, to classify diabetes, we utilized logistic regression, random forest, and MLP. Notably, we fine-tuned MLP for classification due to its promising performance in healthcare, specifically in diabetes prediction.

Second, we implement three widely used machine learning algorithms for diabetes prediction, i.e., moving averages, linear regression, and LSTM. Mainly, we optimized LSTM for crime prediction due to its outstanding performance in real-world applications, particularly in healthcare. The implementation details of the proposed algorithms are as follows.

The remarkable advancements in biotechnology and public healthcare infrastructures have led to a momentous production of critical and sensitive healthcare data. By applying intelligent data analysis techniques, many interesting patterns are identified for the early and onset detection and prevention of several fatal diseases.

Diabetes mellitus is an extremely life-threatening disease because it contributes to other lethal diseases, i.e., heart, kidney, and nerve damage. In this paper, a machine learning based approach has been proposed for the classification, early-stage identification, and prediction of diabetes. Furthermore, it also presents an IoT-based hypothetical diabetes monitoring system for a healthy and affected person to monitor his blood glucose (BG) level. For diabetes classification, three different classifiers have been employed, i.e., random forest (RF), multilayer perceptron (MLP), and logistic regression (LR). For predictive analysis, we have employed long short-term memory (LSTM), moving averages (MA), and linear regression (LR). For experimental evaluation, a benchmark PIMA Indian

Diabetes dataset is used. During the analysis, it is observed that MLP outperforms other classifiers with 86.08% of accuracy and LSTM improves the significant prediction with 87.26% accuracy of diabetes. Moreover, a comparative analysis of the proposed approach is also performed with existing state-of-the-art techniques, demonstrating the adaptability of the proposed approach in many public healthcare applications.

By exploiting the advantages of the advancement in modern sensor technology, IoT, and machine learning techniques, we have proposed an approach for the classification, early-stage identification, and prediction of diabetes in this paper. The primary objective of this study is twofold. First, to classify diabetes into predefined categories, we have employed three widely used classifiers, i.e., random forest, multilayer perceptron, and logistic regression. Second, for the predictive analysis of diabetes, long short-term memory (LSTM), moving averages (MA), and linear regression (LR) are used. To demonstrate the effectiveness of the proposed approach, PIMA Indian Diabetes is used for experimental evaluation. We concluded that, in experimental evaluation, MLP achieved an accuracy of 86.083% in diabetes classification as compared to the other classifiers and LSTM achieved a prediction accuracy of 87.26% for the prediction of diabetes. Moreover, we have also performed a comparative analysis of the proposed approach with existing state-of-the-art approaches. The accuracy results of our proposed approach demonstrate its adaptability in many healthcare applications.

METHODOLOGY

In this section we shall learn about the various classifiers used in machine learning to predict diabetes. We shall also explain our proposed methodology to improve the accuracy.

The variation in glucose levels is cause of diabetes. Insulin balances the blood glucose level in the body, deficiency of which cause diabetes. For the prediction of diabetes machine learning is used, these have many steps like image pre-processing/data preprocessing followed by a feature extraction and then classification. We can use any of the mentioned machine learning classifiers to predict this disease. In the above section we have learning about many classification algorithms, we can either use any one of these to predict the disease or we can explore the techniques to use the hybrid methodology to improve the accuracy over using a single one. Currently, the researches have used the a single classification algorithm and have come up to accuracy of 70 to 80% for detection of the diabetes disease.

We cannot differentiate which of the algorithms are superior or not. classification algorithms used are

- Logistic Regression: Logic regression is used for Predictive Learning Model. To determine output in this classifier, we use a statistical method to analyse the dataset. These data set can have one or more than one independent values. The output is calculated with a data in which there could be two outputs. The aim of this classification algorithm is to find the relationship between the dichotomous category and predictor variables.

- K Nearest Neighbour: As the name suggests the nearest neighbour algorithm is based on the nearest neighbour and this classification algorithm is supervised. It is also called as k- nearest neighbour classification algorithm. A cluster of labelled points are used to understand how the other points should be labelled. For labelling a new point it checks the already labelled points which could be closest to the point to be labelled, i.e closest to the neighbour. In this way depending on the votes of the neighbour the new point is labelled the same label which most of neighbours have. In in algorithm k is the number of neighbours which are checked.

- Decision Trees: This classification algorithm builds the regression models. These models are builded in form of structure which is similar to tree – a tree like structure is created by this classifier. It keeps on dividing the data set into subsets and smaller subsets which develops an associated tree, incrementally. The decision tree is finally created which has decision nodes and leaf nodes. In this tree the leaf node will have details about the classification or the decision taken for classification whereas the decision will have branches. The highest decision node which will be at the top of the tree will correspond to the root node. This will be the best predictor.

The diabetes prediction algorithm consists of three fundamental steps. First, weights are initialized and a sigmoid unit is used in the forget/keep gate to decide which information should be retained from previous and current inputs (C_{t-1}, h_{t-1}, and x_t). The input/write gate takes the necessary information from the keep gate and uses a sigmoid unit which outputs a value between 0 and 1. Besides, a $\mathrm{Tan_h}$ unit is used to update the cell state C_t and combine both outputs to update the old cell state to the new cell state.

