IJCSN - International Journal of Computer Science and Network, Volume 8, Issue 4, August 2019
ISSN (Online): 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Implementation of Machine Learning Classifiers for Predicting the Diabetes Mellitus

[1] Ruchita Gudipati; [2] Muvva Vennela Sai; [3] K Radha

[1] IV-CSE-B.TECH, GITAM University Department, Rudraram, Hyderabad, Telangana, India

[2] IV-CSE-B.TECH, GITAM University Department, Rudraram, Hyderabad, Telangana, India

[3] Assistant Professor,CSE,GITAM University Department, Rudraram, Hyderabad, Telangana, India

Abstract - Nowadays, Diabetes has become a constant chronic disease affecting the mankind. Various causes such as bacterial or viral infection, toxic or chemical contents mix with the food, auto immune reaction, obesity, bad diet, change in lifestyles, eating habit, environment pollution, etc. are responsible in increasing number of victims suffering from Diabetes. Hence, It would be very helpful in predicting this disease at early stage and diagnosing the disease effectively. In health care, this process is carried out using machine learning algorithms to analyze medical data to build to carry out medical diagnoses. Diabetes Mellitus or Diabetes is a serious chronic disease which results in increase of blood sugar. It has always been tedious to identify diabetes, but with emergence of machine learning the identification process has become simpler. Three machine learning algorithms namely SVM, Decision Tree and Naive Bayes are used to detect Diabetes in earlier stages. Algorithms are experimented and evaluated on measures like precision, Accuracy-measure and Recall. The results obtained show Naive Bayes performs better with 76.30% compared to other algorithms. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.

Keywords – Diabetes, Diabetes Mellitus, SVM, Decision Tree, Naïve Bayes

1. Introduction

Machine learning is a emerging tool that allows us to handle petabytes of data .It is a powerful tool incorporating Artificial Intelligence and resulting in promising results. There has been an unprecedented growth in the capabilities of machine learning and has transformed a wide variety of industries. The scope of machine learning is enormous with more and more applications that we never imagined.

Nowadays, diabetes has become a constant chronic disease affecting the mankind. Various causes such as bacterial or viral infection, toxic or chemical contents mix with the food, auto immune reaction, obesity, bad diet, change in lifestyles, eating habit, environment pollution, etc are responsible in increasing number of victims suffering from Diabetes. Hence, It would be very helpful in predicting this disease at early stage and diagnosing the disease effectively. The process of examining and identifying the hidden patterns from large amount of data to draw conclusions is data analytics. In health care, this process is carried out using machine learning algorithms to analyze medical data to build to carry out medical diagnoses. Youngsters are a constant victims of diabetes nowadays Diabetes is caused by the increase level of the sugar in the blood. It can be classified into two categories such as type 1 diabetes and type 2 diabetes.

Type I diabetes: This type occurs when the body fails to produce insulin. People with type 1 diabetes are insulin-dependent, which means they must take artificial insulin daily to stay alive.

Type 2 diabetes: Type 2 diabetes affects the way the body uses insulin. While the body still makes insulin, unlike in type I, the cells in the body do not respond to it as effectively as they once did. This is the most common type of diabetes, obesity could be a effect to it. Diabetes is considered as one of the major health challenges all over the world. The prevalence of diabetes is increasing at a fast pace , deteriorating human, economic and social fabric.

Diabetes prediction is becoming a subject of interest in Healthcare community . Even though a number of decision making support systems are being designed which utilize data mining techniques for prediction of diabetes. The conventional systems are narrowed in scope with just on a

single classifier or a plain combination thereof. Extensive developments are being made for improving the accuracy of such systems using ensemble classifiers. This research implements the Adaboost and bagging ensemble techniques using J48 (c4.5)decision tree as a base learner along with standalone data mining technique J48 to classify patients with diabetes using diabetes risk factors. This classification is conducted on three different ordinal groups in Canadian Primary sentinel surveillance network. According to Canadian Diabetes Association (CDA), between 2010 and 2020, the number of victims diagnose with diabetes in Canada is anticipated to escalate from 2.5 million to about 3.7 million7.Intrestingly,the world picture was almost close to this study. The prediction of Data mining techniques plays a vital role in strategy preparation to prevent communicable as well as non-communicable diseases in located area.

