COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS FOR DIABETIC PREDICTION

G. Sathar[1], Saladi Naveen[2], D. Vithal Varma[3], Md. Reshma[4], J. Anji Nayak[5]
1Assitant Professor, [2]Student, [3]Student, [4]Student, [5]Student
Computer Science Department
Anil Neerukonda Institute of Technology and Sciences, Andhra Pradesh, India

Abstract : Glucose, or commonly called sugar, is an important energy source that is needed by all the cells and organs of our bodies. Excessive growth of blood sugar or glucose levels above the desired level on a sustained basis leads to diabetes. People undergo blood sugar tests to diagnose diabetes. In the field of human health, Computer vision plays a major role by reducing human judgment and providing accurate results. The primary aim of this research is to provide a better classification of diabetes. In this manuscript, we focus on studying the comparative analysis of algorithms to enhance the accuracy of the prediction model using machine learning and data mining techniques. We used the "Pima Indians Diabetes Dataset" standard which was supported by the UCI machine learning repository. Feature selection is performed to increase the potentiality of the dataset. Algorithms like Support Vector Machine (SVM), Naïve Bayes, Decision Tree, K-Nearest Neighbors (KNN) and Logistic Regression are applied on the dataset to evaluate the performance of the model. Evaluation metrics like Precision, Recall, Specificity and mean absolute error for each model are calculated to suggest the best classifier for the sample dataset. Waikato environment for knowledge Analysis toolkit was used to compare the accuracy of the model. Cross-Validation is performed to generalize accuracy. The conclusion depicts an accuracy of 78% in the case of the SVM algorithm. Thus, the work seems to be beneficial for predicting type 2 diabetes (T2D).

Index Terms: *Diabetes, Data mining, and Classification.*

1. INTRODUCTION

Diabetes is a disease that diminishes the body's potentiality to secrete insulin. The food we eat is converted into glucose in the blood. Insulin is used to regulate the glucose levels by pushing the glucose into the cells. These cells convert glucose into energy and perform their specialized function. Insufficient secretion of insulin leads to an increase in blood glucose levels [1].

Diabetes can be majorly classified into 3 types: Type 1, Type 2 and Gestational.

Type 1 diabetes is caused because of inadequate secretion of insulin. Immunity system destroys the beta cells within the body that produce insulin. The quantity of insulin produced is very little, which means that insulin should be furnished to the body through injections to maintain the blood glucose level. It is most predominant in children.

Type 2 diabetes is also referred to as insulin resistance because it does not use the glucose produced in the body. Glucose is accrued in huge amounts within the blood, which causes diabetes. It is most widespread in adults.

Gestational diabetes is caused during the time of pregnancy. Change in the hormones produced leads to this scenario. It happens only during pregnancy. The conceived baby is likely to be infected by type 2 diabetes in the future.

Excess amount of sugar leads to malfunctioning of heart and kidney [2]. It also affects the nervous system.

WHO has published a report, stating the increase of diabetic people from 108 million in 1980 to 422 million in 2014 [1]

The global prevalence of diabetes revealed by the International Diabetes Federation reports the increase of diabetes above the age group of 18 from 4.7% in 1980 to 9.3% in 2019. It is estimated that it would reach 10.2% by 2030 and 10.9% by 2045 [2].

Diabetes prevalence has increased by 64 percent across India over the quarter-century, according to a November 2017 report by the Indian Council for Medical Research [3].

In general, Diagnosis of diabetes is done in accordance with fasting blood tests, which is performed after having a fast of eight hours [3]. It requires much effort for testing. Recent improvement in technology has revolutionized various field of society using data mining [5]. Data mining, also known as Knowledge Discovery in Databases (KDD) is defined as the process of analyzing the data and recognizing the patterns in a large dataset. Data mining along with machine learning has shown some major outbreaks in the medical field for predicting diseases. This helps the people in making a preliminary judgment about diseases based on their physically examined data. These predictions also serve as a reference for doctors [4]. As a large number of people are affected by type2 diabetes, we hereby focused on improving the accuracy of the model based on type 2 diabetes dataset.

2. RELATED WORKS

Increase of Diabetes in the recent years has achieved great attention of people across the world. Recent developments in technology have boosted the medical researchers to adopt latest techniques to predict diseases. There were several existing models, which have been implemented for the classification of diabetes dataset.

