iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6138

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Prediction and Analysis of Diabetes Mellitus on Different Features using Machine Learning Algorithm

Akshatha A Nayak[1], Dr. Karuna Pandit[2]
[1]*PG Scholar,* [2]*Professor, Department of MCA, NMAMIT, Nitte,*

Abstract: ***Data Mining is a current trending process in software engineering to analyze the different pattern, correlation and outlier detection in a given data set. Data mining is applicable in various cross points of statistics, database and machine learning. Nowadays, the application of machine learning has increased in all the field, and it is playing a significant role in the medical sector. Using machine learning algorithm diagnosis of the disease is turned out to be more accessible. In this paper, we have carried out a Diabetes Mellitus diagnosis based on a different feature and various machine learning algorithm. The accuracy of multiple algorithms is outlined with the result analysis.***
Keywords: Classifiers, Diabetes Mellitus, Data Mining, Disease Prediction, Machine Learning.

I. INTRODUCTION

In the current era, technology expansion has risen the growth of data in many areas like entertainment, communication, medical field etc. which has highlighted the need for data analysis and processing.

Medical data specifies detailed knowledge associated with health. It is a heap of data regarding the clinical program and patient symptoms and cases.

The scooping and analysis of medical data can solve many health-related issues; it can easy the classification and diagnosis of diseases and promotes faster disease cure.

These days machine learning, convolution neural network and data mining has advanced widely, and these algorithms are used in disease analysis.

In our paper, we have focused on diabetes. Diabetes is also known as diabetes mellitus. It is a disease related to metabolism, results in high sugar level in blood. According to a recent survey, it found 30 million people in India have diabetes. In general diabetes mellitus, is classified into Type 1 diabetes, is also known as an autoimmune disease. It affects pancreas cell, and 10% of people in India face Type 1 diabetes. Type 2 diabetes is also known as pre-diabetes, or gestational diabetes affects in the body resultant to insulin.

Its observed different factors and features affect the sugar level of a human body; some of them holds the primary role. The features are pregnancy, glucose, blood pressure, Insulin, Skin thickness body weight, family background etc. In this paper, we have analysed diabetes occurrence chances with different factor.

The machine learning consists of some well-known algorithm, useful for classification and prediction of diabetes.

1) *KNN Algorithm:* This algorithm is applicable for both regression and classification, it performs based on the proximity of similar things.
2) *Naive Bayes Algorithm:* It is a probabilistic classifier and a collection of the algorithm. It assumes the value of each feature is independent.
3) *Support Vector Machine Algorithm:* Support vector machine is highly recommended supervised algorithm. It classifies the feature and data point based on the hyperplane.
4) *Random Forest Algorithm:* It is a collection of decision tree works well on extensive data set, and each tree is built based upon random vector. The voting system is used to choose a favorite class.
5) *Logistic regression Algorithm:* Logistic regression is a statistical model for a binary dependent variable. It is identical to the linear algorithm. In this, coefficients are combined with an input value to predict the output value.

The objective of our paper is to identify the accuracy of some well-known machine algorithm in the prediction of diabetes mellitus and the diabetes occurrence chances concerning feature. The article is outlined as follows; section 2 explains some existing work on diabetes mellitus in machine learning algorithm, section 3 methodology section 4 discuss result analysis and section 5 conclusion.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

II. SURVEY ON DIABETES MELLITUS PREDICTION

In this section, the study of various existing work on diabetes mellitus using machine and data mining outlined.

K Vijaykumar[1], presented an idea to predict diabetes in the early stage. The work is carried out using a random forest machine-learning algorithm for the classification of diabetes mellitus.

Comparison of the accuracy of the random forest algorithm and other algorithms performed and it is proven random tree forest gives good result.

Amani yahyaoui et al. [2], presented a decision support system for diabetes. They illustrated a machine learning technique using an algorithm like support vector machine and random forest. The designed machine learning technique is compared with other deep learning approaches. Data set of 768 samples with eight features experimented. And the highest accuracy achieved is 83.67% for the random forest algorithm.