Cluster patterns can be drawn from the diseases like hypertension, diabetes mellitus, cardiovascular diseases; stroke etc in order to locate etiological area and to diagose. MsIs take et al. developed a prototype Intelligent Heart Disease Prediction System using three data mining modelling techniques, namely, Decision Trees, Naïve Bays and Neural Network [7]. Classification strategies are significantly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier. Diabetes is a disease which influences the ability of the body in producing the hormone insulin, which affects the metabolism of carbohydrate abnormal which in turn raise the levels of glucose in the blood. A person with diabetes suffers from high blood sugar. Excessive thirst and hunger, Frequent urination (from urinary tract infections or kidney problems), Weight loss or gain are some of the common symptoms of Diabetes. Many complications are yet to be treated. Some severe complications include ketoacidosis and non-ketonic hyperosmolar coma [1]. Diabetes is examined as a vital serious health matter during which the measure of sugar substance cannot be controlled. Diabetes is not only affected by various factors like weight, height, hereditary factor and insulin but the major reason considered is sugar concentration among all factors. This subject became a topic of research and various machine learning algorithms are been used in order to predict this disease in earlier stages. Algorithms like decision tree, decision table, Naive bayes were found out to contribute in better prediction of these Diabetes .Since large amount of data from different sources are involved,data mining and machine learning can efficiently handle large amount of data integrating background information in the study. Research is performed on a pregnant woman suffering with diabetes. Machine learning algorithms,like Naive Bayes, SVM and Decision Tree. Experimental performance of all the three algorithms are compared on various measures and achieved good accuracy [2]. Many technologies are being implemented on data mining in the sector of health care. It is believed that data mining has the potential to influence the doctor's ability to make drug recommendations. Data mining analysis can be to make business decisions that would improve cost, revenue and operational efficiency of healthcare industry while maintaining high levels of patient care. Data mining techniques are majorly used in healthcare management for, Diagnosis and Treatment, Healthcare Resource Management, CustomerRelationship Management and Fraud and Anomaly Detection. Data mining techniques contribute to identify best treatments in terms of care and cost. Some of the famous data mining methods are classified as: On-LineAnalytical Processing (OLAP), Classification, Clustering, Association Rule Mining, Temporal Data Mining, Time Series Analysis, Spatial Mining, Web Mining etc. The data source can range from data warehouse, database, flatfile or any text file.

The algorithms may be Statistical Algorithms, Decision Tree based, Nearest Neighbor, Neural Network based, Genetic Algorithms based, Ruled based, Support Vector Machine etc. Data mining applications in healthcare can have humongous potential and usefulness. The major concern with health care data mining is the availability of a clean dataset. The data collected must be precise ,clean ,easy to store and ease to mine.. Possible solutions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications .Diabetes is a group of metabolic diseases in which a person has high blood sugar .This high blood sugar produces the symptoms of frequent urination, increased thirst. Insulin is one of the most important hormones in thebody which regulates the blood sugar content in the body. However, improper production of insulin, leads to loss of redundant amount of sugar Via urination. Obesity and lack of exercise appear to possibly play significant roles.

There are three main types of diabetes :

Type I diabetes: this type occurs when the body fails to produce insulin. People with type 1 diabetes are insulin-dependent, which means they must take artificial insulin daily to stay alive.
Type 2 diabetes: Type 2 diabetes affects the way the body uses insulin. While the body still makes insulin, unlike in type I, the cells in the body do not respond to it as effectively as they once did. This is the most common type of diabetes, obesity could be a effect to it.

Type 3: Gestational diabetes is the third main form and occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level .Classification is done to know the exactly how data is being classified Data classification is a two phase process in which first step is the training phase where the classifier algorithm builds classifier with the training set oftuples and the second phase is classification phase wherethe model is used for classification and its performance is analyzed with the testing set of tuples [3]. The Classify Tab is also supported which shows the list of machine learning algorithms. These algorithms generally perform on a classification algorithm and it is excecuted multiple times manipulating algorithm parameters or input data .Weight to increase the accuracy of the classifier.

2. Machine Learning and Data Mining Methods in Diabetes Research

Machine learning technique has significantly contributed to make predictions through training the system initially. Machine learning involves learning structures from the data which is extracted. The term Automatic learning has gathered a greater amount of interest in medical domain due to less amount of time for detection and less interaction with patient, enhancing patient's care [5]. Both type 1 and 2 diabetes as well as other rare forms of diabetes that are directly inherited, including MODY and diabetes due to mutations in mitochondrial DNA, are caused by a combination of genetic and environmental risk factors. Unlike some traits, diabetes does not seem to be inherited in a simple pattern. However, some people are born prone to developing diabetes more comparative to others. It is said that environmental factors contribute to the etiology of T1D. The recent elevated number of T1D incidents projects a changing global environment, which acts either as initiator or accelerator of beta cell auto immunity rather than variation in the gene pool. Several genetic factors are involved in the development of the disease .

