PERFORMANCE EVALUATION OF VARIOUS SUPERVISED MACHINE LEARNING ALGORITHMS FOR DIABETES PREDICTION

Janhavi R Raut
Research Scholar, Jagdishprasad Jhabarmal Tibrewala University, Rajasthan

Dr. Yogesh Sharma
HOD and Associate Professor, Research coordinator, Department of Computer Science and Engineering

Dr. Vinayak D. Shinde
Head and Associate Professor, Department of Computer Engineering

Abstract- Diabetes is a chronic disease, causes due to increasing high level of sugar in blood and it directly impact on human body organs. When level of sugar increase in blood, body doesn't produce required amount of insulin or sometimes it does not use sufficient available insulin that impact on too much glucose are present in blood. This can cause various health issues that may be dangerous to the human life. Every time person visits to diagnosis center and they spent money for diagnosis of disease. Various Machine Learning algorithms with help of data mining techniques can solve this problem. The data mining techniques play important role in healthcare industries to prediction of disease such as diabetes disease, heart disease, kidney disease etc. The goal of this paper is three machine learning algorithms namely K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Random Forest (RF) is used to for early prediction of diabetes disease. Also, all these three algorithms are compared based on performance metrics.

Keyword: Diabetes, KNN, Support Vector Machine, Random Forest, Data Mining

I. INTRODUCTION

Diabetes Mellitus which is a chronic disease is a globally health issues, millions of people in world are affected by diabetes disease. Diabetes disease cause due to lack of sugar present in the blood. The diabetes disease categorized into two types such as Type I and Type II diabetes. Type I diabetes occur due to body does not produce sufficient insulin in human body. Most of the children are suffering from type I diabetes. Whereas adult is suffering from Type II diabetes which cause due to either body does not produce insulin or body cells not react to the insulin.[8] Therefore, need to diagnosis and early prediction of diabetes disease save the human life and save money. With help of data mining and machine learning algorithm it is possible to find out early prediction of any disease.

Data Mining techniques are important to prediction of diabetes disease. Data mining is an analytical process using huge amount of data it can discover hidden pattern. In banking sector, healthcare industries, education system, customer relationship management data mining plays important role. In healthcare sector, classification, clustering, association, prediction, regression this different data mining techniques are helpful for discover useful pattern from medical dataset to prediction of disease.[4]

Various Machine learning algorithm such as K-Nearest Neighbour, Support Vector Machine, Random Forest, Logistic Regression, Decision Tree, Naïve Bayes are helpful for predicting the disease.

II. LITERATURE REVIEW

Machine learning is used to early prediction of diabetes disease. For experiment purpose KNN, Decision tree, support vector machine, Naïve bayes, random forest and logistic regression algorithms implemented on Pima Indian Diabetes Dataset. KNN is supervised learning algorithm used for classification and regression. Decision tree algorithm predict the targeted value by applying decision rules on data. SVM algorithm employed in classification and regression. It separates two classes with help of hyperplane in training dataset. Random forest algorithm work on constructing multiple decision tree. The model is assessed based on accuracy, recall, f1-measure, misclassification rate and ROC Curve. The paper conclude that Logistic Regression algorithm perform better compared to others algorithm. It gives 77.6% accuracy, 75% f-score, 76% recall rate, 23.8% misclassification rate and 73.6 ROC score [3].

Model developed based on Decision Tree, ANN, Naive Bayes and SVM classification algorithms. Decision tree is easy to understand and accuracy is 74%. ANN gives better prediction in simple data, in this model provide 82% accuracy. Naïve Bayes algorithm is robust and it also handles missing value in dataset and show 80% accuracy. And SVM algorithm is good when we have no idea on dataset attributes. The SVM algorithm provide 82% accuracy [6].

The proposed system uses AdaBoost algorithm for classification with Decision Stump as a base classifier. The Support vector machine, Naïve Bayes and Decision tree these three algorithms also use as a base classifier with AdaBoost algorithms. The result is compared with all algorithms and it shows AdaBoost algorithm with decision stump give highest accuracy which is 80.72% [10].

KNN, SVM, Multilayer Neural Networks Perceptron Artificial (MLP), Classification Naive Bayes (CNB) Classifier and Linear Discriminant Analysis (LDA) this five-classifier comparing to obtain higher accuracy. The result obtain that Multilayer Perceptron Artificial Network show highest accuracy of 96% as compared to others classifiers [5].

Attributes selection method reduces the time of proposed system which is done by Extreme Learning Machine (ELM) algorithm due to its faster learning capabilities. The existing technique compared with proposed system; result prove that proposed system has more efficient than existing system [1].

