Performance Analysis of Data Mining Classification Algorithm to Predict Diabetes

First Author
Gajendra Sharma
Department of Computer Science & Engineering, Kathmandu University, Dhulikhel, Nepal
Email: gajendra.sharma@ku.edu.np
Second Author
UmeshHengaju
Department of Computer Science & Engineering, Kathmandu University, Dhulikhel, Nepal
Email: hengajumesh@gmail.com

---**ABSTRACT**--

In Data mining, Classification and prediction are the two very essential forms of data analysis. They are widely used for extracting models for describing important data classes. This paper aims in designing classifier models based on five different classification algorithms namely, Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Random Forest and Support Vector Machines (SVM), to classify and predict patients with diabetes. These classifiers are experimented with 10 fold Cross Validation and their performances are evaluated by computing Accuracy, Precision, F-Score, Recall and ROC measures. The test experiment shows that the accuracy given by classifier models developed by using Decision Tree, KNN, Naïve Bayes, SVM and Random Forest are 73.82%, 71.65%, 76.30%, 65.10% and 68.74 % respectively. Similarly, their precisions and recall are 0.705, 0.552, 0.759, 0.424, 0.538 and 0.738, 0.763, 0.82, 0.651, 0.804 respectively. Thus, this study shows that the Naïve Bayes algorithm provides the better accuracy in predicting diabetes as compared to other techniques. And, the data set chosen for this study is "Pima Indian Diabetic Dataset" taken from University of California, Irvine (UCI) Repository of Machine Learning databases.

Keywords: **Data Mining, Diabetes, Classification, Prediction, KNN, Naive Bayes, Random Forest, SVM, Accuracy, Precision, F-Measure, Recall**

I. Introduction

Diabetes mellitus, commonly known as diabetes, is a group of metabolic disorders characterized by high blood sugar levels than the normal level over a prolonged period, which is caused by defective insulin secretion or its impaired biological effects, or both. It is a condition in which human body does not properly process food. Diabetes can lead to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves. Diabetes is categorized mainly into two types namely, Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D). Type 1 diabetes is developed in younger people, mostly less than 30 years old and the typical clinical symptoms for this type of diabetes are increased thirst and frequent urination, high blood glucose levels. This type of diabetes cannot be cured effectively with oral medications alone and the patients are required insulin therapy. Type 2 diabetes occurs more commonly in middle-aged and elderly people, which is often associated with the occurrence of obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases [1].

With the development of living standards, diabetes is increasingly common in people's daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy studying. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels. The earlier diagnosis is obtained, the much easier we can control it. Data mining techniques can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data. This paper presents the study of predicting diabetes using data mining classification algorithm. It mainly focuses on implementing different data mining techniques on diabetes dataset and preparing models for prediction of diabetes disease using supervised learning [2, 12].

Number of data mining classification algorithms has been purposedto classify, predict and diagnose diabetes. But the comparative study of the performance of those algorithms has not been studied effectively. It has not been studied whichamong the available classifier modelscan provide best prediction for diabetes. This paper concentrates on preparing classifier models based on several data mining techniques and implementing these models for predicting diabetes on the available dataset and then makes comparative analysis of the performances of the classifier models in predicting diabetes with better accuracy. The data mining classification algorithms used in this study for designing classifier models are Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Random Forest and Support Vector Machines (SVM) [3].

The respite of the paper is organized as follows: Section 2 briefs Related Work,Section 3 describes the details of chosen dataset, Section 4 describes the methodology and

technique applied on dataset as well as presentsthe performance analysis of algorithms, Section 5 provides the discussion on results of the study and finally Section 6 concludes the study.

II. Related Work

In [5], Pradhan et al. concentrate on developing the classifier model based on ANN, Logistic regression and J48 to diagnose diabetes. According to the researchers, J48 machine learning technique provides efficient and better accuracy among others, for ANN to perform better onset prediction of diabetes, the researcher suggested using the concept of Multilayer Perceptron and Genetic algorithm for feature selection. The researcher did not evaluate classification algorithm using Cross validation evaluation method.

In [6], Orabi et al. designed a system for diabetes prediction, whose main aim is the predict diabetes in a patient. This system is based on the concept of machine learning classification algorithm, namely, decision tree. And the results of this system in predicting diabetes were satisfactory. It contributes an accuracy of 84% in predicting diabetes. Besides this, the system only predicts the diabetes incidents at a particular age. It is not reliable for predicting diabetes in people at different ages.

In [7], Santhanam and Padmavathi implementedK-means and Genetic programmingfor prediction of diabetes. Results achieved using Genetic Programming gives optimal accuracy as compared to K-means. 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.