It is a fact that advances in genotyping technology, over the past few years, have facilitated rapid progress in large-scale genetic studies. Data mining and machine learning emerge as a key process providing insight into possible relationships among molecules and conditions such as gene–gene, protein–protein, drug–drug, drug–disease or gene–disease, etc .From the perspective of DM, although there are several types of diabetes, the overall results suggest that the articles reviewed refer to T1Dand T2D, with T2D representing the majority of the articles. A few articles refer to pre-diabetes and only one pertains to the metabolic syndrome, which is a term for metabolism-related patho-physiology. The types of data used in each case of the present collection were either clinical , genetic, electrochemical, chemical or medical. Only a few articles used clinical data in combination with genetic data. In addition, it is worth mentioning that the vast majority of the articles reviewed handled only clinical datasets. When it comes to prediction, the main biomarkers used involve anthropometric parameters, demographic characteristics, known risk factors, medical and drug history data, laboratory measurements and epidemiological data. The most common biomarker seems to be blood glucose levels, as expected, since its detection is the basic step toward diagnosis and classification of a candidate diabetic patient [6].

3. Related work

Sajida et al. in [10] discusses the role of Bagging and Adaboost ensemble machine learning methods [11] using J48 decision tree as the basis for classifying the Diabetes and patients as diabetic or non diabetic, based on diabetes risk factors. It was found out that, Adaboost machine learning ensemble technique performs well compared to bagging as well as a J48 decision tree. Orabi et al. in [12] designs a algorithm for diabetes prediction, which aims at the prediction of diabetes a candidate is suffering at a particular age. Decision tree in machine learning was used to develop the concept.

The results which were Obtained were successful as the designed system works well in predicting the diabetes at a particular age, with higher accuracy using Decision tree[13], [14].Pradhan et al in [15] used Genetic programming (GP) for the training and testing of the database for prediction of diabetes by employing Diabetes data set which is sourced from UCI repository. Results achieved using Genetic Programming[16],[17]gives optimal accuracy as compared to their implemented techniques. There can be significant improvement in accuracy as less time was taken for classifier generation. The diabetes prediction was found out to be cost effective. Rashid et al. in [18] designed a prediction model with two sub-modules to predict diabetes-chronic disease. Artificial Neural Network was used in the first module and Fasting Blood Sugar also called as FBS is used in the second module. Decision Tree (DT)[19] was used in detecting the symptoms of diabetes on patient's health.Nongyaoetal.in[20] used four renowned machine learning classification methods namely Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes algorithm which classifies the risk of diabetes .Bagging and Boosting techniques were integrated to enhance the robustness of the model designed.

Experimentation results shows the Random Forest algorithm gives optimum results among all the algorithms employed.

4. Prediction Technique in Data Mining for Diabetes Mellitus

The main drawback in diabetes data classification is insufficiency of resources and improper data mining. Hence, pre-processing must be done to remove redundancy and noisy data from the dataset. Sajida et al. in [9] discusses the role of Bagging and Adaboost ensemble machine learning methods [10] using J48 decision tree as the basis for classifying the Diabetes and patients as diabetic or non diabetic, based on diabetes risk factors. It was found out that, Adaboost machine learning ensemble technique performs well compared to bagging as well as a J48 decision tree. Orabi et al. in [11] designs a algorithm for diabetes prediction, which aims at the prediction of diabetes a candidate is suffering at a particular age. Decision tree in machine learning was used to develop the concept. The results which were Obtained were successful as the designed system works well in predicting the diabetes at a particular age, with higher accuracy using Decision tree.

Major report statistics from various health organisations

• In 2017, National Diabetes Statistic Report for Center Disease Control and Prevention(CDC), gives the facts give an account of the United States that 30.3 million individuals have diabetes, among that 23.1 are analysed and 7.2 million are undiscovered individuals .

• In 2018, the American Diabetes Association models of therapeutic care in diabetes discharges a report about "Order and finding of diabetes" which incorporates the arrangement of diabetes, diabetes care, treatment objectives, criteria for conclusion test ranges and dangers esteems, chance engaged with diabetes.

• In 2017, Global provides details regarding Diabetes by world wellbeing association, it expresses the weight of diabetes, hazard components and inconveniences of diabetes. Likewise, gives the data about counteracting diabetes in individuals with high hazard and overseeing diabetes at beginning times with fundamental solutions to be taken.

There can be significant improve in accuracy by taking less time for classifier generation. It proves to be useful for diabetes prediction at low cost.