Hun wa [6] has conducted a research that aims to extract knowledge from the given set of data and to generate model using improved K-means and the logistic regression algorithm. Waikato Environment for Knowledge Analysis toolkit was used to compare the results. The constructed model has achieved good performance over three different datasets.

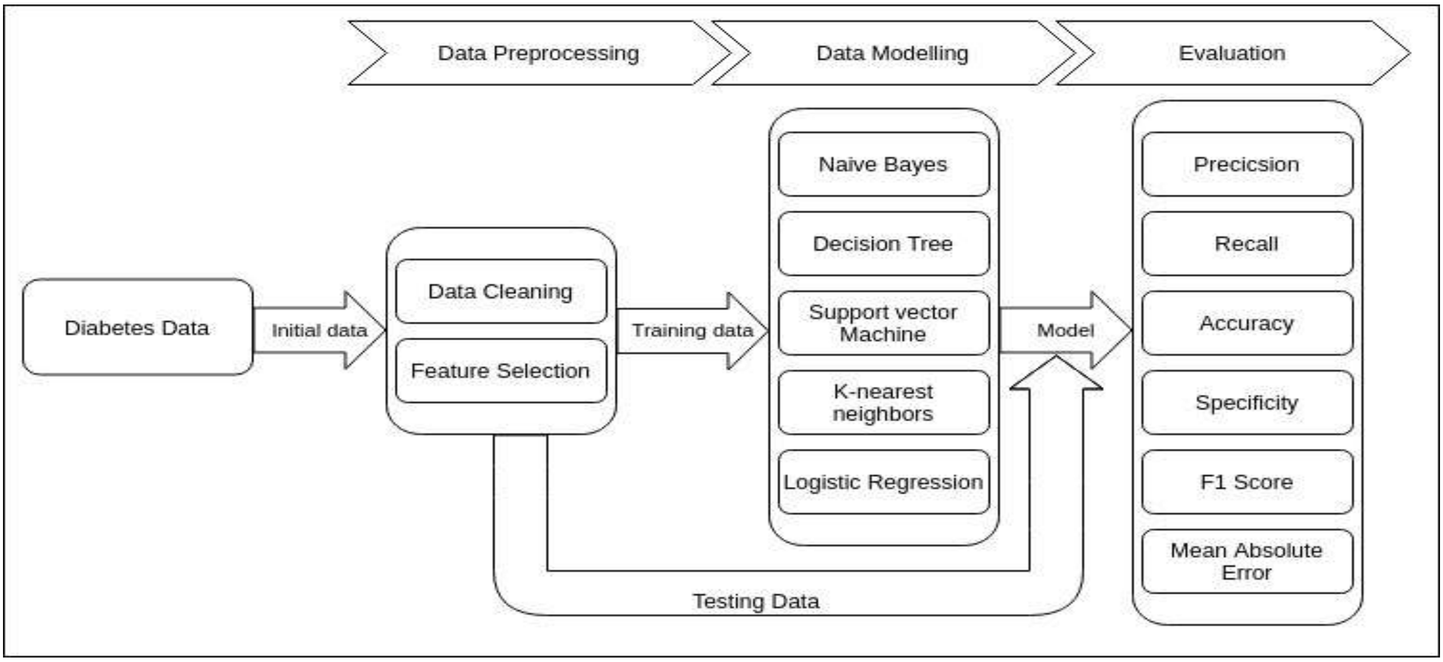
N Sneha [7] has proposed work on Diagnosing diabetes through data mining classification techniques like decision tree, support vector machine, Naïve Bayes, random forest and k-nearest neighbors. They used pima Indian dataset. Model evaluation methods like sensitivity and specificity are used to evaluate the performance of classification.

Debadri Dutta [8] conducted a study on discovering the critical elements that are responsible for diabetes. They improved the quality of the dataset by incorporating the attributes that show major impact using Random Forest Classifier. The proposed work consists of methodologies like SVM, Logistic Regression and Random Forest Classifiers. They used f1-score and recall score for evaluating the model.

Hina S [9] has researched different classifying algorithms such as Naïve Bayes, J.48, ZeroR, Random Forest, and Regression to depict the results using the Pima Indian dataset. The work mainly emphasis on reducing the classification error.

Sadri Sa'di [10], proposed a model for diagnosis of type II diabetes using data mining algorithms like Naive Bayes, RBF Network, and J48. The dataset used for the diagnosis of type II diabetes includes 768 samples from diabetic patients taken from Pima Indians Dataset. The algorithm's performance diagnosis is done using the Weka tool.