Lejla Alic et al. [3], designed a prediction model to predict diabetes. The major work achieved on diabetes type 2. Using the SVM algorithm prediction model is developed. The data training is performed by 10-fold cross-validation technique. The validation accuracy of the system is 84%

Nahal H Bakat et al. [4], illustrated a diabetes prediction system using the SVM algorithm. The significant advantage of the system is, it consists of a black box model as an additional module. The approach has given promising result with 94% accuracy.

Debadri Dutta et al. [5], carried out work to identify elements reason for diabetes. They have analysed different variables and feature that develop the chance of diabetes. They have carried out analysis using various machine learning algorithm like SVM, Logistic regression and Random Forest. It is observed random forest gives a promising result.

Chunguan Huang [6], surveyed SVM algorithm and its efficiency in classifying and predicting diabetes. The primary survey was on bio-heat transfer theory of diabetes and two main features like blood pressure and metabolic function rate. And they observed SVM gives 90% of efficiency.

Gridhar Gopal Ladha et al. [7], surveyed to explore more features and factors used for diabetes prediction. The focus of the paper was to find out the research gap in diabetes prediction.

Bakshi Rohit Prasad et al. [8], presented the preferred data mining technique to identify desirable feature and attribute of diabetes. This approach helps to trace symptoms of diabetes.

The methodology used behind the proposal is the voting technique to pick the most fitting classifier, and this approach gives the encouraging accuracy.

Deeraj Shetty et al. [9], designed a model to help the physician in predicting diabetes. To analyse the attribute of diabetes from a given extensive data set algorithm like KNN and Bayesian are used, and its observed designed approach provides better efficiency.

Shetty S et al. [10], proposed an approach for disease prediction when the data set available is in low condition. They designed a deep learning framework connected with the knowledge base to boost the accuracy of text mining. The approach also holds good for unstructured radiology free-text report.

III. METHODOLOGY

In the proposed methodology, we have adopted the Pima Indians Diabetes dataset. 70% of the data set is applied for validation and training, and 30% of the data set for the testing.

The approach is built using python. The objective of the procedure is to spot a practical algorithm for diabetes mellitus prediction with high accuracy.

Programming of Machine learning Algorithms like K Nearest Algorithm, Naive Bayes, Random Forest, Support Vector Machine and Logistic regression is achieved with features like Pregnancy, Glucose, Blood pressure, Skin Thickness, Insulin, BMI, Pedigree and Age. Each algorithm is tested with a different set of features, and accuracy obtained is noted.

The KNN algorithm is tested with the above mentioned eight features. We have considered a different subset of features. First, we examined the KNN algorithm with individual characteristics and accuracy is noted. Further, we tested the KNN algorithm with two features, three features and so on till eight elements.

The best accuracy obtained is listed and shown in Table 1.

We have examined all other above mentioned machine learning algorithm with the different set of features and best accuracy obtained is listed in Table 2, Table3, Table 4 and Table 5.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Table 1: The KNN Classifier Best Accuracy with Following Features

Total Features	Feature Name	Accuracy Value
1	Pregnancies	67.3177%
2	Glucose and Pedigree	68.8802%
3	Pregnancies, Glucose and Age	71.4844%
4	Pregnancies, Glucose, Insulin and Age	71.0938%
5	Skin thickness, Glucose, BMI, Pedigree and Age	69.7917%
6	Pregnancies, Glucose, BMI, Pedigree, Blood Pressure and Age	69.1406%
7	Pregnancies, Glucose, BMI, Pedigree, Blood Pressure, Age and Insulin	69.2708%
8	Pregnancies, Glucose, BMI, Pedigree, Blood Pressure, Age, Skin Thickness and Insulin	70.1823%