In [8], Nongyao et al. applied an algorithm which classifies the risk of diabetes mellitus. To fulfill the objective, authors employed four following renowned machine learning classification algorithms namely Decision Tree, Logistic Regression, Random Forest and Naive Bayes. For improving the robustness of designed model, Bagging and Boosting techniques were used. Their experimentation result shows that Random Forest algorithm provides reliable results.

In[9], Kumar et al. used support vector machines(SVM), Logistic Regression, and Naïve Bayes method of classification algorithm along with 10 fold cross validation to predict diabetes. The main aim of this paper is to find out best model from different algorithm that can be used to predict diabetes by applying on the data set of the patients. For this they compared performance of the algorithm based on the accuracy measure and the researchers conclude that Support Vector Machines provides best accuracy than the other algorithm.

In [10], Rashid et al. designed a prediction model with two sub-modules to predict diabetes-chronic disease. ANN (Artificial Neural Network) is used in the first module and FBS (Fasting Blood Sugar) is used in the second module. Decision Tree is used to detect the symptoms of diabetes on patients. The accuracy of the technique vary from before pre-processing and after pre-processing. This indicates in prediction of diseases the pre-processing of data set has its own impact on the performance and accuracy of the prediction.

Data Description

The dataset analyzed in this study is the "Pima Indian Diabetes Dataset" taken from University of California, Irvine (UCI) machine learning repository which is an open source data archive. The link to this dataset is *https://www.kaggle.com/uciml/pima-indians-diabetes-database*.The data set mainly contain 9 attributes and 768 number of instances. The attributes in the data set are named with the variables A1, A2, A3, ……., A9. The first 8 variables specifies various attributes about patients which are submitted into the application and the 9th variable contains the outcome of the application, either positive for diabetesrepresented by "Yes" or negative for diabetesrepresented by "No".

III. Methodology

The methodology implemented for the study is very simple. It is summarized in figure-1 below as a proposed workflow of this study.
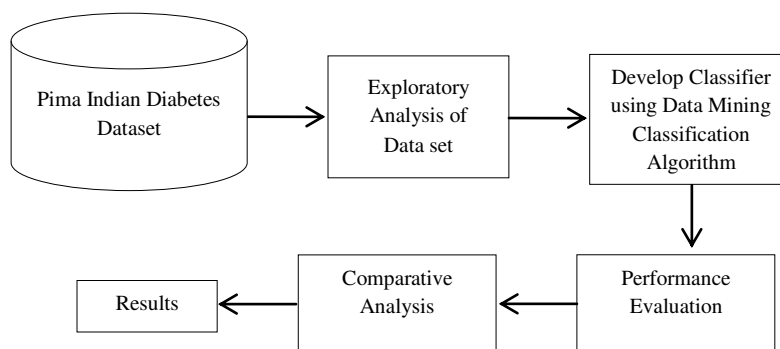

Figure 1: Proposed Workflow

Exploratory Analysis of Dataset
Attributes Naming

From the source of the dataset, it was observed that the names and values of the attributes have been changed to some generic and meaningless symbols to ensure the

confidentiality and privacy of the patients. So as to avoid confusion and for the sake of simplicity, the labels of the variables in our analysis have been assumed to be some working names [6, 11] according to the values in the attributes like A1 changed to Pregnancy, A9 changed to Outcome. Below is the short summary of the changed variable names which are used throughout the analysis.

Table 1: Attribute list from data sets with Names

Variables	Name	Description
A1	Pregnancy	Number of times pregnant
A2	Glucose	Glucose concentration a 2 hours in an oral glucose tolerance test
A3	Pressure	Blood pressure (mm Hg)
A4	Skin	Skin fold thickness (mm)
A5	Insulin	2hrs serum insulin (mu U/ml)
A6	BMI	Body mass index (weight in kg/(height in m)^2)
A7	Pedigree	Diabetes pedigree function
A8	Age	Age (in years)
A9	Outcome	Class variable (Yes or No)

Data Transformation
The data in this analysis contains character values which are transformed to numeric values or binary values. The attribute with name Pregnancy and Age have values in character data type in the original data set. It is transformed into numeric type. Similarly, Outcome attribute have values "Yes" and "No" in original dataset and we made assumption that "Yes" means positive for diabetes then changed to 1 and "No" means negative for diabetes which is changed to 0. They are transmitted from character to binary data type.

Table 2: Data type transformation of attributes

Attributes	Original Data Type	Transformed Data Type
A1/ Pregnancy	character	numeric
A8/ Age	character	numeric
A9/ Outcome	character	binary

Missing Data Treatment
The Pima Indian Diabetes dataset, in the form extracted from the UCI repository, contains missing values over attributes Glucose, Pressure and BMI. These missing values are found in 53 of the 768 cases representing almost 7% of the data. These missing values can be fixed with various approaches such as plug mean/median/ mode values in the place of NA's. In this study, we substitute NA's using prediction method.We used the mice (Multivariate Imputation via ChainedEquations) package in R to approximate missing values of Glucose, Pressure and BMI.