3. SYSTEM ARCHITECTURE AND ALGORITHMS



In this paper, we have proposed a three-level architecture model for predicting type2 diabetes. It consists of Data pre-processing, Data Modelling, and Evaluation metrics as core methodologies. Fig.1 shows the architecture of the entire system.

In the first phase, Data Pre-processing techniques like Data cleaning and Feature selection are included to improve the accuracy of the model. It is done to eliminate the problem of overfitting. The data is then split into two subsets, Training data, and Testing data.

In the second phase, different algorithms like Naive Bayes, Decision Tree, Support Vector Machine, K-nearest neighbors, and Logistic Regression are applied over the training dataset. The constructed model is then tested with the testing dataset.

In the third phase, Evaluation metrics like Precision, Recall, Specificity, Accuracy and Mean Absolute Error are calculated to predict the accuracy of the model.

3.1 DATA MINING TOOL KIT:

Waikato environment for knowledge Analysis (WEKA) was developed by Waikato University, New Zealand. It consists of a cluster of several machine learning algorithms for data mining tasks. WEKA contains different data mining techniques like data pre-processing, classification, clustering, regression, association, and visualization [11].

3.2 DATASET:

The quality of data, to a major extent, affects the predicted result. The accuracy depends mainly on the data considered. We used the "Pima Indians Diabetes Dataset" standard which was supported by the UCI machine learning repository [12]. This is a standard dataset that has drawn the values from the real instances.

Dimensions of the dataset: (768, 9)

It consists of 768 instances and each instance is associated with 9 attributes which are all numeric values. The table shows the names, description and value range of each attribute.

S.NO	Name	Description	Unit	Value Range
1	Pregnancy	No of times pregnant	Numeric value	0-9
2	Glucose	Glucose content	Numeric value	0-199
3	Blood pressure	Diastole blood pressure	mmHg	0-122
4	Skin	Triceps skin fold thickness	Mm	0-99
5	Insulin	2-hours serum insulin	Mu/Uml	0-846
6	BMI	Body mass index	Weight in kg Height in m	0-67.1
7	Pedigree	Pedigree function	Numeric value	0.08-2.42
8	Age	Age	Numeric value	21-81

9	Class	Diabetes mellitus type 2	Numeric value	Positive=1
				Negative=0

3.3 DATA PREPROCESSING:

Data Preprocessing is the process of collecting and modifying the data into desired format. The collected data may consist of redundant, inconsistent and noisy data. Preprocessing of data helps in resolving such type of data. Data Preprocessing techniques like Data cleaning and Feature selection are applied over the dataset to improve the accuracy of model.

3.3.1 DATA CLEANING:

In the dataset, we can see some missing values. Most of the inaccurate experimental results were caused by these meaningless values. These values can be replaced by the average of values either mean, median, mode of the attribute. For example, in the original dataset, the values 0, indicates that the real value was missing. We replace them by using the mean of the attributes.

3.3.2 FEATURE EXPLORATION:

Feature selection also called as attribute selection is the process of gathering relevant features for constructing the model. The features are selected based on the correlation between the attributes. It helps in reducing the irrelevant data and improve the prediction accuracy. It is also used to combine different features to produce more sophisticated features. We have used random forest classifier for selecting the features.

RANDOM FOREST CLASSIFIER:

Random Forest is a multi-tree classifier that can be used for both Regression and Classification problems. Decision trees form the basic building blocks of the random forest model. It splits the data into different samples in a random fashion and constructs a decision tree for each data sample. The predictions of each tree are put for voting. In classification problems, the class having the highest number of votes will be considered. In Regression problems, the average of all the class predictions will be considered. The accuracy of the model mainly depends upon the number of trees constructed.

3.4 MODEL SELECTION:

Model selection or algorithm selection is the process of selecting a model that better classifies the dataset. Model construction is completely dependent on the algorithm. There are also some algorithms which do not construct any model and mostly relies on the dataset. In conclusion, we select a model that gives the highest accuracy.

3.4.1 NAÏVE BAYES ALGORITHM:

Naïve Bayes is a classification algorithm that is based on Bayes theorem with naïve independence assumptions. It assumes that the features are independent of the given class. Bayes Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge.