5. Data Mining Techniques

5.1. Support Vector Machine (SVM)

The concept of SVM is given by Vapnik *et al.*, which is based on statistical learning theory. SVMs were initially developed for binary classification but it could be efficiently extended for multiclass problems. SVM or sequential minimal optimization (SMO) is a learning system that uses a hypothesis space of linear functions in a high dimensional space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory [19]. SVM uses a linear model to implement non-linear class boundaries by mapping input vectors non-linearly into a high dimensional feature space using kernels. The training examples that are closest to the maximum margin hyper plane are called support vectors. All other training examples are irrelevant for defining the binary class boundaries. The support vectors are then used to construct an optimal linear separating hyper plane (in case of pattern recognition) or a linear regression function (in case of regression) in this feature space [1]. Tao et al. Algorithms which are used in machine learning have various power in both classification and predicting. Saba et al. there is no single technique gives better performance and accuracy for all diseases, whereas one classifier provides or shows highest performance in a given dataset, another method or approach outdoes the others for other diseases.

The new study or the proposed study concentrates on a novel Combination or hybridization of different classifiers for diabetes Mellitus (DD) classification and prediction, thus, overcoming the problem of individual or single classifiers. The new proposed study follows the different machine learning techniques (MLTs) to predict diabetes Mellitus (DM) at an early stage to save human life. Such algorithms are SVM, Naïve Net, Decision Stump, and PEM to predict and increase the prediction accuracy and performance. various algorithm was explained using different parameters such as Glucose, Blood Pressure (BP), Skin Thickness (ST), insulin, Body max index (BMI), Diabetes Pedigree function(DPF), and age.

SVM is considered to be a standard set of supervised machine learning model employed in classification. .The main goal of the support vector machine algorithm is to find a hyper-plane in an N-dimensional space which distinctly classifies the data points. In order to separate the two classes of data points, there are many possible hyper-planes that could be chosen. The main objective is to find a plane that has the maximum distance between data points of both classes. Maximizing the margin distance

provides some reinforcement so that future data points can be classified with more confidence. . Hyper-plane should be selected which is far from the data points from each category for better generalization. The points that lie nearest to the margin of the classifier are the support vectors [21].

5.2. Naive Bayes Classifier

It is a classification technique based on Bayes theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

5.3. Decision Tree Classifier

Decision tree algorithm falls under the category of the supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes [2]. The evaluated performance of Decision Tree technique can be depicted using a Confusion Matrix.

6. Results

Table 1. Displaying the First 5 Rows of the DataSet

```
In [4]: #Printing the first 5 rows of the data
        data.head()
```

Out[4]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	A
0	6	148	72	35	0	33.6	0.627	
1	1	85	66	29	0	26.6	0.351	
2	8	183	64	0	0	23.3	0.672	
3	1	89	66	23	94	28.1	0.167	
4	0	137	40	35	168	43.1	2.288	

```
In [5]: #define X and y
        feature_cols=['Pregnancies','Glucose','BloodPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age']
        x=data[feature_cols]
        y=data.Outcome

In [6]: # split X and y into training and testing sets
        from sklearn.model_selection import train_test_split
        X_train, X_test, y_train, y_test = train_test_split(x, y, random_state=0)
```

Fig.1. splitting your data into a training set and a test set

In [8]:
```
from sklearn.naive_bayes import GaussianNB
from sklearn import svm
from sklearn.tree import DecisionTreeClassifier

# Create Decision Tree classifer object
clfdt = DecisionTreeClassifier()

# Train Decision Tree Classifer
clfdt = clfdt.fit(X_train,y_train)

# make class predictions for the testing set
y_pred_classdt = clfdt.predict(X_test)

#Create a svm Classifier
clfsvm = svm.SVC(kernel='linear',probability=True) # Linear Kernel

#Train the model using the training sets
clfsvm.fit(X_train, y_train)
probas_ = clfsvm.fit(X_train, y_train).predict_proba(X_test)
# make class predictions for the testing set
y_pred_classsvm = clfsvm.predict(X_test)

#Create a Gaussian Classifier
clfnb = GaussianNB()

# Train the model using the training sets
clfnb.fit(X_train, y_train)

# make class predictions for the testing set
y_pred_classnb = clfnb.predict(X_test)
```

Fig.2. Implementing our Classifiers

In [16]:
```
print('DECISION TREE CONFUSION MATRIX')
print(metrics.confusion_matrix(y_test, y_pred_classdt))
print('SVM CONFUSION MATRIX')
print(metrics.confusion_matrix(y_test, y_pred_classsvm))
print('NAIVE BAYES CONFUSION MATRIX')
print(metrics.confusion_matrix(y_test, y_pred_classnb))
```

```
DECISION TREE CONFUSION MATRIX
[[99 31]
 [24 38]]
SVM CONFUSION MATRIX
[[117  13]
 [ 25  37]]
NAIVE BAYES CONFUSION MATRIX
[[114  16]
 [ 29  33]]
```