Data Distribution
Data Distribution is used to visualize the value distribution of the attributes. Since the dataset has continuous attributes, visual analysis of those attribute values is conducted to have a sense of the nature of the dataset. Different types of data visualization techniques such as function graph, bar diagram, scatter plot etc. are implemented on the data set to gain idea of the possible relationships between the attributes and to observe visible effect of each factor on whether the medical diagnosis for diabetes test is positive or negative [6].Figure below shows a function graph for the attributes values of the dataset.
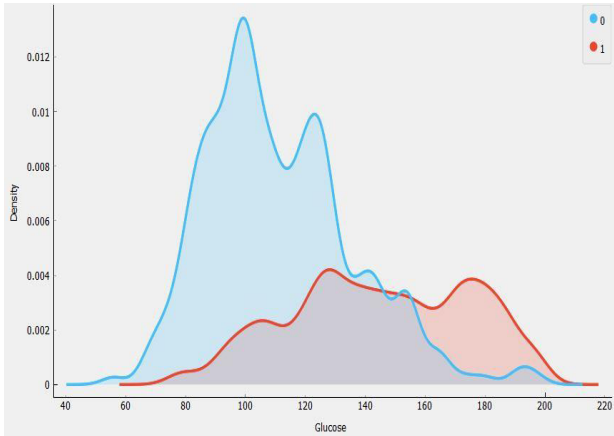
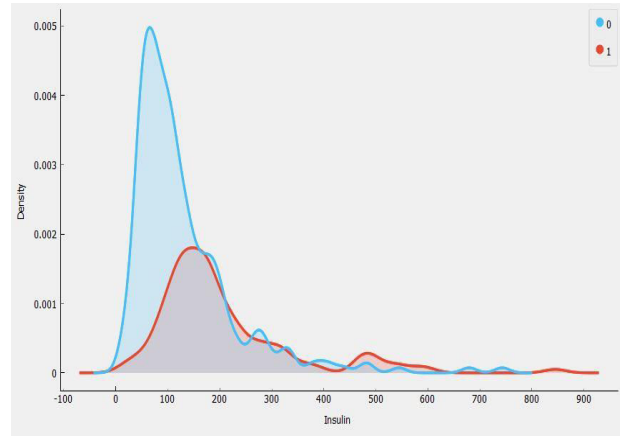Figure 1: Distribution of Glucose
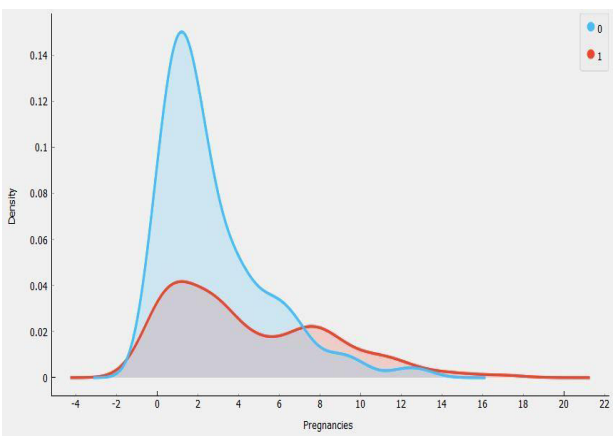