Bayes' Theorem is stated as:

$$\boldsymbol{p(h/d) = p(d/h) * \frac{p(h)}{p(d)}}$$

Where

$p(h/d)$ is the probability of hypothesis h given the data d. This is called the posterior probability.

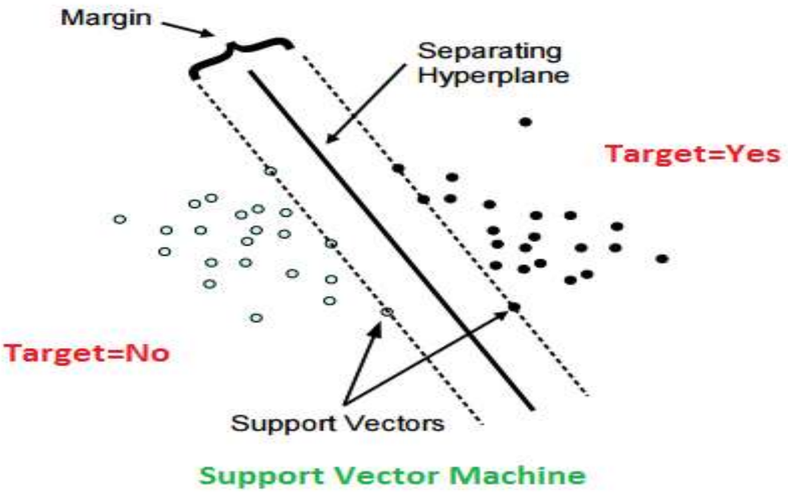
$p(d/h)$ is the probability of data d given that the hypothesis h was true.

$p(h)$ is the probability of hypothesis h being true (regardless of the data) This is called the prior probability of h.

$p(d)$ is the probability of the data (regardless of the hypothesis)

3.4.2 SUPPORT VECTOR MACHINE:

SVM is a classification algorithm that maps the data item into points in the n-dimensional space. The main objective is to find a linear decision boundary (hyperplane) that segregates the data into two classes. The optimal hyperplane is said to be optimal based on maximum marginal boundary i.e., Euclidean distance between the support vectors is maximum. SVM mainly emphasizes on risk minimization. Using of SVM can help in achieving greater performance as it uses significantly less data when compared to other classifiers.



3.4.3 LOGISTIC REGRESSION:

Logistic Regression is a classification technique that works on a probability basis. It is much similar to Linear Regression which gives continuous value. It is named after the logistic function which is the core method. The Logistic function [13], also called a sigmoid function is an S-shaped curve that takes a real number and transforms it into a value that lies between 0 and 1. The sigmoid function is mathematically represented as

$$\mathbf{F(x)} = \frac{1}{1+e^{-x}}$$

When using linear regression, we used a formula of the hypothesis

$$\mathbf{H\ (x) = \beta_0 + \beta_1 X}$$

For logistic regression, we apply sigmoid function to the hypothesis of linear regression.

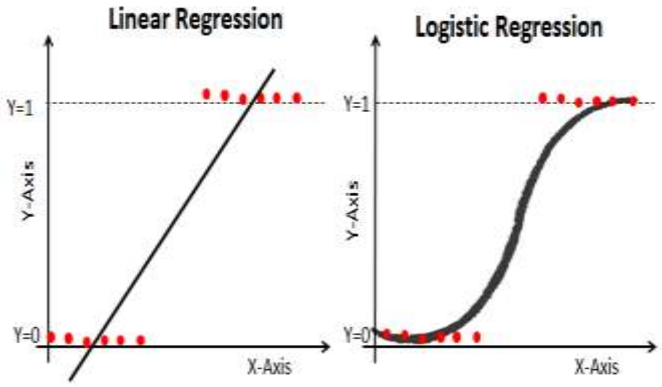
$$\boldsymbol{\sigma(Z) = \sigma(\beta_0 + \beta_1 X)}$$

To transform these values to a discrete class, we select a threshold value, above which we will classify values into class A and below which we classify values into class B.

