Journal of Innovative Engineering and Research (JIER)
Vol.- 4,Issue -2, October 2021, pp. 6-10 (5 pages)

Predicting Diabetes through Machine Learning

MOHAMMED ALI SHAIK[1], DHANRAJ VERMA [1], SANTOSH PAWAR[2], RITESH YADAV[2]

[1] *Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University, Indore 452016, India*
[2]*School of Engineering, Dr. A. P. J. Abdul Kalam University, Indore 452016, India*
Corresponding Author Email: niharali@gmail.com

Abstract— ***Advancement in health science and biotechnology has lead to a significant increase in the production and collection of healthcare data. Thus it is way easier to track the history of the patients and look into their medical records. Machine learning is being used more than ever to study and predict patterns in various diseases. Extensive research is being done on diabetes as it is a common chronic disease and affects millions of people worldwide and this is a cause of concern. This paper focuses primarily on surveying AI approaches used to predict early diabetes, such as decision tree, pruning, random forest, and logistical regression. Initially, the data collection used comes from the National Diabetes and Digestive and Kidney Diseases Institute (NDDKDI). The goal of dataset is to perform analysis of prediction that a patient has infected with diabetes or not.***

Index Terms— Diabetes, Machine Learning, Research, Dataset.

I. INTRODUCTION

In India more than 30 million people have now been diagnosed with diabetes. The CPR (Crude prevalence rate) in India's urban areas is thought to be 9%. In rural areas, the prevalence is around 3% of the total population. India's population now reaches 1000 million: this helps give an idea of the size of the issue. The estimation of the actual number of diabetics in India is about 40 million. IGT (Impaired Glucose Tolerance) is also a growing issue in India. IGT is reported to have a prevalence of about 8.7% in urban areas and 7.9% in rural areas and in rural areas, 7.9%, although this number might be too high. It is estimated that about 35% of IGT sufferers tend to develop type 2 diabetes, hence India is currently facing a healthcare crisis. In India, the type of diabetes is substantially different from that of the Western world. Type 1 is much rarer, and only around 1/3 of type II diabetes are overweight or obese.

Across India, diabetes also begins to occur much earlier in life, suggesting that long-term complications are becoming more universal. There are immense consequences for the Indian healthcare system, which is why diabetes prediction is a much-needed effort to solve the unpredictable and household diabetes problem. This paper survey of AI methods and machine learning methods is performed and used to foresee the diabetes problem. Machine learning consists of 3 different types of learning and are supervised learning, unsupervised learning, and reinforced learning. Unsupervised learning is a form of machine learning which searches for previously undetected patterns in a data set with no pre-existing labels and a minimum of human supervision.

Unlike supervised learning, which typically uses human-labeled data, unsupervised learning, also known as self-organization, makes it possible to model the probability densities over inputs & Reinforcement Training is machine learning model training to make a series of decisions. Within an unknown, potentially challenging world the agent learns to meet a target. An artificial intelligence faces a game like condition in reinforcement learning. The machine uses trial and error to arrive at a solution to the problem. The artificial intelligence obtains either bonuses or punishments for the acts it performs to get the system to do what the programmer wishes. The aim is to optimize overall efficiency. We will be researching supervised learning here, because we have the real outcome of comparing it with the tests and measuring our accuracy.

Supervised learning is the task of learning a function that maps an input to an output based on input-output pairs of examples. It infers a feature consisting of a collection of training examples from the labeled training data. The dataset consists of multiple variables of medical predictors and one target variable originally obtained from “National Diabetes and Digestive and Kidney Diseases Institute (NDDKDI)” [12].

The intent of selecting this dataset determines the patient data who are diagnosed whether the patient is diagnosed with diabetes or not as the evident analytical procedures comprises of dataset. Most of the restrictions are implied for selecting distinct illustrations over a broader record of database where most of the patients are females related to Pima Indian descent who are at least 21 years old as the variables of predictors includes various frequency measures [13].