III. RESEARCH METHODOLOGY

Classifier

A) K-Nearest Neighbour (KNN): KNN is a supervised machine learning algorithm used in classification and regression problem. The algorithm called non-parametric algorithm, because it cannot predict any conclusion based on underlying or training data. The algorithm is also known as lazy learner algorithm, first it stored the dataset and when classification task performs, that time KNN algorithm will be implemented on training set. The KNN algorithm classify a new data on previously stored data basis on similarity. The KNN algorithm used when we have new sample of data then t decides new available sample lie on which categories of data. The closeness of data point is calculated with help of Euclidean distance between data points.

The Euclidean distance between point x and y that is X (x1, x2, …., xn) and Y (y1, y2, …,yn) is defined the equation [4]

$$\mathrm{d}(X, Y) = \sum_{i=1}^{n} (xi - yi)^2$$

B) Support Vector Machine (SVM):

SVM is supervised machine learning algorithm which exist in linear and non- linear form. The algorithm creates a best hyperplane that makes classification of classes which help to put new data point to desired category. SVM support classification as well as regression problem. The SVM uses three types of kernel such as linear kernel, polynomial kernel and gaussian kernel. [2] SVM algorithm used vector point for creating best hyperplane. The linear SVM for separating linearly data whereas non-linear type separate data in non-linear fashion.

C) Random Forest (RF):

Random Forest supervised machine learning classification algorithm. It creates a forest with the help of multiple trees. The performance of algorithm is depending on number of trees in the forest. The Random Forest includes important parameters

1. The method used to split the leaf

2. Which Type of predictor are applied

3. Number of predicted samples in each split [7]

The highest number of trees in forest give highest accuracy of algorithm. Random Forest avoid overfitting problem. The limitation of Random Forest classifier is prediction process is slow because of multiple number of trees available than can affect model becomes slower.

IV. Description about Dataset:

The dataset is taken from UCI repository. [9] The dataset having 768 number of instances available with 9 attributes. In dataset, the first eight attributes define the input value whereas last attribute that is Outcome define whether patient is diabetic positive or negative.

Sr. No.	Attributes	Description
1.	Pregnancies	It represents number of times Pregnant
2.	Glucose	It describes Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3.	Blood Pressure	It shows Diastolic blood pressure of a patient (mm Hg)
4.	Skin Thickness	It represents Triceps skin fold thickness level (mm)
5.	Insulin	It shows 2-Hour serum insulin (mu U/ml)
6.	BMI	It represent Body mass index of patient (weight in kg/(height in m)^2)
7.	Diabetes Pedigree Function	Diabetes pedigree function
8.	Age	It represents age of the patient
9.	Outcome	It shows patient had diabetic or not 0 represent positive and 1 represent negative

Table 1: PIDD attributes and its description

V. Model Evaluation and Analysis:

Comparing machine learning algorithm to evaluated diabetes disease prediction. Performance matric evaluated using correct and incorrect classified instance of the dataset. Total 192 number of instances of training dataset. The KNN algorithm having 71.35% accuracy, SVM showing 73.43% accuracy and Random Forest algorithm give highest accuracy which gives 74.47% accuracy. So, as compared to others algorithms experiment result shows that Random Forest algorithm is best classification techniques.

Performance Measures

Accuracy (A)

Accuracy is a fraction of prediction for correct instance. It is a ratio of correct prediction number to the total number of input present.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ number\ of\ samples}$$

Precision (P)

Precision is the ratio of True Prediction value divided by the number of positive results by the classifier algorithm.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall (R)

Recall is used to measure the classifiers completeness.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F1-Score

F1-Score is harmonic mean of Recall and the precision.

$$F1Score = \frac{2 * (\text{Recall} * \text{Precision})}{Recall + Precision}$$

Result:

Classifier Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
KNN	71.35	87.80	72.97	79.70
SVM	73.43	78.86	79.50	79.18
RF	74.47	80.48	79.83	80.16