Fig.3. Confusion Matrix of Three Classification Algo's

```
In [15]: # calculate accuracy=percentage of correct predictions
from sklearn import metrics

print('Decision Tree Accuracy : ', metrics.accuracy_score(y_test, y_pred_classdt))
print('Support Vector Machines Accuracy :  ', metrics.accuracy_score(y_test, y_pred_classsvm))
print('Naive Bayes Accuracy : ', metrics.accuracy_score(y_test, y_pred_classnb))
```

```
Decision Tree Accuracy :  0.7135416666666666
Support Vector Machines Accuracy :  0.8020833333333334
Naive Bayes Accuracy :  0.765625
```

```
In [18]: print('Decision tree precision score : ',metrics.precision_score(y_test, y_pred_classdt))
print('svm precision score : ',metrics.precision_score(y_test, y_pred_classsvm))
print('naive bayes precision score : ',metrics.precision_score(y_test, y_pred_classnb))
#print('Recall : ',metrics.recall_score(y_test,y_pred_class))
#print('F - Measure : ',metrics.f1_score(y_test,y_pred_class))
```

```
Decision tree precision score :  0.5507246376811594
svm precision score :  0.74
naive bayes precision score :  0.673469387755102
```

Fig.4. Precision Scores of 3 Classification Algorithms

```
In [19]: print('Decision tree Recall score : ',metrics.recall_score(y_test, y_pred_classdt))
print('svm Recall score : ',metrics.recall_score(y_test, y_pred_classsvm))
print('naive bayes Recall score : ',metrics.recall_score(y_test, y_pred_classnb))
```

```
Decision tree Recall score :  0.6129032258064516
svm Recall score :  0.5967741935483871
naive bayes Recall score :  0.532258064516129
```

```
In [20]: print('Decision tree F1-Score score : ',metrics.f1_score(y_test, y_pred_classdt))
print('svm F1-Score score : ',metrics.f1_score(y_test, y_pred_classsvm))
print('naive bayes F1-Score score : ',metrics.f1_score(y_test, y_pred_classnb))
```

```
Decision tree F1-Score score :  0.5801526717557252
svm F1-Score score :  0.6607142857142857
naive bayes F1-Score score :  0.5945945945945945
```

Fig.5.Recall and F1-Score of 3 Classification Algo's

```
In [15]: # store the predicted probabilities for class 1
y_pred_prob = clf.predict_proba(X_test)
```

```
In [16]: # Compute ROC curve and area the curve
fpr, tpr, thresholds = metrics.roc_curve(y_test, probas_[:, 1])
# roc_auc = auc(fpr, tpr)
# print ("Area under the ROC curve : %f" % roc_auc)

# Plot ROC curve
plt.clf()
plt.plot(fpr, tpr)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.0])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiverrating characteristic example')
plt.legend(loc="lower right")
plt.show()
```