Figure 3: Distribution of Insulin
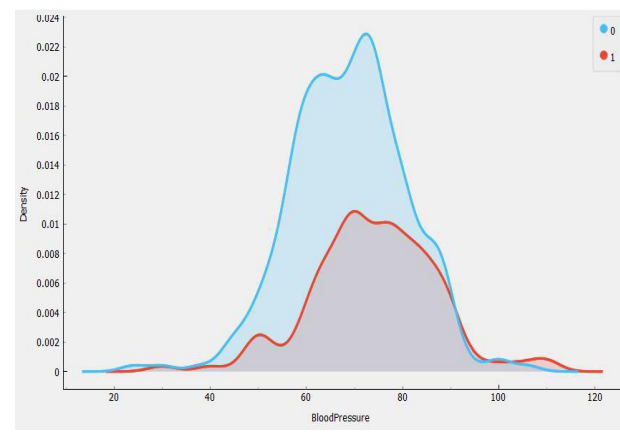

Figure 4: Distribution of Pregnancies
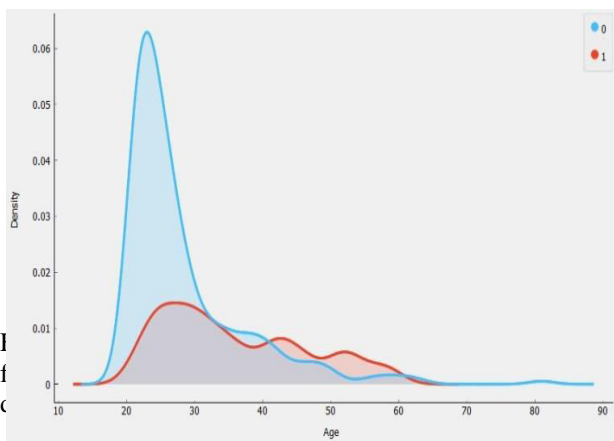

Figure 5: Distribution of Blood Pressure
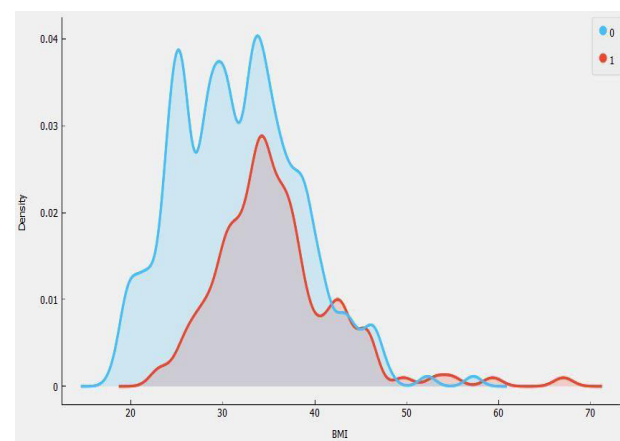

Figure 6: Distribution of Age
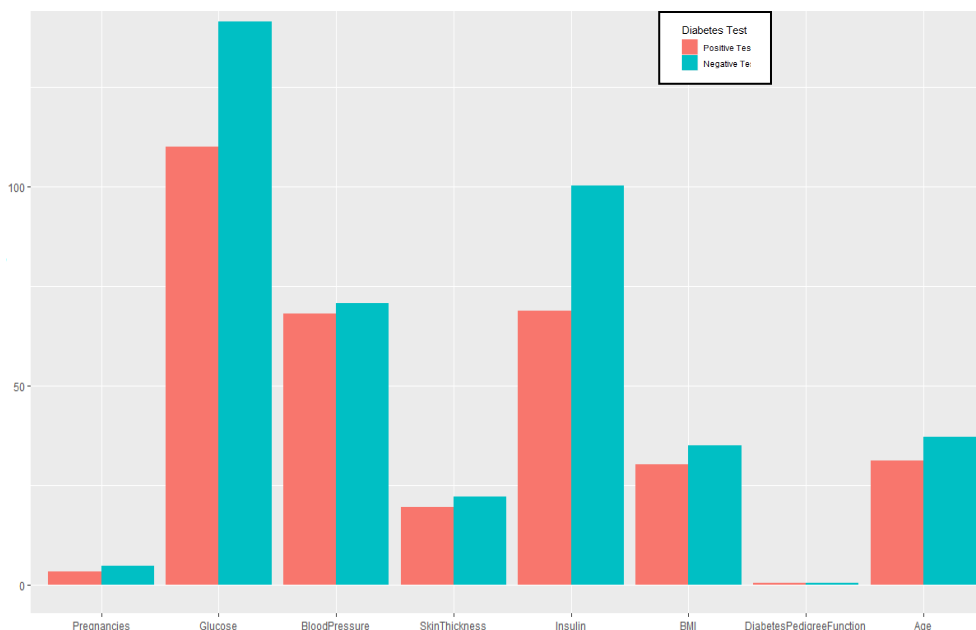

Figure 7: Distribution of BMI

Figure 8: Visualizing positive and negative tests of Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pedigree and Age for diabetes via bar diagram.