II. LITERATURE REVIEW

In this Broad efforts were made to identify publications on diabetes research using AI comprises of methods of data mining with two distinct databases comprises of (July 15 2016) which is one of the usually implemented in biomedical sciences with PubMed along with the Bibliography of DBLP Computer Science as it comprises of over 3.5 million of articles along with the conference papers along with other

publications in the computer science area (July 2016). Some of the major motives of using DBLP was to ensure that high-effect international scientific journals are not indexed by PubMed in the field of computer science, although the suggested published methods are applicable to biomedical databases in some cases.

Russell & Norvig[1], published in 1995, focuses on artificial intelligence problem-solving, learning, reasoning, neural networks, Knowledge and other decision-making elements in their seven-part book "Artificial Intelligence: A Modern Approach".

Mehryar Mohri, Afshin Rostamizadeh & Ameet Talwalka [2] are in their 2012 published book "Machine Learning Foundations" provides descriptions of many modern algorithms, provides the theoretical background to these algorithms, and explains key aspects of their implementation. The chapters are primarily devoted to regression, multi-class classification, and ranking.

In their research paper "Machine Learning and Data Mining Methods" that are related to "Diabetes Research" similarly Tsave et al[3] aims to handle by performing a procedural review over the machine learning technologies, data mining techniques and diabetes research tools. A wide variety of machine learning algorithms have been employed. Support vector machines (SVM) emerge as the algorithm most popular and commonly used. Clinical databases were primarily used with respect to data type.

In their paper "Data- Driven simulation and prediction of blood glucose dynamics: Machine learning applications for type 1 diabetes" published in 2019, Woldaregay et al [4] screened and analyzed 417 records and articles based on the inclusion and exclusion criteria, which excluded another 204 papers, leaving 213 papers pertinent. Following a full-text review, there were only 55 papers which were evaluated critically. A Cohen Kappa test was used to assess the inter-rater trust, and disputes were resolved through discussion.

Uddin S., A. Khan[5], In their paper 'Comparing various supervised machine learning algorithms for disease prediction' published in 2019, et al performed a thorough analysis and research and estimated the number of times the algorithm is used for prediction, i.e. its frequency and efficiency performance and, as a result, SVM is most commonly used (29 times) and naive bays (23 times).

In their research paper "Study of diabetes mellitus is used to perform the initial stages of prediction to be implemented over the optimal feature selection process" being published in 2019. N. Sneha and Tarun Gangil [6] carried out empirical work on chronic Diabetes Mellitus and their proposed work analyzes various features that are generated from the datasets which tends to identify the optimal features that are built or generated on the values of correlation as the decision tree algorithm is being implemented over the random forest through the maximum precision value of 98.31% and 97.59% correspondingly to perform the data analysis of diabetic records.

The Support vector machine (SVM) and Naive Bias (NB) techniques has provides with the accuracy of 78.32% and 74.11% percent as per the available system and based on these data the proposed system will enhance the accuracy of the classification techniques considered. Based on these the improvised SVM provides the 78% accuracy and the NB provided with 83% of accuracy as the features will convert the dimensions from higher to lower accurately and effectively.

This provides the data which provides the finest match for identifying the diabetic and non-diabetic patients as the SVM's highest level of disease incidence is estimated at 45.7 percent.

S. Saru, S. Subhashree [7] used WEKA software as a mining tool for the diagnosis of diabetes in their research work "Analysis and Prediction of Diabetes Using ML. " as the proposed approach in this reference will provide with the maximum accuracy along with a precision value of 91.25% as the resultants show that the decision making stipulate us with minimal accuracy than that of other methodologies exist with 83.51% of constant accuracy.

Rajesh K. et al [8] conducted work to identify Diabetes Clinical data and estimate a patient's probability of developing Diabetes. The training dataset used to classify the process of data mining implied over distinct classification methodologies identified using c4.5 technique of classification algorithm which is the most preferred algorithm for classifying the data collection.