$$\mathbf{p>=0.5, class=1}$$

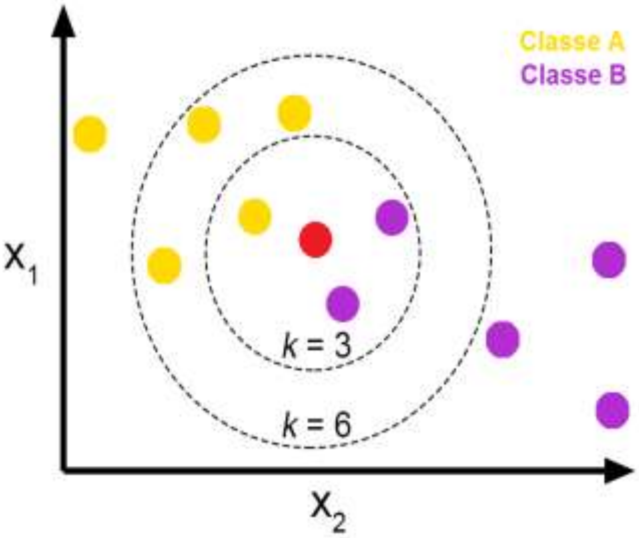
$$p<0.5, \quad class=0$$

In conclusion, we decided to use the logistic regression as one part of our proposed model.



3.4.4 KNN ALGORITHM:

KNN is an instance-based learning algorithm that uses the entire training data for prediction. It does not require any learning as it doesn't have any model. It is also called as a lazy algorithm as it does not perform any generalization. It is based on the theory of feature similarity**.** In this method, each sample should be classified similarly to the surrounding samples. Classification is done based on minimum distance from the new point to the K nearest neighbors i.e., Euclidean or Manhattan distance. The performance of a KNN classifier is primarily determined by the choice of K as well as the distance metric applied. With increase in the value of K, the new point is calculated more accurately as more neighbors are included.

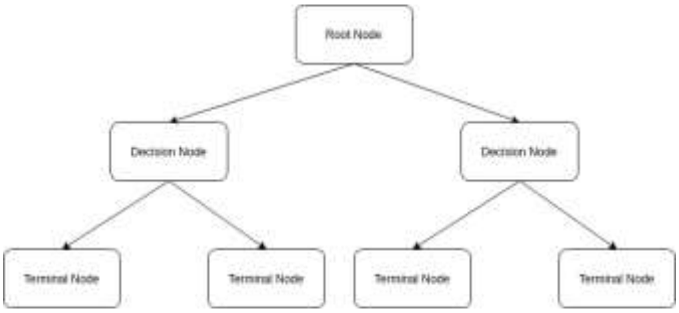


3.4.5 DECISION TREE:

A decision tree is a pictorial representation of all the attributes in the dataset to achieve a more generalized conclusion. It is a prediction algorithm that is used for classification purposes. The dataset is divided into different subsets based on the most prominent attribute. Several algorithms like Hunt's, ID3, and CART algorithms were used to implement the decision tree. The way the dataset is split is completely dependent on the algorithm. A decision tree is drawn upside down with its root at the top. Each node except the leaf node consists of two branches, yes and No. Based on the decision, the tree is constructed. It is done until all the attributes are included in the tree. Tree Pruning method is done at the end to eliminate the unwanted data. This helps in improving the accuracy of the model. For example, in the ID3 algorithm, the attributes with highest information gain is selected as root node. The entropy of each attribute is also calculated. Mathematical Expressed as follows