Fig.6. Implementation of ROC Curve code for SVM

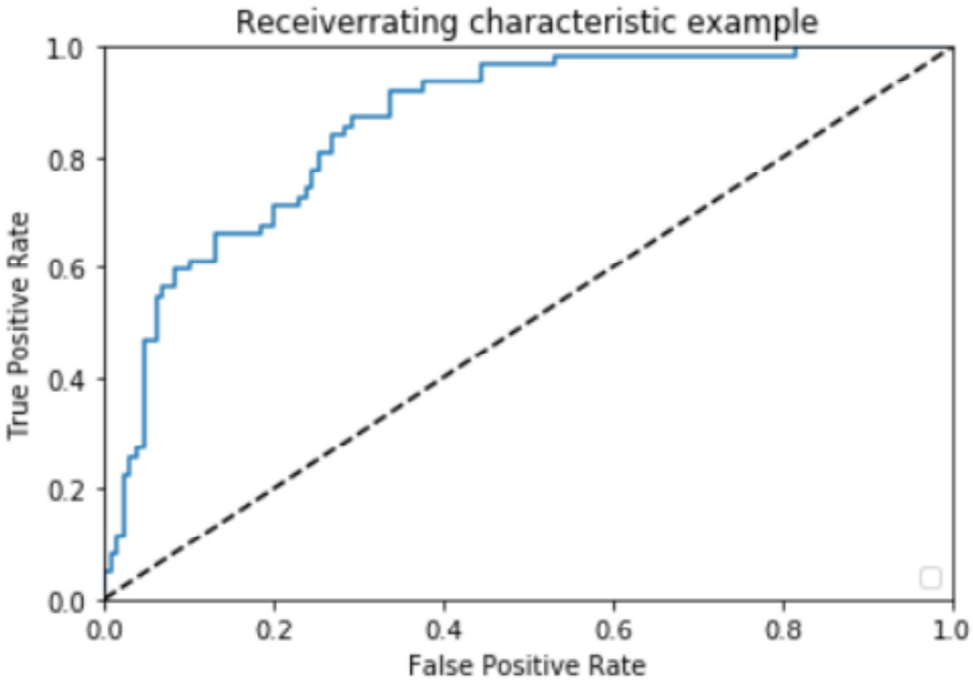

Fig.7. Output of Roc Curve of SVM

```
In [17]: # define a function that accepts a threshold and prints sensitivity and specificity
         def evaluate_threshold(threshold):
             print('Sensitivity:', tpr[thresholds > threshold][-1])
             print('Specificity:', 1 - fpr[thresholds > threshold][-1])

In [18]: evaluate_threshold(0.5)

         Sensitivity: 0.5967741935483871
         Specificity: 0.9153846153846154

In [20]: #AUC is useful as a single number summary of classifier performance.
         #If you randomly chose one positive and one negative observation, AUC represents the likelihood that yo
         #AUC is useful even when there is high class imbalance (unlike classification accuracy).


         # calculate cross-validated AUC
         from sklearn.model_selection import cross_val_score
         cross_val_score(clf, x, y, cv=10, scoring='roc_auc').mean()

Out[20]: 0.8277264957264958
```

Fig.8.Roc Score of SVM classifier

```
In [22]: # store the predicted probabilities for class 1
         y_pred_prob = model.predict_proba(X_test)[:, 1]

In [23]: # IMPORTANT: first argument is true values, second argument is predicted probabilities
         fpr, tpr, thresholds = metrics.roc_curve(y_test, y_pred_prob)
         plt.plot(fpr, tpr)
         plt.xlim([0.0, 1.0])
         plt.ylim([0.0, 1.0])
         plt.title('ROC curve for diabetes classifier')
         plt.xlabel('False Positive Rate (1 - Specificity)')
         plt.ylabel('True Positive Rate (Sensitivity)')
         plt.grid(True)
```

Fig .9.Implementation of ROC Curve code for Naïve Bayes

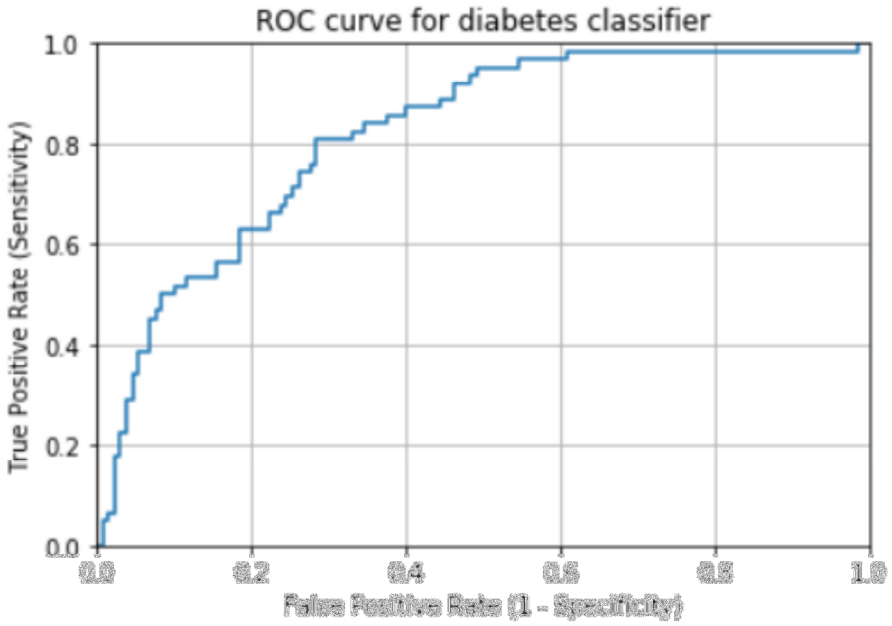

Fig.10.Output of ROC Curve for Naïve Bayes Classifier

```
In [24]: # define a function that accepts a threshold and prints sensitivity and specificity
         def evaluate_threshold(threshold):
             print('Sensitivity:', tpr[thresholds > threshold][-1])
             print('Specificity:', 1 - fpr[thresholds > threshold][-1])

In [25]: evaluate_threshold(0.5)

         Sensitivity: 0.532258064516129
         Specificity: 0.8846153846153846

In [26]: evaluate_threshold(0.3)

         Sensitivity: 0.6290322580645161
         Specificity: 0.7769230769230769

In [27]: #AUC is the percentage of the ROC plot that is underneath the curve:
         # IMPORTANT: first argument is true values, second argument is predicted probabilities
         print(metrics.roc_auc_score(y_test, y_pred_prob))

         0.8176178660049629
```