III. PROPOSED METHODOLOGY

The Methodology proposed is as follows:

1. Correlation between variables
2. Establishing a Data Tree
3. Error rate of Decision Tree
4. Pruning of the tree
5. Error rate of pruned Tree
6. Random forest technique
7. Error rate of random Forest
8. Logistic regression

A.We start off by determining the correlation between the variables given in the dataset.

It helps to know the relationship between the variables and determine the dependent and independent variable so that we can further proceed with the elements which have more impact on the outcome. Now that we have determined which variables are helpful, we start by dividing the current set of data into two parts 75% for training and 25% to test so that when we check for accuracy it doesn't do it on the same dataset on which the model was trained.

Journal of Innovative Engineering and Research (JIER)
Vol.- 4,Issue -2, October 2021, pp. 6-10 (5 pages)

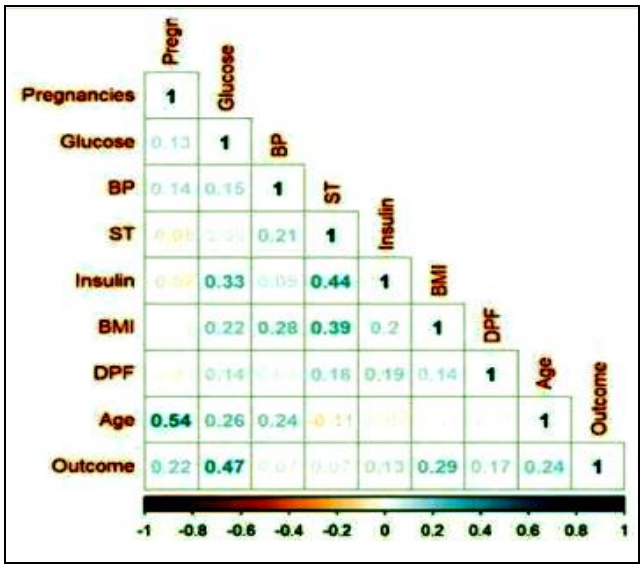


Figure 1: Correlation between the Variables

B. Decision tree

While performing the decision analysis the utilization of decision tree is performed to identify the process of decision-making process which is visually and clearly implementable. Like the name goes, it uses a decision model that resembles a tree. As a commonly used data mining technique to extract a strategy to accomplish a particular objective, it can also be widely used in machine learning. A tree can be "learned" by splitting the source set into sub-sets based on an attribute value check.

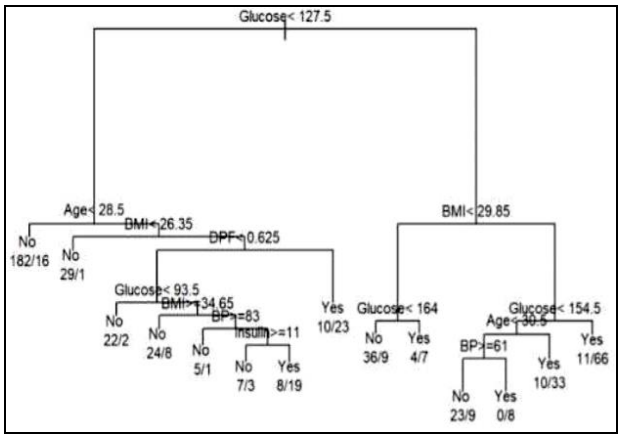


Figure 2: Decision tree

This process is replicated in a recursive fashion identified as the process of recursive partitioning over the process of each of the resultant subset in an iterative process that is accomplished using the subset at a specific node by imparting the similar target variable value that splits the enhanced value of predictions which are not anymore. The construction of a decision tree classifier involves no domain knowledge or setting of parameters and is therefore ideal for the exploration of explorative knowledge. Decision trees can manage data of the large dimensions. Tree classifier has excellent accuracy in general decision taking. Decision tree induction is a standard inductive method for learning classification information.