$$\text{Entropy(S)} = \sum_{i=1}^{n} - p_i \log_2 p_i$$

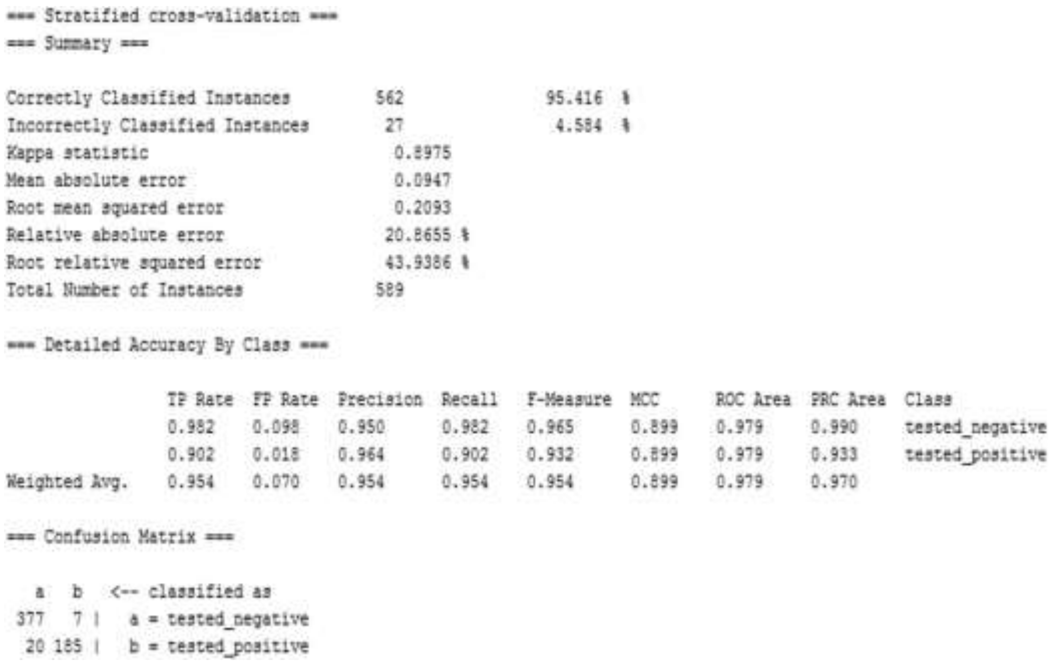
$$\text{Gain (S, J)} = \text{Entropy(S)} - \sum x \in values(J) \frac{|S_x|}{|S|} \text{Entropy } (s_x)$$



4. Evaluating Methods:

We will be evaluating the model by splitting the data set into two portions, training set, and testing set. The training set is used to train the model and the testing set is used to test the model. After being processed by classification algorithms, we evaluate the accuracy of the model.

4.1 WEKA TOOL:

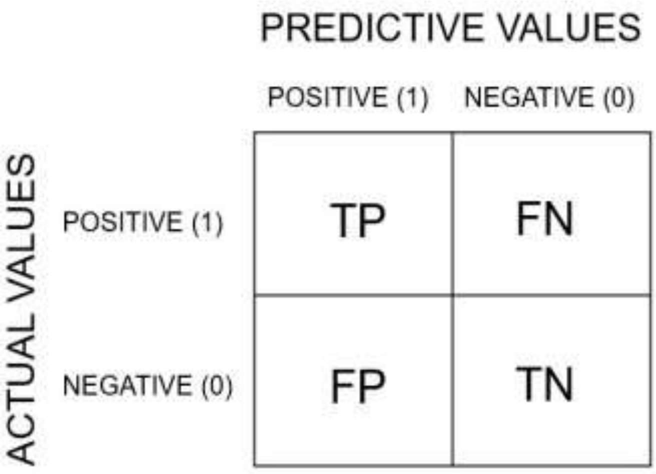


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         562               95.416  %
Incorrectly Classified Instances        27                4.584  %
Kappa statistic                          0.8975
Mean absolute error                      0.0947
Root mean squared error                  0.2093
Relative absolute error                 20.8655 %
Root relative squared error             43.9386 %
Total Number of Instances              589

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.982    0.098    0.950      0.982    0.965      0.899    0.979     0.990     tested_negative
                 0.902    0.018    0.964      0.902    0.932      0.899    0.979     0.933     tested_positive
Weighted Avg.    0.954    0.070    0.954      0.954    0.954      0.899    0.979     0.970