Fig.11.RocCurve Score for Naïve Bayes Classifier

```
In [18]: # store the predicted probabilities for class 1
         y_pred_prob = clf.predict_proba(X_test)[:, 1]

In [19]: # IMPORTANT: first argument is true values, second argument is predicted probabilities
         fpr, tpr, thresholds = metrics.roc_curve(y_test, y_pred_prob)
         plt.plot(fpr, tpr)
         plt.xlim([0.0, 1.0])
         plt.ylim([0.0, 1.0])
         plt.title('ROC curve for diabetes classifier')
         plt.xlabel('False Positive Rate (1 - Specificity)')
         plt.ylabel('True Positive Rate (Sensitivity)')
         plt.grid(True)
```

Fig.12.Implementation of Roc Curve Code for Decision Trees

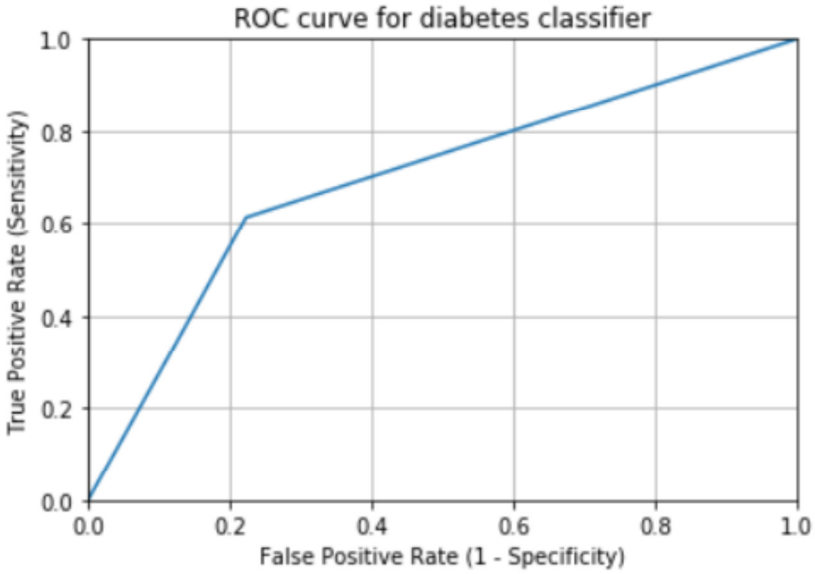

Fig.13.Roc Curve for Decision Tree Classifier

```
In [20]: # define a function that accepts a threshold and prints sensitivity and specificity
         def evaluate_threshold(threshold):
             print('Sensitivity:', tpr[thresholds > threshold][-1])
             print('Specificity:', 1 - fpr[thresholds > threshold][-1])

In [21]: evaluate_threshold(0.5)

         Sensitivity: 0.6129032258064516
         Specificity: 0.7769230769230769

In [22]: evaluate_threshold(0.3)

         Sensitivity: 0.6129032258064516
         Specificity: 0.7769230769230769

In [23]: #AUC is the percentage of the ROC plot that is underneath the curve:
         # IMPORTANT: first argument is true values, second argument is predicted probabilities
         print(metrics.roc_auc_score(y_test, y_pred_prob))

         0.6949131513647643
```

Fig.14.Roc Score of Decision Tree Classifier

6. Conclusion

Diabetes has become a chronic disease claiming many lives hence, detection of it in early stages is vital. In the conducted study, initiatives are taken to predict diabetes. Machine learning classification algorithms were used and Naive Bayes was found to outperform the remaining two classification algorithms with accuracy over 76.30%. Results were obtained from Pima Indians Diabetes Database and classification algorithms. The above results indicate a promising results and the scope of fiels like Machine learning and Data mining. In near future, machine learning classification algorithms can be used to predict and diagnose other diseases. Automation can be used to perform better diabetes analysis.

References

[1] Kumar,D.A.,Govindasamy,R.,2015.PerformanceandEvaluationofClassificationDataMiningTechniquesinDiabetes.InternationalJournalofComputerScienceandInformationTechnologies,6,1312–1319.