C. In Figure 2

"Diabetes" appears to be Glucose after the first split criterion for the branch (e.g., Glucose < 127.5). Predict the answer on the test data and construct a confusion matrix that compares the test labels to the test labels expected. The test error rate is 23.96%. In other words, accuracy is 76.04%.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  105  22
##        Yes  24  41
##
##               Accuracy : 0.7604
##                 95% CI : (0.6937, 0.8189)
##    No Information Rate : 0.6719
##    P-Value [Acc > NIR] : 0.00474
##
##                  Kappa : 0.461
##
##  Mcnemar's Test P-Value : 0.88278
##
##             Sensitivity : 0.8140
##             Specificity : 0.6508
##          Pos Pred Value : 0.8268
##          Neg Pred Value : 0.6308
##              Prevalence : 0.6719
##          Detection Rate : 0.5469
##    Detection Prevalence : 0.6615
##       Balanced Accuracy : 0.7324
##
##        'Positive' Class : No
##
```

Figure 3: Error rate in the Decision tree

D. Pruning of Tree

Pruning is a technique used in machine learning and search algorithms that reduces the size of decision trees by eliminating parts of the tree that have little power for instance classification. Pruning reduces the final classifier's complexity, and thus increases predictive accuracy by reducing overfitting.

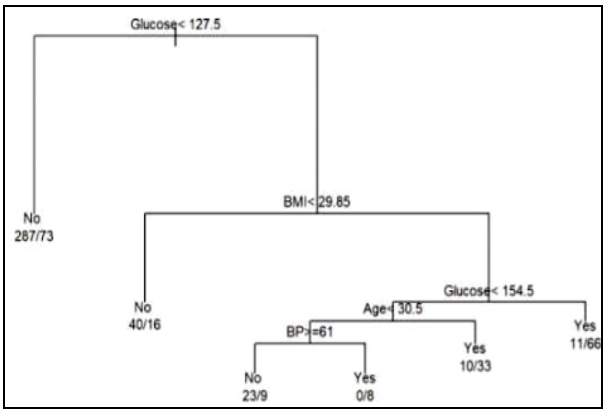


Figure 4: Pruned tree

The redundant branches are removed in order to improve the efficiency and accuracy of the decision tree.

E. Error rate of Pruned Tree

New accuracy is 77.6% which is marginally better than the previous result

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction  No Yes
##        No  118  32
##        Yes  11  31
##
##               Accuracy : 0.776
##                 95% CI : (0.7104, 0.8329)
##    No Information Rate : 0.6719
##    P-Value [Acc > NIR] : 0.001015
##
##                  Kappa : 0.4447
##
##  Mcnemar's Test P-Value : 0.002289
##
##             Sensitivity : 0.9147
##             Specificity : 0.4921
##          Pos Pred Value : 0.7867
##          Neg Pred Value : 0.7381
##              Prevalence : 0.6719
##          Detection Rate : 0.6146
##    Detection Prevalence : 0.7812
##       Balanced Accuracy : 0.7034
##
##        'Positive' Class : No
##
```

Figure 5: The error rate in the new tree

F. Random Forest Technique

The most ensemble learning technique of classification or regression is the random forests or the random decision forests are primarily implementing multitude over the decision trees at a specific training time to generate the classes over a specific class mode by performing classification or to predict the mean value by implementing the aspect of regression over distinctive trees to overfit to their training collection.