Table 2: The Naive Bayes classifier Best Accuracy with Following Features

Total Features	Feature Name	Accuracy Value
1	Glucose	75%
2	Glucose and BMI	76.4323 %
3	Blood Pressure, Glucose and BMI	76.8229%
4	Pedigree, Glucose, Age and BMI	77.424%
5	Pregnancies, Glucose, Blood Pressure, Pedigree and BMI	77.7344%
6	Pregnancies, Glucose, Blood Pressure, Skin thickness, Pedigree and BMI	77.2135%
7	Pregnancies, Glucose, Blood Pressure, Age, Insulin, Pedigree and BMI	76.8229%
8	Pregnancies, Glucose, BMI, Pedigree, Blood Pressure, Age, Skin Thickness and Insulin	76.3021 %

Table 3: The Random Forest classifier Best Accuracy with Following Features

Total Features	Feature Name	Accuracy Value
1	Glucose	70.7031%
2	Glucose and BMI	72.7865%
3	Pregnancies, Glucose and BMI	73.9583%
4	Pregnancies, Glucose, Blood Pressure and BMI	75.3906%
5	Blood Pressure, Glucose, BMI, Pedigree and Age	75.7813%
6	Pregnancies, Glucose, BMI, Pedigree, Blood Pressure and Age	76.4323%
7	Glucose, Skin thickness, BMI, Pedigree, Pregnancies, Age and Insulin	76.3021%
8	Pregnancies, Glucose, BMI, Pedigree, Blood Pressure, Age, Skin Thickness and Insulin	75.7813%

Table 4: The Support Vector classifier Best Accuracy with Following Features

Total Features	Feature Name	Accuracy Value
1	Glucose	74.7396%
2	Glucose and BMI	75.5208%
3	Pregnancies, Glucose and BMI	76.3021%
4	Pregnancies, Glucose, Blood Pressure and BMI	76.8229%
5	Pregnancies, Glucose, Blood Pressure, Insulin and BMI	77.2135%
6	Pregnancies, Glucose, Blood Pressure, Insulin, Pedigree and BMI	77.424%
7	Pregnancies, Glucose, Blood Pressure, Skin thickness, Insulin, Pedigree and BMI	77.474%
8	Pregnancies, Glucose, BMI, Pedigree, Blood Pressure, Age, Skin Thickness and Insulin	77.3438 %

Table 5: The Logistic Regression classifier Best Accuracy with Following Features

Total Features	Feature Name	Accuracy Value
1	Insulin	77.2135 %
2	Pregnancies and Age	77.2135 %
3	Pedigree, Glucose and BMI	77.0833 %
4	Pedigree, Glucose, Age and BMI	77.7344%
5	Glucose, Insulin, Pedigree, Age and BMI	77.8646%
6	Glucose, Blood Pressure, Insulin, Age , Pedigree and BMI	77.7344%
7	Glucose, Blood Pressure, Skin thickness, Age, Insulin, Pedigree and BMI	77.7344%
8	Pregnancies, Glucose, BMI, Pedigree, Blood Pressure, Age, Skin Thickness and Insulin	77.2135 %

After analyzing different classification algorithm with distinct features, its observed KNN Classifier algorithm will give better accuracy for the combination of three features pregnancy, Glucose and Age. The accuracy achieved is 71.48%. The Naive Bayes Classifier will give exceptional accuracy value of 77.73% for the sequence of five features pregnancy, Glucose, Blood, pressure, pedigree and BMW. The Random Forest Classifier is tested with a combo of six features; pregnancy, Glucose, BMI, Pedigree, Blood, Pressure and Age, and we attain improved accuracy of 76.43%. The support vector machine stands for a combination of seven features Pregnancy, Glucose, Blood Pressure, Skin thickness, Insulin, Pedigree and BMI with promising accuracy of 77.47%. The Logistic Regression algorithm will give the best result for five features Glucose, Insulin, Pedigree, Age and BMI with an accuracy of 77.68%.

The best accuracy value of each algorithm with no. of features is listed in Table 6 with a graphical representation and its noticed Naive Bayes, Logistic Regression, and SVM classifier gives better classification
output and the graphical representation of same is shown in Fig 1

Table 6: Comparative Analysis of Accuracy of Each Algorithm.