[2] Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge

Management Process 5, 1–14. doi:10.5121/ijdkp.2015.5101, arXiv:1502.03774.

[3] Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 09, 1–16. doi:10.4236/jilsa.2017.91001.

[4] Orabi,K.M.,Kamal,Y.M.,Rabah,T.M.,2016.EarlyPredictiveSystemforDiabetesMellitusDisease,in:IndustrialConferenceonDataMining,Springer.Springer.pp.420–427.

[5] Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.

[6] Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.

[7] Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", Procedia Computer Science 82 (2016) 115 – 121.

[8] Harleen,Dr. Pankaj Bhambri "A Prediction Technique in Data Mining for Diabetes Mellitus", Journal of Management Sciences and Technology, 4 (1), October – 2016 ISSN -2347-5005.

[9] Misra, B.B. G. (2007). "Simplified Polynomial Neural Network for classification task in data mining". International Conf. on Evolutionary Computation, 2007, pp 721 – 728.

[10] Perveen,S.,Shahbaz,M.,Guergachi,A.,Keshavjee,K.,2016.PerformanceAnalysisofDataMiningClassificationTechniquestoPredictDiabetes.ProcediaComputerScience82,115–121.doi:10.1016/j.procs.2016.04.016.Issue 1, pp 10, 2010.

[11] NaiArun,N.,Sittidech,P.,2014.EnsembleLearningModelforDiabetesClassification.AdvancedMaterialsResearch931-932,1427–1431.doi:10.4028/www.scientific.net/AMR.931-932.1427.

[12] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933-5.

[13] Priyam,A.,Gupta,R.,Rathee,A.,Srivastava,S.,2013.ComparativeAnalysisofDecisionTreeClassificationAlgorithms.InternationalJournalofCurrentEngineeringandTechnologyVol.3,334–337.doi:JUNE2013,arXiv:ISSN2277-4106.

[14] Esposito, F., Malerba, D., Semeraro, G., Kay, J., 1997. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 476–491. doi:10.1109/34.589207.

[15] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933-5.

[16] Sharief,A.A.,Sheta,A.,2014.DevelopingaMathematicalModeltoDetectDiabetesUsingMultigeneGeneticProgramming.InternationalJournalofAdvancedResearchinArtificialIntelligence(IJARAI)3,54–59.doi:doi:10.14569/IJARAI.2014.031007.

[17] Pradhan,P.M.A.,Bamnote,G.R.,Tribhuvan,V.,Jadhav,K.,Chabukswar,V.,Dhobale,V.,2012.AGeneticProgrammingApproachforDetectionofDiabetes.InternationalJournalOfComputationalEngineeringResearch2,91–94.

[18] TarikA.Rashid,S.M.A.,Abdullah,R.M.,Abstract,2016.AnIntelligentApproachforDiabetesClassification,PredictionandDescription.AdvancesinIntelligentSystemsandComputing424,323–335.doi:10.1007/978-3-319-28031-8

[19] Han, J., Rodriguez, J.C., Beheshti, M., 2008. Discovering decision tree based diabetes prediction model, in: International Conference on Advanced Software Engineering and Its Applications, Springer. pp. 99–109.

[20] NaiArun,N.,Moungmai,R.,2015.ComparisonofClassifiersfortheRiskofDiabetesPrediction.ProcediaComputerScience69,132–142.doi:10.1016/j.procs.2015.10.014.

[21] Sisodia,D.,Singh,L.,Sisodia,S.,2014.FastandAccurateFaceRecognitionUsingSVMandDCT,in:ProceedingsoftheSecondInternationalConferenceonSoftComputingforProblemSolving(SocProS2012),December28-30,2012,Springer.pp.1027–1038.

Authors –

Ruchita Gudipati : currently studying 4th year of engineering in Computer Science from GITAM University, Hyderabad. I have an inclination towards the field of research and towards contributing my part of knowledge especially to the domain of Machine Learning

Muvva Vennela Sai : currently in 4th year pursuing engineering in Computer Science from GITAM University, Hyderabad. Apart from my various interests in this field of study, the subject of Machine Learning has inspired me to research further.

K Radha : completed her BTECH,MTECH at JNTUH. Pursuing PhD in KL University,Guntur. She has 12 years of Teaching Experience and 3 Years of Research Experience. She has applied for DST Research Projects. She has published 25 papers in International Journals, SCOPUS journals and Springer and IEEE Conferences. Her Research Interests are Cloud Computing, Big Data Analytics, Machine Learning, Deep Learning, and Artificial Intelligence.