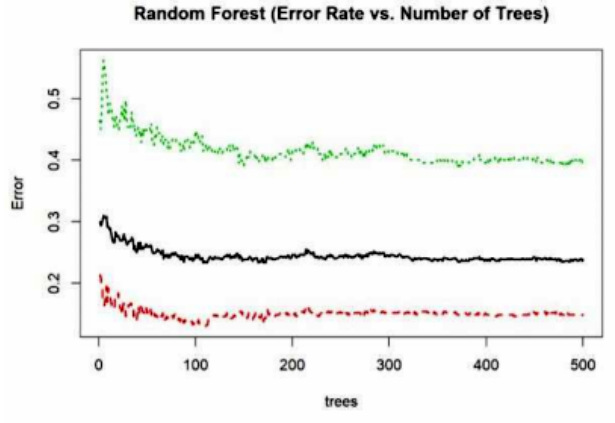


Figure 6: Random Forest graph

G. Error Rate of Random Forest Tree

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction  No Yes
##        No  114  23
##        Yes  15  40
##
##               Accuracy : 0.8021
##                 95% CI : (0.7386, 0.856)
##    No Information Rate : 0.6719
##    P-Value [Acc > NIR] : 4.405e-05
##
##                  Kappa : 0.5361
##
##  Mcnemar's Test P-Value : 0.2561
##
##             Sensitivity : 0.8837
##             Specificity : 0.6349
##          Pos Pred Value : 0.8321
##          Neg Pred Value : 0.7273
##              Prevalence : 0.6719
##          Detection Rate : 0.5938
##    Detection Prevalence : 0.7135
##       Balanced Accuracy : 0.7593
##
##        'Positive' Class : No
##
```

Figure 7: Error rate of Random Forest

H. Logistic Regression

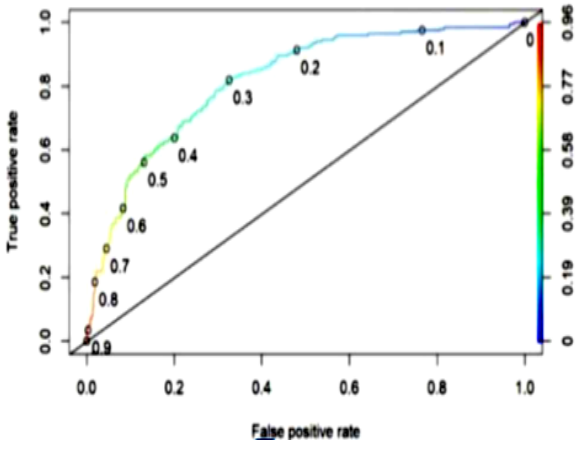
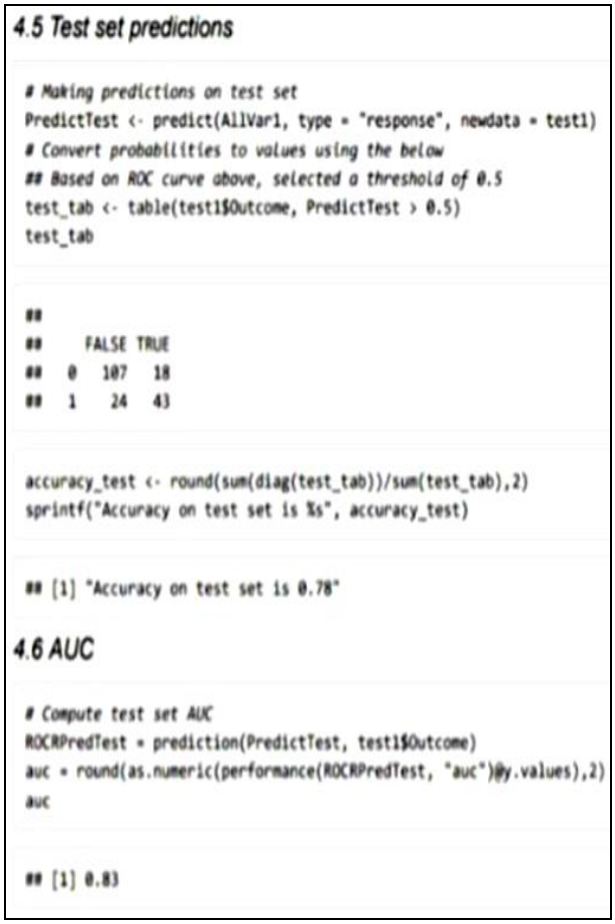


Figure 8: ROC curve

The Receiver Operating Characteristic Curve is a graph that displays binary classifier system's diagnostic capacity as the threshold is varied to obtain the best possible performance.



4.5 Test set predictions