Figure below is a scattered plot showing the effects of attributes on the positive and negative test for diabetes. Scatter Plot provides a 2-dimensional visualization for attributes. The data is displayed as a collection of points, each having the value of the x-axis attribute determining the position on the horizontal axis and the value of the y-axis attribute determining the position on the vertical axis.
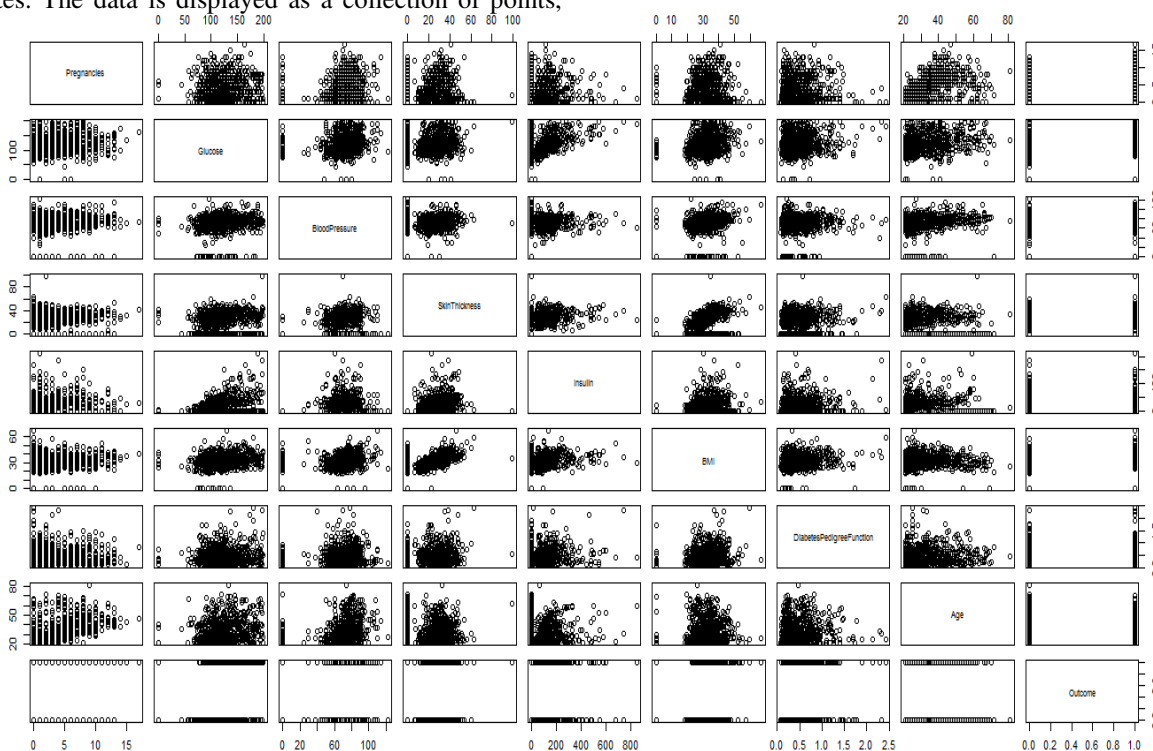

Figure 9: Visualizing the effects of Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pedigree and Age on diabetes via Scatter Plot

In scatter plot, it is observed that the diabetes test seems positive on those patients having higher glucose and Blood Pressure concentration levels and low insulin concentration level. The other attributes are not showing any correlation with the diabetes across various levels so they might not be impacting.

Develop Classifier models using Data Mining Classification Algorithm

In this study, we designed the several classifier models using different data mining classification algorithm namely, Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Random Forest and Support Vector Machines (SVM), to classify the PIMS Diabetes Data set and to predict patients with diabetes.

Decision Tree

Decision tree is one of the simpler data mining technique used in classification, clustering and predictive task. Decision trees are a very effective method of supervised learning. It aims in partitioning dataset into groups as homogeneous as possible in terms of the variable to be predicted. It takes as input a set of classified data, and outputs a tree that resembles to an orientation diagram where each end node (leaf) is a decision (a class) and each non- final node (internal) represents a test. Each leaf represents the decision of belonging to a class of data verifying all tests path from the root to the leaf [4].

The algorithm of Decision Tree that is implemented in this study for developing a classifier model is as follows:

Step 1: Check for base cases

Step 2: For each attribute 'a', implement selection measure to select the best attribute. The three popular Attribute Selection Measures are:

- **Information Gain**

It is defined as the difference between the original information requirement and the new requirement. Mathematically,

$$\text{Gain (A)} = \text{Info (D)} – \text{Info}_A(D)$$

Here, D is data to be partitioned;
$\text{Info (D)} = -\sum_{i=1}^{m} p_i \log_2(p_i)$ is Entropy of D;
$\text{Info}_A(D) = \text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Info }(D_j)$ is expected information required to classify tuple from D based on the partitioning by A.

- **Gain Ratio**

It applies a kind of normalization to information gain using split info valued defined analogously with Info (D). Mathematically,

$$\text{Gain Ratio} = \frac{\text{Gain(A)}}{\text{SplitInfo}_A(D)}$$

Where, $\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2 (\frac{|D_j|}{|D|})$

- **Gini Index**

The Gini Index measures the impurities of D, data partition or set of training tuples. Mathematically,

$$\text{Gini (D)} = 1 – \sum_{i=1}^{m} p_i^2$$

Here, pi = $|c_i$, D|/|D| is probability that an arbitrary tuple in D belongs to class c_i.