Total Features	Feature Name	Accuracy Value	Classifier Algorithm
3	Pregnancies, Glucose and Age	71.4844%	KNN
5	Pregnancies, Glucose, Blood Pressure, Pedigree and BMI	77.7344%	Naïve Bayes
6	Pregnancies, Glucose, BMI, Pedigree, Blood Pressure and Age	76.4323%	Random Forest
7	Pregnancies, Glucose, Blood Pressure, Skin thickness, Insulin, Pedigree and BMI	77.474%	SVM
5	Glucose, Insulin, Pedigree, Age and BMI	77.8646%	Logistic Regression

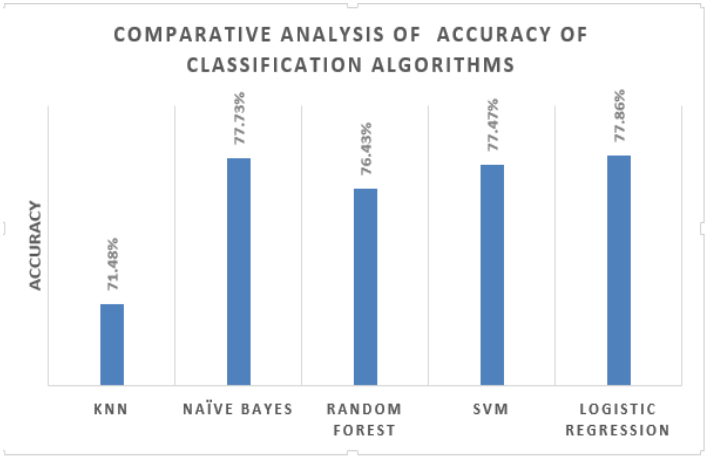


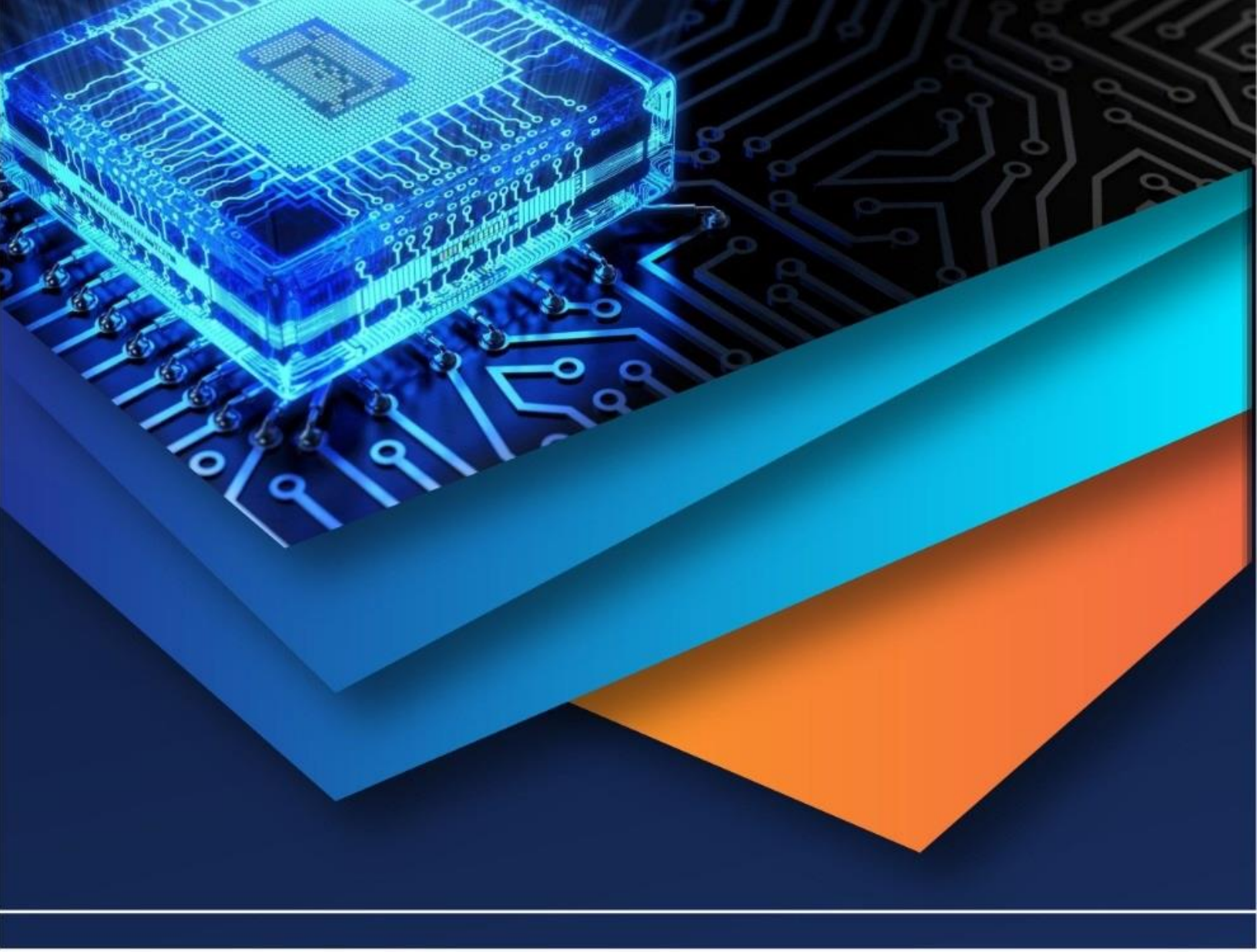
Fig 1: Comparative Analysis of Accuracy of Classification Algorithms