=== Confusion Matrix ===

   a   b   <-- classified as
 377   7 |   a = tested_negative
  20 185 |   b = tested_positive
```

4.2 Confusion matrix

The information about the actual and predicted classification system is held by the Confusion matrix. It demonstrates the accuracy of the solution to a classification problem. Table1 shows the confusion matrix for a binary classifier. The entries in the confusion matrix have the following meaning in the context of our study.

- Tp is the number of correct predictions that an instance is positive.
- Fn is the number of incorrect predictions that an instance is negative.
- Fp is the number of incorrect predictions that an instance is positive and
- Tn is the number of correct predictions that an instance is negative.



		PREDICTIVE VALUES	
		POSITIVE (1)	NEGATIVE (0)
ACTUAL VALUES	POSITIVE (1)	TP	FN
	NEGATIVE (0)	FP	TN

Fig 1. The Confusion Matrix for a two class Classifier

4.3 Precision

Precision is the ratio of correctly predicted true positive events to the total number of predicted positive events. It predicts the percentage of events that are relevant among all the predicted events.

$$Precision = \frac{tp}{tp+fp}$$

4.4 Mean absolute error (MAE)

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - y_i|$$

4.5 Recall/true positive rate/sensitivity:

Sensitivity is the ratio of correctly predicted true positive events to the total number of actual positive events. The recall is also known as sensitivity or true positive rate.

$$Recall = \frac{tp}{tp + fn}$$

4.6 True negative rate/specificity

Specificity is the ratio of correctly predicted true negative events to the total number of actual negative events. Specificity is also known as a True negative rate.

$$Specificity = \frac{tn}{tn+fp}$$

4.7 Accuracy

The ratio of correctly classified samples to the total number of samples is known to be Accuracy (AC). It shows the overall effectiveness of the classifier.

AC = No of correct predictions/ total no of predictions

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

4.8 F1 Score

F1 Score is the harmonic mean of precision and recall. It gives a better measure of the incorrectly classified cases than the Accuracy Metric. F1-score is a better metric when there are imbalanced classes.

$$F1 = \ 2 * \frac{precision \ \times \ recall}{precision \ + \ recall}$$

5. Cross validation

Cross-validation is a statistical method used to find the generalized accuracy of each model. Using this method, we split our dataset into 'n' equal parts. "n-1" parts are used in training while one part is used for testing the data. All possible combinations are used for both testing and training of data. Different accuracy scores are obtained for different combinations of data. The mean of these scores gives a generalized accuracy.

6. Comparison of classification algorithms

Table 2 provides the accuracy of classification algorithms before and after performing feature selection. We have neglected two attributes from the original dataset to increase the accuracy.

Classifier	Accuracy	Modified Accuracy
Support vector machine	**0.77**	**0.78**
Naïve Bayes	**0.74**	**0.74**
Decision tree	**0.67**	**0.69**
K nearest neighbors	**0.75**	**0.75**
Logistic regression	**0.77**	**0.77**

Graphical representation of the above table.

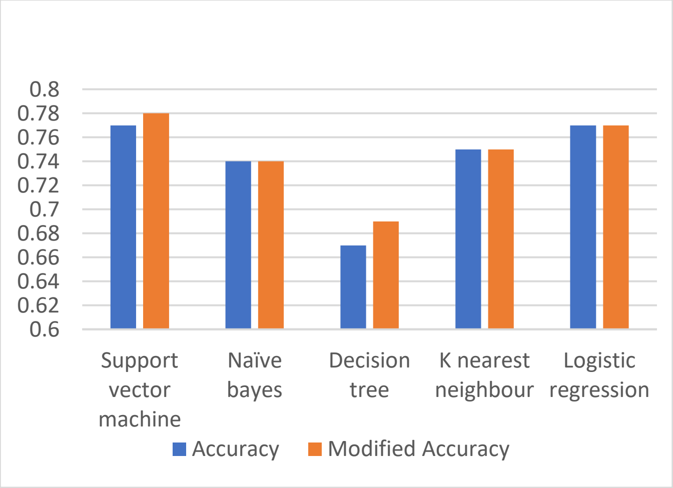
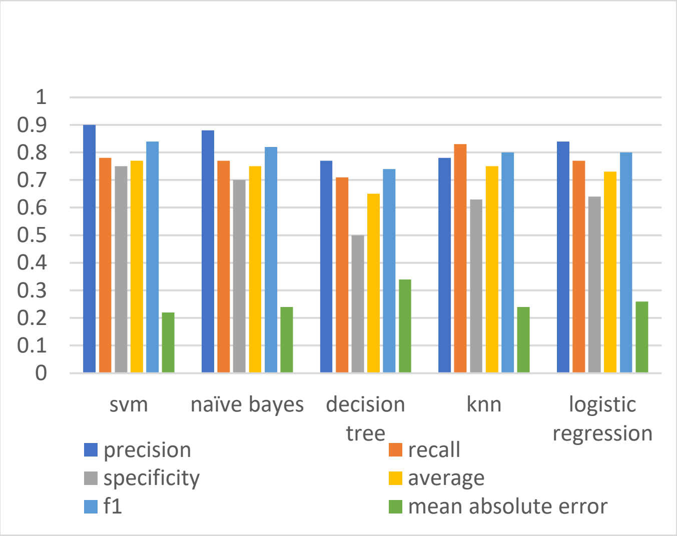


Table 3 provides the statistical information of models in terms of evaluation measures like Precision, Recall, Specificity, Accuracy, and Mean Absolute error.

Algorithm	Precision	Recall	Specificity	Average	F1	Mean Absolute Error
Support Vector Machine	0.90	0.78	0.75	0.77	0.84	0.22
Naïve Bayes	0.88	0.77	0.70	0.75	0.82	0.24
Decision tree	0.77	0.71	0.50	0.65	0.74	0.34
K Nearest Neighbors	0.78	0.83	0.63	0.75	0.80	0.24
Logistic Regression	0.84	0.77	0.64	0.73	0.80	0.26

Graphical representation of the above table gives an insight into the various machine learning models and their predictive accuracy in terms of performance.



7. CONCLUSION