Step 3: Choose an attribute having the best score for the measure.

Step 4: It creates a decision node with that attribute.

Step 5: This process is repeated with sub list of the nodes and added to its child node.

Thus, the Decision Tree classifier model predicts the positive or negative test for diabetes by calculating either Information Gain or Gain Ratio or Gini Index. Higher is

K- Nearest Neighbor (KNN)

KNN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). In this, a case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its KNN measured by a distance function. KNN algorithm uses the

the value of these parameters, test for diabetes in patient is positive.

closest data points for estimation. It is analytically tractable and highly adaptive to local information [6].

The algorithm that has been implemented for developing a classifier model using KNN classification technique is as follows:-

Step 1: Each of the new instances is checked with the already available cases, based on Euclidean distance and classified using k value. The Euclidean distance between two points say (x_1,y_1) and (x_2,y_2) is given by;

$$\text{Distance (D)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Step 2: The distance will be less, if the instances are more similar and vice versa.

Step 3: Observe the distance, k -value and instance. Based on these observations instances are assigned to a specific class.

Step 4: The prediction is based on the k-value. So KNN classifier is k-dependent. Here k represents the number of nearest neighbors and for different values of k, outcome may not be the same.

Step 5: Determine the value of k for Pima Indian Diabetic Dataset (PIDD) for classification accuracy.

Thus for predicting positive or negative test of diabetes, KNN classifier model calculates Euclidean distance between the attribute points and determines k values i.e. nearest neighbor. Less is the distance between attribute points, neighbor points are more close to each other and more is the number of closest data points, test for diabetes is positive.

Naïve Bayes
Naive Bayes is one of the classification techniques that rely on Bayes theorem with hypothesis among predictors for individualistic assumptions. It is simple and dynamic algorithm which is particularly useful for classifying very large data sets. It deals with probability theory to classify data [7].

The basic steps used in this study for developing a classifier model using Naive Bayes algorithmis:

Step 1: Convert the data set into a frequency table

Step 2: Create Likelihood table by finding the probabilities

Step 3: Now, use Naive Bayesian equation to calculate posterior probability for each class.

Step 4: The class with the highest posterior probability is the outcome of prediction.

Thus, Naïve Bayes based classifier model predicts the diabetes in patients by calculating the posterior probabilities of the effect of attributes in the given dataset. Higher is the posterior probability, test for diabetes is positive.

Naive Bayes is a strong and powerful predictor. It execute efficiently in the case of absolute input variables contrast to number variables. The visualization distributed for Naive Bayes are very much straightforward to understand. Therefore, it could be used for inventions in predictions for real time model as well. But, Naive Bayes renders considerable outcomes when implemented on categorical data set.It makes assumption of independent predictors, so some time it could provide bad prediction [8].

Support Vector Machines
SVMs are set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize thegeometric margin. So SVM called Maximum Margin Classifiers. SVM is based on the Structural Risk Minimization (SRM) [9].
The basic steps carried out in this study for developing the classifier models using support vector machines are:

Step 1: Identifying the right hyper plane

Step 2: Maximize the distances between neighbor data point

Step 3: Add a feature as $z = x^2 + y^2$ in order to generalize errors.

Step 4: Apply SVM classifier to classify the binary class.

Therefore, SVM classifier maps input vector to a higher dimensional space so that a two parallel hyper planes with a maximal separatingcan be constructedon each side of the hyper plane. The separation of the hyper planes depicts the positive or negative test for diabetes. Larger the margin or distance between these parallel hyper planes the better the generalization error of the classifier will beand better is the efficiency of classifier to predict diabetes .

Random Forest
A random forest is an ensemble learning algorithm. The basic premise of the algorithm is that building a small decision-tree with few features is a computationally cheap process. If we can build many small, weak decision trees in parallel, we can then combine the trees to form a single, strong learner by averaging or taking the majority vote. In practice, random forests are often found to be the most accurate learning algorithms to date. It runs effectively on large volume data sets. It can handle thousands of input variables without variable deletion. Also it has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing [8].

Random forest algorithm has two parts: Tree bagging and from tree bagging to random forest. The algorithm works as follows: for each tree in the forest, we select a bootstrap sample from S where S (i) denotes the ith bootstrap. We then learn a decision-tree using a modified decision-tree learning algorithm. The algorithm is modified as follows: at each node of the tree, instead of examining all possible feature-splits, we randomly select some subset of the features f⊆ F. where F is the set of features. The node then splits on the best feature in f rather than F. In practice f is much, much smaller than F. Deciding on which feature to split is oftentimes the most computationally expensive aspect of decision tree learning. By narrowing the set of features, we drastically speed up the learning of the tree. It generates an internal unbiased estimate of the generalization error as the forest building progresses [10].