Finally, inputs are processed at the output gate and again a sigmoid unit is applied to decide which cell state should be output. Also, $\mathrm{Tan_h}$ is applied to the incoming cell state to push the output between 1 and −1. If the output of the gate is 1, then the memory cell is still relevant to the required production and should be kept for future results. If the output of the gate is 0, the memory cell is not appropriate, so it should be erased. For the write gate, the suitable pattern and type of information will be determined written into the memory cell. The proposed LSTM model predicts the BG level (h_t) as output based on the patient's existing BG level (X_t).

DATASETS

The primary objective of using this dataset is to build an intelligent model that can predict whether a person has diabetes or not, using some measurements included in the dataset. There are eight medical predictor variables and one target variable in the dataset. Diabetes classification and prediction are a binary classification problem.

The dataset consists of 768 records of different healthy and diabetic female patients of age greater than twenty-one. The target variable outcome contains only two values, 0 and 1. The primary objective of using this dataset was to predict diabetes diagnostically.

In [3]:
```
# import dataset

data = pd.read_csv(r'C:\Users\parik\OneDrive\Desktop\diabities\diabetes.csv')
df = pd.DataFrame(data.drop('Pregnancies',axis = 1))
df
```

Out[3]:

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
2	183	64	0	0	23.3	0.672	32	1
3	89	66	23	94	28.1	0.167	21	0
4	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...
763	101	76	48	180	32.9	0.171	63	0
764	122	70	27	0	36.8	0.340	27	0
765	121	72	23	112	26.2	0.245	30	0
766	126	60	0	0	30.1	0.349	47	1
767	93	70	31	0	30.4	0.315	23	0

768 rows × 8 columns

```
ds.columns
```

```
Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
      dtype='object')
```

```
ds.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

```
ds.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Classification report of various machine learning Algorithms:

1. KNN(K- Nearest Neighbour)

```
In [27]: knn = KNeighborsClassifier(10)
         knn . fit(x_train , y_train)
         y_pred = knn . predict(x_test)
         print('acc:',metrics . accuracy_score(y_test , y_pred))
```

```
acc: 0.7532467532467533
```

---

2. Logistic Regression

```
In [70]: from sklearn.linear_model import LogisticRegression
         logisticRegr = LogisticRegression()
         logisticRegr.fit(x_train, y_train)
         logisticRegr.predict(x_test[0].reshape(1,-1))
         predictions = logisticRegr.predict(x_test)
```

```
In [71]: score = logisticRegr.score(x_test, y_test)
         print(score)
```

```
0.7748917748917749
```

---

3. Decision Tree Classifier

```
In [61]: x = pd.DataFrame(df , columns=['Glucose','BloodPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age']).value
         y = df . Outcome . values . reshape(-1 , 1)
         x_train , x_test , y_train , y_test = train_test_split(x , y ,test_size=0.3 , random_state=0)
         from sklearn.tree import DecisionTreeClassifier
         dtc = DecisionTreeClassifier()
         dtc.fit(x_train,y_train)
         y_predd = dtc.predict(x_test)
         from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
         cf1 = confusion_matrix(y_predd,y_test)
         print(cf1)
         print(classification_report(y_predd, y_test))
         print( "accuracy score: ", accuracy_score(y_predd, y_test))
```

```
[[124  34]
 [ 33  40]]
              precision    recall  f1-score   support

           0       0.79      0.78      0.79       158
           1       0.54      0.55      0.54        73

    accuracy                           0.71       231
   macro avg       0.67      0.67      0.67       231
weighted avg       0.71      0.71      0.71       231

accuracy score:  0.70995670995671
```

RESULT

Indian diabetes dataset is used for analysis for this study. It consists of eight independent attributes and one independent class attribute. The study was implemented by python programming language using jupyter notebook. Machine learning algorithms (Logistic regression, K-NN, Decision tree) are used to predict the diabetics disease in early stages.

Performance Metrics: Three widely used state-of-the-art performance measures (Recall, Precision, and Accuracy) are used to evaluate the performance of proposed techniques, as shown in Table 4. TP shows a person does not have diabetes and identified as a nondiabetic patient, and TN shows a diabetic patient correctly identified as a diabetic patient. FN shows the patient has diabetes but is predicted as a healthy person. Moreover, FP shows the patient is a healthy person but predicted as a diabetic patient. The algorithm utilized 10-fold cross-validation for training and testing the classification and prediction model.

Accuracy of all models are given below:
1) KNN = 0.75
2) Logistic Regression = 0.77
3) Decision Tree Classifier= 0.70

Recall
TP/(TP+FN)

Precision
TP/(TP+FP)

Accuracy
(TP+TN)/(TP+TN+FP+FN)

CONCLUSION

Machine learning can help doctors identify and cure diabetes. We will conclude that improving the accuracy of the classification will help the machine learning models perform better. The performance analysis is in terms of accuracy rate among all the classification techniques such as logistic regression, K-nearest neighbors, SVM, random forest. Machine learning has the great ability to revolutionize diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. The main aim is to design and implement diabetes prediction using machine learning methods and performance analysis of that method. The proposed method approach uses Decision tree, KNN, logistic regression, and random forest. The technique may also help researchers to develop an accurate and effective tool that will reach the table of clinicians to help them make better decision about the disease status.