The paper is aimed to provide a model that better classifies the instances of the dataset. Techniques like Data cleaning and Feature selection has helped to improve the potentiality of the dataset. All the Classifiers have achieved an accuracy of above 67%. Cross-validation is performed on each combination to get the mean accuracy of each model. SVM has achieved an accuracy of 78% while Logistic Regression has achieved 77%. SVM and naïve Bayes are comparatively better in terms of evaluation metrics like precision, recall, f1 score, and specificity. SVM has less mean absolute error when compared to other models. By the comparative analysis, we specify SVM as the best model that fits the dataset concerning the diabetic and non-diabetic persons.

ACKNOWLEDGEMENT:

The project team members would like to express our thanks to every individual of the Computer Science Department at Anil Neerukonda Institute of Technology and Sciences for their suggestions and their help with any problems that arose during our work. We want to thank our Assistant Prof. G. Sathar, for his valuable suggestions and guidance for completing our research work. We must admit that he always inspires and motivated us during the completion of the work. We also thank the other teaching staff in the department for their great efforts and for giving us the advice that has helped us to achieve our project's goals.

REFERENCES

[1] World Health Organization Global Report on Diabetes 2019.https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] International Diabetes Federation.
https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html
[3]http://www.healthdata.org/sites/default/files/files/2017_India_State-Level_Disease_Burden_Initiative_- _ Full_Report%5B1%5D.pdf
[4] The increasing burden of diabetes and variations among the states of India: the Global Burden of Disease Study 1990–2016. https://doi.org/10.1016/S2214-109X(18)30387-5
[5] Kumari, V, Chitra, R. Classification of diabetes disease using support vector machine. Int J Eng Res Appl. 2013;3(2):1797-1801.

[6] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, Type 2 diabetes mellitus prediction model based on data mining, Informatics in Medicine Unlocked,Volume 10,2018,Pages 100-107,ISSN 2352-9148,
[7] Sneha, N., Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. *J Big Data* 6, 13 (2019). https://doi.org/10.1186/s40537-019-0175-6

[8] Dutta, Debadri & Paul, Debpriyo & Ghosh, Parthajeet. (2018). Analysing Feature Importances for Diabetes Prediction using Machine Learning. 924-928. 10.1109/IEMCON.2018.8614871.

[9] Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5.

[10] Guo, Gongde & Wang, Hui & Bell, David & Bi, Yaxin. (2004). KNN Model-Based Approach in Classification.

[11] Kaviani, Pouria & Dhotre, Sunita. (2017). Short Survey on Naive Bayes Algorithm. International Journal of Advance Research in Computer Science and Management. 04.

[12] Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.

[13]Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

[14]Vrushali Y Kulkarni,Pradeep K Sinha,Effective Learning and Classification using Random Forest Algorithm, nternational Journal of Engineering and Innovative Technology (IJEIT), Volume 3, Issue 11, May 2014, ISSN: 2277-3754.

[15] Hina S, Shaikh A, Sattar SA. Analyzing diabetes datasets using data mining. J Basic Appl Sci. 2017;13:466–71.

[16] Sa'di S, Maleki A, Hashemi R, Panbechi Z, Chalabi K. Comparison Of Data Mining Algorithms In The Diagnosis Of Type II diabetes. International Journal on Computational Science & Applications (IJCSA) 2015; 5(5).

[17] https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce

[18] Machine Learning Group at the University of Waikato. Weka 3: Data Mining Software in Java. Retrieved September 4,2016, from http://www.cs.waikato.ac.nz/ml/weka/
[19] Pima Indians Diabetes Data Set. http://networkrepository.com/pima-indians-diabetes.php
[20] Zafar, Faizan & Raza, Saad & Khalid, Muhammad & Tahir, Muhammad. (2019). Predictive Analytics in Healthcare for Diabetes Prediction. 253-259. 10.1145/3326172.3326213.