Thus, the classifier model built by using random forest classifying technique predicts positive or negative test for diabetes by constructing a multitude of decision trees. Since the dataset contains case of either positive test or negative test for diabetes, 2 cases are sampled at random. This sampling will either grow the tree or shrink it. If the

tree grows to largest extent possible then, the test for diabetes is positive otherwise, the test is negative.

Performance Analysis of Algorithms

To analyze the performances of the classifier models in predicating diabetes in patients, first evaluating criteria has been defined. The different evaluating criteria considered

Table 3: Performance Analysis of classifier models

Algorithm	Accuracy (%)	Precision	Recall	F-Score	ROC
Decision Tree	73.82	0.705	0.738	0.736	0.751
KNN	71.65	0.552	0.763	0.659	0.734
Naïve Bayes	76.30	0.759	0.82	0.760	0.819
SVM	65.10	0.424	0.651	0.513	0.50
Random Forest	68.74	0.538	0.804	0.648	0.728

IV. Results and Discussion

Table 3 shows the accuracy, precision, recall, f-score and ROC of different classifier models to predict

in this study were Accuracy, Precision, F-Measure, Recall and ROC. And 10-fold cross-validation has been conducted for validating the results. The performance of considered classification algorithms over Accuracy, Precision, F-Measure, Recall and ROC measures are shown in the table given below:

positive tests for diabetes in PIMS Diabetes data set. In the table, we can clearly see that the accuracy given by classifier models developed by using Decision Tree, KNN, Naïve Bayes, SVM and Random Forest are 73.82%, 71.65%, 76.30%, 65.10% and 68.74 % respectively. Similarly their precisions and recall were 0.705, 0.552, 0.759, 0.424, 0.538 and 0.738, 0.763, 0.82, 0.651, 0.804 respectively. Thus we can say that the classifierModel designed using Naïve Bayes classification algorithm can provide efficient prediction of Diabetes among others.

The result provided by different classifier model in predicting diabetes can be depicted in bar diagram as follows:
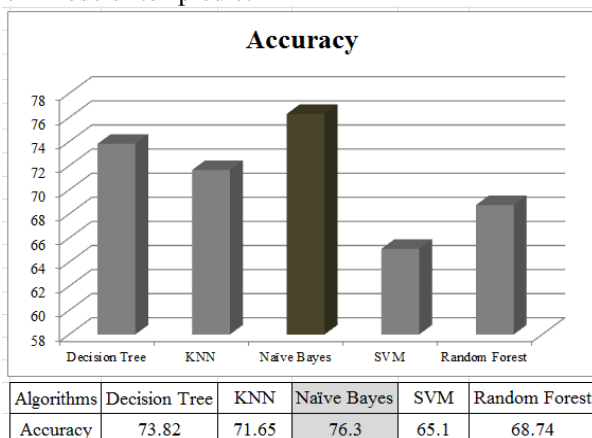

Algorithms	Decision Tree	KNN	Naïve Bayes	SVM	Random Forest
Accuracy	73.82	71.65	76.3	65.1	68.74

Figure 10: Comparison of Accuracy of Decision Tree, KNN, Naïve Bayes, SVM and Random Forest

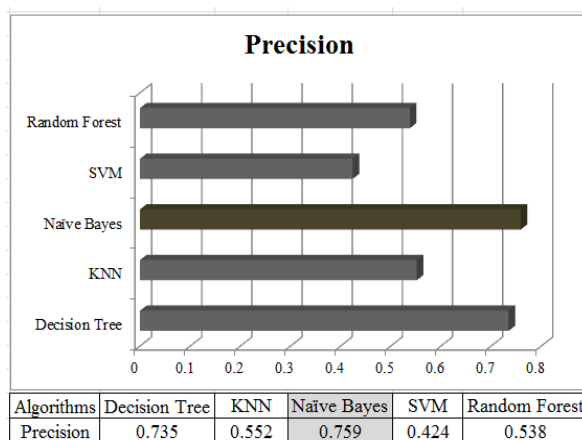| Algorithms	Decision Tree	KNN	Naïve Bayes	SVM	Random Forest
Precision	0.735	0.552	0.759	0.424	0.538

Figure 11: Comparison of Precision of Decision Tree, KNN, Naïve Bayes, SVM and Random Forest
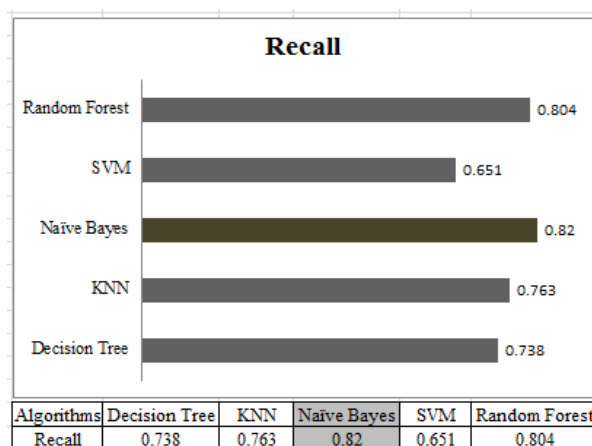
Figure 12: Comparison of Recall of Decision Tree, KNN, Naïve Bayes, SVM and Random Forest

By analyzing the results, it is found that decision tree classifier has ability to break down a complex decision making process into collection of simpler decision, thus providing solution which is easier to interpret. However, Decision tree are also problematic for time-series data values a lot of effort is put into presenting the data in such a way that tends and sequential patterns are made visible. Similarly it is found that KNN based classifier model has higher computational cost and it requires large amount of data set for training. Moreover, it shows poor accuracy rate in multidimensional datasets. Also, SVM suffers from over fitting problem and its computational is inefficiency. And Random Forest based classifier model has problem of balancing error in class population. Because of these limitations, the efficiency of classifier models based on Decision Tree, KNN, SVM and Random Forest algorithms gets degraded in predicting diabetes. But, Naïve Bayes based classifier models are strong and powerful predictors. They solve diagnostic and predictive problems and provide a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. That's why; the Naïve Bayes based classifier model provides better accuracy in predicting diabetes in this study.

V. CONCLUSION

In summary, this study involves comparing the performances of five prediction models for predicting diabetes mellitus using 8 important attributes under two different situations. The prediction models were developed by using five different types of data mining classification algorithms namely, Decision Tree, K-Nearest Neighbors, Naive Bayes, Random Forest and Support Vector Machines. The accuracy of these classifier models for predicting diabetes are 73.82%, 71.65%, 76.30%, 65.10% and 68.74 % respectively. Here the studies conclude that Naïve Bayes classification algorithm based classifier achieves higher accuracy of 76.30 % than other four classifiers.The results have been validated using 10-fold cross-validation procedure. And the data set used in the study was Pima Indian Diabetes Dataset taken from UCI Repository. This study helps in selecting best classifier for predicting diabetes.

REFERENCES

[1].   Han, J., Kamber, M. *Data Mining Concepts and Techniques*, 3rd ed, 2012, 443-491.

[2].   D.Rubben, Jr.Canals, Data mining in health care: current applications and issues, 2009. Available online at: http://mines.humanoriented.com/classes/2010/fall/csci568/papers/Data_Mining_Health.pdf.

[3].   Iyer, A., S, J., Sumbaly, R., diagnosis of diabetes using classification mining techniques, International Journal of Data Mining & Knowledge Management, Process 5, 2015, 1–14.

[4].   Joshi, A., Kaur, R., A review: comparative study of various clustering techniques in data mining, International Journal of Advanced Research in Computer Science and Software Engineering, 3, 2013, 443-491.

[5].   Pradhan, M.,Sahu R. K., Predict the onset of diabetes disease using artificial neural network, International Journal of Computer Science & Emerging Technologies, 2(2), 2011, 303-311.

[6].   Orabi, K.M., Kamal, Y.M., Rabah, T.M., Early predictive system for diabetes mellitus disease, Industrial Conference on Data Mining, Springer, 2016, 420 – 427.

[7].   Santhanam, T., &Padmavathi, M. S., Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis", Procedia Computer Science, 47, 2015, 76-83.

[8]. Nongyao, N., Moungmai, R. Comparison of classifiers for the risk of diabetes prediction, Procedia Computer Science 69, 2015, 132-142.

[9]. Kumar, D.A., Govindasamy, R., Performance and evaluation of classification data mining techniques in diabetes, International Journal of Computer Science and Information Technologies, 6, 2015, 1312–1319.

[10]. Rashid Tarik A., S.M.A., Abdullah, R.M., An Intelligent Approach for Diabetes Classification, Prediction and Description, Advances in Intelligent Systems and Computing,424, 2016, 323–335.

[11]. Sudhakavya B.V., Senthil, S., Classification Algorithm in Data Mining, International Journal of Advanced Networking & Applications, 10(5), 2019, 18-21,

[12]. Deepa, B.G., Senthil S., Singh P. Data Mining on Classifiers Prophecy of Breast Cancer Tissues, International Journal of Advanced Networking & Applications, 10(5), 2019, 8-12.