Table 2: Performance measure of classifier algorithm

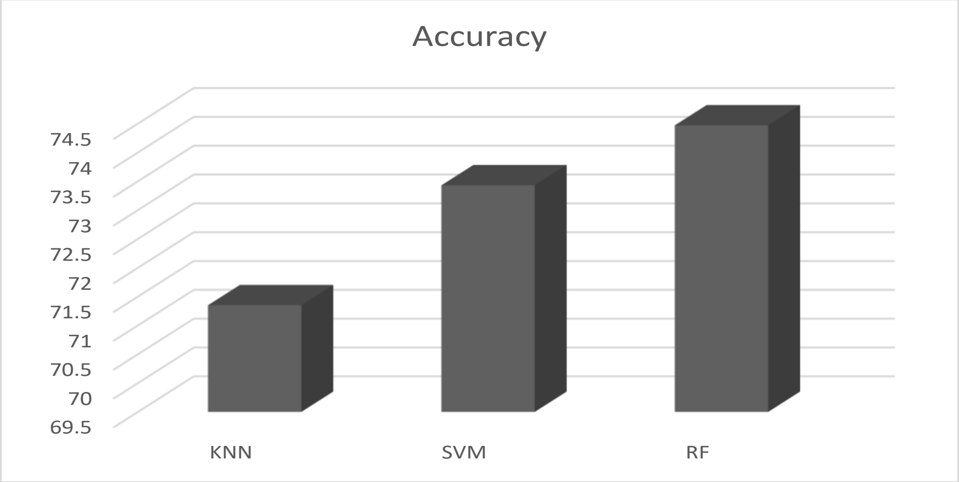

Figure 1: Comparison of algorithm on accuracy

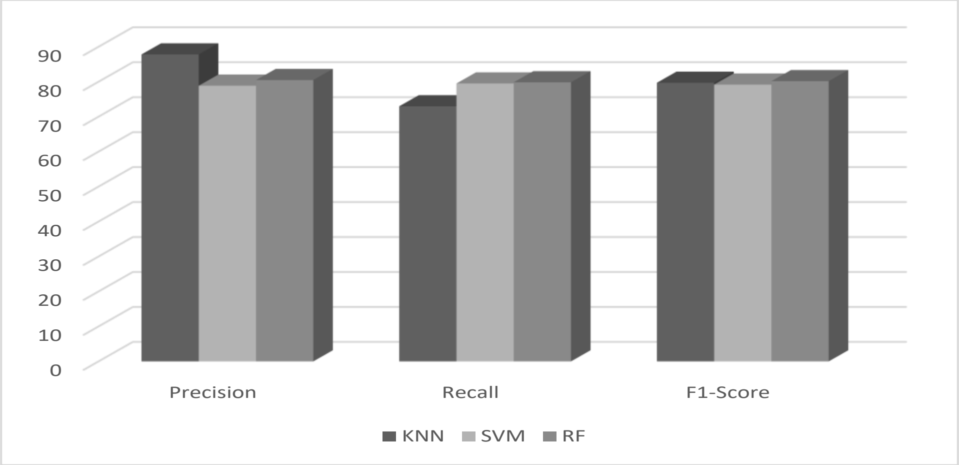

Figure 2: Algorithm comparison on Precision, Recall and F1-Score

VI. CONCLUSION

The experimental study was performing in this paper. Machine learning classification algorithm such as KNN, SVM and Random Forest used for early prediction of diabetes disease. The Pima Indian Diabetes data used for experiment purpose having 768 instance and 8 input attributes which is taken from UCI repository. The experiment result show that Random Forest (RF) algorithms is suitable for classification and early prediction of disease compared to others classifiers. The Random Forest algorithm give highest accuracy of 74.47% with precision of 80.48%, recall 79.83% and 80.16 F-score.

The accuracy of all this algorithm can be improved with help of normalization techniques. The algorithm along with normalization techniques improves the prediction rate of algorithms.

REFERENCES

[1] B. Suvarnamukhi, M. S. (2019). Big Data Processing System for Diabetes Prediction using Machine Learning Technique. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8*(12), 4478-4483.

[2] Debadri Dutta, D. P. (2018). Analysing Feature Importances for Diabetes Prediction using Machine Learning. *IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*.

[3] Jakka, A., & J, V. R. (2019). Performance Evaluation of Machine Learning Models for Diabetes Prediction. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8*(11), 1976-1980.

[4] Krati Saxena, D. Z. (2014). Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm. *International Journal of Computer Science Trends and Technology (IJCST), 2*(4), 36-43.

[5] Mahmood ABED, T. İ. (2019). Comparison between Machine Learning Algorithms in the Predicting the Onset of Diabetes. *International Artificial Intelligence and Data Processing Symposium (IDAP)*.

[6] Priyanka Sonar, P. K. (2019). DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES. *International Conference on Computing Methodologies and Communication (ICCMC 2019)*, 367-371.

[7] Prof. Priya R. Patil, P. S. (2017). Automated Diagnosis of Heart Disease using Random Forest Algorithm. *International Journal of Advance Research, Ideas and Innovations in Technology, 3*(2), 579-589.

[8] S. R. Priyanka Shetty, S. J. (2016). A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique. *I.J. Information Technology and Computer Science*, 26-32.

[9] UCI Machine Learning Repository. http://www.ics.uci.edu/mlearn/MLRepository.html

[10] Veena Vijayan V., A. C. (2015). Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach. *IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 122-127.