```
# Making predictions on test set
PredictTest <- predict(AllVar1, type = "response", newdata = test1)
# Convert probabilities to values using the below
## Based on ROC curve above, selected a threshold of 0.5
test_tab <- table(test1$Outcome, PredictTest > 0.5)
test_tab
```

```
##
##     FALSE TRUE
##   0   107   18
##   1    24   43
```

```
accuracy_test <- round(sum(diag(test_tab))/sum(test_tab),2)
sprintf("Accuracy on test set is %s", accuracy_test)
```

```
## [1] "Accuracy on test set is 0.78"
```

4.6 AUC

```
# Compute test set AUC
ROCRPredTest = prediction(PredictTest, test1$Outcome)
auc = round(as.numeric(performance(ROCRPredTest, "auc")@y.values),2)
auc
```

```
## [1] 0.83
```

Figure 9: Accuracy of Logistic Regression

IV. CONCLUSION

With the survey of all the methods applied in the project the logistic regression is the best mathematical model that can be used as shown above to predict diabetes with precision of at least 83 per cent.

REFERENCES

[1] K. Kalpakis, D. Gada, and V. Puttagunta. "Distance measures for effective clustering of arima time-series". In Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, California, USA, pp. 273–280, 2001.

[2] P. Praveen, B. Rama and T. Sampath Kumar, "An efficient clustering algorithm of minimum Spanning Tree," 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics , Chennai, pp. 131-135.

[3] Mohamed A. Mahfouz, d M. A. Ismail. "Fuzzy Relatives of the CLARANS Algorithm With Application to Text Clustering". International Journal of Electrical and Computer Engineering. 2009 370-377.

[4] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim. "ROCK: A Robust Clustering Algorithm for Categorical Attributes". In: Proceedings of the 15th international Conference on Data Engineering, pp. 512-521, 1999.

[5] Mohammed Ali Shaik, Dhanraj Verma, Agent-MB-DivClues: Multi Agent Mean based Divisive Clustering, Ilkogretim Online - Elementary Education, Vol. 20 (5), pp. 5597-5603.

[6] Mohammed Ali Shaik and Dhanraj Verma, Enhanced ANN training model to smooth and time series forecast IOP Conf Ser: Mater Sci Eng Vol. 981, 022038

[7] Mohammed Ali Shaik Dhanraj Verma P Praveen K Ranganath and Bonthala Prabhanjan Yadav, RNN based prediction of spatiotemporal data mining IOP Conf Ser: Mater Sci Eng Vol. 981, 022027

[8] Mohammed Ali Shaik and Dhanraj Verma, Deep learning time series to forecast COVID-19 active cases in INDIA: a comparative study IOP Conf Ser: Mater Sci Eng Vol. 981, 022041

[9] Mohammed Ali Shaik, Time Series Forecasting using Vector quantization International Journal of Advanced Science and Technology IJAST Vol. 29 (4), pp. 169-175, 2020.

[10] Mohammed Ali Shaik, A Survey on Text Classification methods through Machine Learning Methods International Journal of Control and Automation, Vol. 12 (6), pp. 390-396, 2019.

[11] Mohammed Ali Shaik T Sampath Kumar P Praveen R Vijayaprakash, Research on Multi-Agent Experiment in Clustering International Journal of Recent Technology and Engineering, Vol. 8 (1S4) pp. 1126-1129, 2019.