IV. CONCLUSION

In this paper, we tried different machine learning classification algorithms with varying attributes of Diabetes Mellitus to predict the Diabetes Occurrence. It is noticed the accuracy of algorithm increase with the increase in the number of features. Naive Bayes, SVM and Logistic Regression give better efficiency compared to other algorithms. It is analyzed the obtained result is better and need to be improved. In future work, we are planning to design a novel classification algorithm to achieve superior accuracy.

REFERENCES

[1] K. VijiyaKumar, B. Lavanya, I. Nirmala and S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2019, pp. 1-5.
[2] A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 2019, pp. 1-4.
[3] H. Abbas, L. Alic, M. Rios, M. Abdul-Ghani and K. Qaraqe, "Predicting Diabetes in Healthy Population through Machine Learning," 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Cordoba, Spain, 2019, pp. 567-570.
[4] N. Barakat, A. P. Bradley and M. N. H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," in IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 4, pp. 1114-1120, July 2010.
[5] D. Dutta, D. Paul and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, 2018.
[6] C. Huang, G. Jiang, Z. Chen and S. Chen, "The research on evaluation of diabetes metabolic function based on Support Vector Machine," 2010 3rd International Conference on Biomedical Engineering and Informatics, Yantai, 2010, pp. 634-638.
[7] G. G. Ladha and R. Kumar Singh Pippal, "A computation analysis to predict diabetes based on data mining: A review," 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2018, pp. 6-10.
[8] B. R. Prasad and S. Agarwal, "Modeling risk prediction of diabetes — A preventive measure," 2014 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, 2014, pp. 1-6.
[9] D. Shetty, K. Rit, S. Shaikh and N. Patil, "Diabetes disease prediction using data mining," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, 2017, pp. 1-5.
[10] Shetty S., Ananthanarayana V.S., Mahale A. (2020) Medical Knowledge-Based Deep Learning Framework for Disease Prediction on Unstructured Radiology Free-Text Reports Under Low Data Condition. In: Iliadis L., Angelov P., Jayne C., Pimenidis E. (eds) Proceedings of the 21st EANN (Engineering Applications of Neural Networks) 2020 Conference. EANN 2020. Proceedings of the International Neural Networks Society, vol 2. Springer, Cham.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY