International Journal of Multidisciplinary Engineering in Current Research
ISSN: 2456-4265, Volume 7, Issue 2, February 2022, http://ijmec.com/

COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS IN PREDICTING EARLY RISK OF DIABETES.

[1] ***R. V. Subbaiah***, *Professor,Department of CSE,Rise krishna Sai Prakasam groups of Institutions,Valluru.*
[2] ***Uma Devarakonda***, *Asst.Professor, Department of CSE, Rise krishna Sai Gandi groups of Institutions,Valluru.*
[3] ***Dr.Bharathi Devi Matta***,*Assoc.Professor,Department of CSE,Rise krishna Sai Gandi groups of Institutions,Valluru.*

***Abstract*—Diabetes is one of the most occurring diseases at present. Irrespective of age and gender, many people are endan- gered to suffer from diabetes. Diabetes is a medical condition in which the sugar levels in the blood rise beyond limits. When people eat food, glucose is the source of energy for the cells of the human body. However, if the chemical called insulin is not released by the pancreas, the levels of glucose in the human body rise and affect the metabolism of the human body. Therefore, it is highly important to predict diabetes in its early stages and takes steps to prevent it. This is where Machine Learning comes into play. Machine Learning is known to be highly useful in cases where prediction is key to the solution. In the case of predicting diabetes, the system analyses feature such as sudden weight loss, genital thrush, and delayed healing. In this project work, four classification machine learning algorithms are trained and tested using an open-source dataset collected from the internet, and their performance is evaluated and compared.**

***Index Terms*—Accuracy, Decision Tree, Logistic Regression, Random Forest, Support Vector Machine.**

I. INTRODUCTION

At present, many people suffer from diabetes due to their lifestyles. Diabetes is a chronic disease where the pancreas in the human body does not release insulin or the human body does not use it in the right way. People suffering from diabetes experience various symptoms such as excess urination, weight loss, tiredness, and slower healing. According to World Health Organisation, people suffering from diabetes rose from 108 million in 1980 to 422 million in 2014 and there has been an increase of 3% in the mortality rate between 2000 and 2009. The most important aspect regarding diabetes is that people irrespective of age are suffering from it.

Motivation: As mentioned, the need to predict the occurrence of diabetes is very important. Technology on the other hand is being used for all sorts. Machine learning can be highly useful in the case where something is to be predicted using past data. In the prediction of diabetes, machine learning can be used to predict the occurrence of diabetes by analyzing the various features collected through surveys or medical analyses. In this project, features such as Age, Polyuria, Polydipsia, sudden weight loss, genital thrush, delayed healing, and obesity are used to predict if the

person has the risk of diabetes. Therefore, we have decided to conduct a thorough literature review to make sure that the algorithm that we choose has accurate results.

In this research, we want to compare a few classification algorithms that are the most suitable for predicting the early risk of Diabetes. So the first part of the paper discusses the literature review conducted to select the

models for comparison. The later parts discuss how the models are built and implemented. The final section is the results and the analysis of the outcomes computed.

. BACKGROUND

Before getting deep, let us explain a few terms on whichthis project is based on.

A. *Supervised machine learning*

One of the three types of machine learning categories can be described as machine learning knowledge used to predict a data value by using an existing data value. As this project's aim is to classify if a person is at risk of diabetes, this prediction problem can be called a classification problem. Mathematically, it can be described as mapping a data variable 'p' to a class 'q'.

B. *Support Vector Machine*

Abbreviated as SVM is one of the widely used machine learning algorithms. It is proven by many researchers and programmers that the support vector classifier is a great way to solve classification problems. The SVM model generates a hyperplane that is the best fit for all the data variables and classifies data on a multi-dimensional hyperplane. [9]

C. *Random Forest*

It is a supervised machine-learning algorithm that is usedfor both classification and regression problems. As the name suggests, it is an inclusion of multiple decision trees that are implemented by considering a few factors. The classification outcomes of the implemented decision trees are then voted
to finally classify the data variable. This is done to make thealgorithm more accurate and robust [2].

D. *Logistic Regression*

Although the name logistic regression, it can also be used for many classification problems. It tries to draw a line, linear or curve, that classifies the data into separate classes. It generates a logistic function that maps the relationship between data and the predicted data. This function is also used to calculate the probability of a data variable belongingto a certain class [5].

E. *Decision Tree*

A type of supervised machine learning algorithm which is analogous to a flow chart-like structure that is a series of decisions made based on a few factors and finally reaching a conclusion. It is very useful in classification problems because the main idea of decision trees is decision-making starting from the root node to the child node that can becalled a class [6].

III. RELATED WORK

As already mentioned, we have done a literature review regarding the machine learning algorithms that can be usedto accurately predict the risk of diabetes for a person. The below-mentioned papers were highly related to our research.

Deepthi Sisodia et al, 2018, have done a comparative analysis of classification algorithms for the early-stage risk prediction of diabetes [8]. In their research, they have compared classification algorithms such as the Naive Bayes algorithm, Decision tree, and Support vector machine. Upon analysis and comparing the models using performance metrics such as accuracy score, recall, F1 measure, and precision, they have concluded that the decision tree and Naive Bayes algorithm have achieved almost similar performance with naive Bayes having a bit more precision.

Jodeba Jamal Khanam et al, 2021, have done a comparativeanalysis of multiple classifications, regression, and Neural Networks models in the prediction of diabetes [4]. In their research, they have compared models such as

International Journal of Multidisciplinary Engineering in Current Research
ISSN: 2456-4265, Volume 7, Issue 2, February 2022, http://ijmec.com/

Decision trees, Logistic Regression, Support Vector Machines, Random Forests, K-nearest Neighbours, and Artificial neural networks (ANN). They concluded their work by saying that Logistic regression and SVM are the most suitable algorithms for the problem as they achieved an accuracy of 78.8% and 78.2% respectively.

Jingyu Xue et al, 2020, have published a research paper that compares machine learning models for early-stage diabetes prediction [10]. In their study, they compared classification algorithms such as the Naive Bayes classifier, SVM, and
LightGBM using a direct questionnaire survey conducted at Sylhet Diabetes Hospital in Sylhet. Their paper concluded by mentioning that SVM achieved the highest accuracy of 96.54% and is better than the other two in the early-stage diabetes prediction problem.

S. Saru et al, 2019, have done research in which they compared classification algorithms with hybrid combinations such as Logistic regression with SVM, Decision Tree,KNN (k=1), and KNN (k=3) [7]. They have compared the performance of the algorithms by comparing their accuracy before and after bootstrapping. Finally, they concluded bysaying that the Decision tree got an accuracy of 78.43% before bootstrapping and 94.4% after bootstrapping which are higher when compared to other considered algorithms.

A. Models selected

After performing the literature review, we have decided to compare the classification algorithms Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest because these algorithms were frequently compared in most of the classification problems and are competitive in accuracy battles.

IV. METHOD

Let us discuss the experimentation part of the project. Inthis section, we describe how we preprocess the data, and implement the classifiers for the considered machine learning algorithms and performance metrics.

A. Data

The dataset used in the research is collected from Kaggle and is ”diabetes.csv”. The dataset contains 768 instances and9 attributes. All the attributes are numerical type and did not require any data type conversion.

B. Data Preprocessing

Data Preprocessing is said to be one of the most important steps in Machine Learning. It is a step that is used to makesure that accuracy does not decrease due to mistakes in the dataset. In the data-cleaning step, null values or missing values in the dataset are corrected by using methods suchas imputations or deleting the entire row or column [3]. Fortunately, the dataset that we have collected has no missing values or null values. So, the dataset is clean and ready to be used to train machine learning algorithms.

C. Data Scaling

Data Scaling is a data manipulation technique that is used to scale all the numerical data into a specific range using the concept of normalization. For our implementation, wehave used MaxMinScaler to scale the data values into a
range of 0 to 1. Data Scaling is performed to make sure that large numbers do not affect the accuracy of any algorithms because there are algorithms such as SVM that are distance dependent. Therefore, it is important to normalize and scalethe data.

D. Implementing the machine learning models

As the data is preprocessed, it can be used to train and test the data. We have split the data into 80:20,

International Journal of Multidisciplinary Engineering in Current Research
ISSN: 2456-4265, Volume 7, Issue 2, February 2022, http://ijmec.com/

where 80% of the data is used to train the models and 20% ofthe data is used to test or validate the performance of the models. The classifiers for the considered algorithms are also implemented to perform classification. Hyper-parameter tuning is performed to increase the performance of the models.

It is also very important to find the best estimators in most classification problems because it ensures the accuracy of the classifier is at its best. Before finding the best estimators, it is also a basic step to define a parameter grid for each classifier. Parameter grip is a way of defining parameters for the classifier being implemented. The best estimator can be found using an inbuilt attribute provided by the sci-kit library called the "GridSearchCV". The best estimator can be displayed by printing the "grid name.best estimator " [1].

The best estimators for the machine learning models were:

1) SVM - (C=10, kernel='linear')2) LR - (c=10)
3) RF-(max depth=10,max features='auto', min samples leaf=4,min samples split=5,n estimators=300)
4) DT - (max depth=5, min samples leaf=2)

Therefore, the models are hence implemented and ready to be compared. The performances of the machine learning models can be compared using various metrics called "Performance Metrics". There are various performance metrics such as Accuracy score, F1 measure, Recall, Area under the curve (ROC), and Confusion matrix. We have chosen the accuracy score and F1 measure to compare the performance of the models.

E. *Performance Metrics*

As mentioned, we chose to use the accuracy score and F1 measure as our performance metrics. Below is a briefdefinition and mathematical description of the two.

1) Accuracy - it is the ratio of correct prediction to total predictions. It is mathematically written as

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

where

a) TP - True Positive, the total number of correct predictions to a certain class.
b) TN - True negative, the total number of correct predictions to another class
c) FP - False Positive, the total number of incorrect predictions
d) FN - False Negative, the total number of correct prediction

2) F1 score - it is the harmonic mean of recall and precision. It is mathematically written as

$$F1\ score = 2\ (precision * recall)/(precision + recall) \quad (1)$$

where

a) Precision - the ratio of True positives to totalpredictions.
b) Recall - the ratio of True positives to actual posi-tives.

V. RESULTS AND ANALYSIS

Let us analyze the performance metrics and compare the performance of the models implemented.

Upon computing the performance of all the algorithms, we can compare the algorithms by comparing their accuracies. SVM got an accuracy and F1 score of 75.34%. Logistic Regression got an accuracy of 75.9% and an F1 score of 0.6666. Random Forest got an accuracy of 75.97% and an F1 score of 0.679. Decision Tree got the highest accuracyof 78.57% and an F1 score of 0.673. After analyzing the results, Decision Tree with hyper-tuned parameters got the best results and is the most suitable machine learning classification technique to predict the early stage diabetes risk.

International Journal of Multidisciplinary Engineering in Current Research
ISSN: 2456-4265, Volume 7, Issue 2, February 2022, http://ijmec.com/

Model	Accuracy Score	F1 Score
SVM	75.3%	0.753
LR	75.9%	0.666
RF	75.8%	0.627
DT	78.5%	0.673

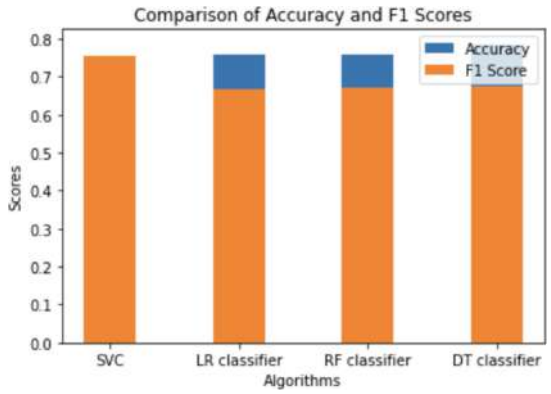

Fig. 1. Comparison between Algorithms

VI. CONCLUSION

The main aim of this research is to analyze and compare a few machine learning models that are the most suitable for predicting early-stage diabetes using various factors. A data set "diabetes.csv" has been collected from Kaggle. A literature review has been conducted for model selection and algorithms like SVM, Logistic Regression, Random Forest, and Decision Tree have been selected. The algorithms are trained and models and their classifiers have been built. Performance metrics accuracy and F1 score have been used to evaluate the performance of the models. Upon performance evaluation, we conclude that Decision is the most suitable solution for the problem as it achieved an accuracy of 78.5%.

VII. CONTRIBUTION

Both of us have worked together equally. We have together performed the Literature review and experimentation. Both of us equally shared the report writing part and completed the project without any issues.

REFERENCES

[1] "sklearn.model selection. Grid Search CV."[Online].Available: https://scikit-learn/stable/modules/generated/sklearn.model*selection.GridSearchCV.html*

[2] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp.197–227, 2016.

[3] I. F. Ilyas and X. Chu, *Data cleaning*. Morgan & Claypool, 2019.

[4] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021.

[5] T. G. Nick and K. M. Campbell, "Logistic regression," *Topics in biostatistics*,pp. 273–301, 2007.

[6] J. R. Quinlan, "Learning decision tree classifiers," *ACM Computing Surveys (CSUR)*, vol. 28, no. 1, pp. 71–72, 1996.

[7] S. Saru and S. Subashree, "Analysis and prediction of diabetes using machine learning," *International journal of emerging technology and innovative engineering*, vol. 5, no. 4, 2019.

[8] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, vol. 132, pp. 1578–1585, 2018.

[9] S. Vishwanathan and M. Narasimha Murty, "Ssvm: a simple svm algorithm," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, vol. 3, 2002, pp. 2393–2398vol.3.

[10] J. Xue, F. Min, and F. Ma, "Research on diabetes prediction method based on machine learning," in *Journal of Physics: Conference Series*, vol. 1684, no. 1. IOP Publishing, 2020, p. 012062.