Ilkogretim Online - Elementary Education Online, 2020; Vol 19 (Issue 4): pp.4214-4223
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2020.04.764828

Statistical Analysis and Prediction of Diabetes Disease Using Machine Learning Algorithms

Saddam Hussain, Shanxi University of Finance and Economics, china, saddamhussain.stat885@gmail.com
Saira Raiz, National College of Business Administration and Economics, Pakistan, ajmscs@yahoo.com
Anum Iftikhar, Shanxi University of Finance and Economics, china, anumiftikhar885@gmail.com

Abstract: Diabetes is a chronic condition or a series of metabolic disorders where a person hurts from a higher blood glucose level in the body. The insulin making is insufficient because the body's cells do not respond suitably to insulin. Constant diabetes hyperglycemia is associated with long-term damage, brokenness and failure of multiple organs, particularly the kidneys, ewyes, heart, veins, and nerves. This study aims to present a comparative analysis among three popular machine taxonomy processes namely Support Vector Machine, Naive Bayes, and Decision Tree, used to perceive diabetes at a primary stage in a patient. In this work, we have tried to brief the most important machine learning algorithm with full accuracy to predict diabetes disease in a patient.

Keywords: Diabetes, Decision Tree, Support Vector Machine, Naive Byes, Decision Tree; Accuracy; Machine Learning, Central tendency, Dispersion.

I. Introduction

Disease, such as diabetes, is a common trouble that can be very severe in old ages and people not having a healthy lifestyle. Diabetes happens when there is too much blood glucose, also called blood sugar. Efficiently mining the diabetes data is a crucial concern. Regular check-ups and diagnoses and maintaining a proper eating habit can avert it to some level. Diabetes is measured one of the lethal diseases. It is an illness that disturbs the capacity of the body to produce the hormone insulin. Having too much glucose in your blood over time will lead to health difficulties, such as:

- Greater Thirst.
- Regular Urination.
- Great Hunger.
- Inexplicable Weight Loss.

If not treated on time, Diabetes may cause serious health problems in a person like:

- Diabetic Acidosis.
- Hyperosmolar Hyperglycemic State.
- Heart Disease.
- Kidney Disease.
- Nerve Damage.
- Blindness.

Diabetes, leads to long-term challenges and severe health issues. A study by the World Health Organization (WHO) [1] speaks diabetes and its issues that impact families financially, medically, and socially. The report says that 1.2 million deaths due to mortality because of the unregulated process of wellbeing. Around 2.2 million deaths have happened due to diabetes risk complications such as coronary and other disorders. About one in four people with diabetes has no idea that they have the disorder. Prediabetes is estimated to be found in 84.1 million Americans at or over 18 years of age. Diabetes is a condition caused by the elevated level of fixation with sugar in the blood. Prediction using machine learning algorithms is the primary focus. In many industry uses, such as e-commerce and many more, machine learning is widely used today. To guarantee a specific result, machine learning can predict valuable knowledge about large quantities of data. Due to the handling of a vast volume of data to merge data from many sources and to incorporate the study's context knowledge, several machine learning algorithms gain power. In the last few years, several scholars have made a major progress in the advancement of methods of disease diagnosis. Using diverse machine learning algorithms, we perform experiments to diagnose diseases, though machine learning algorithms function well in the diagnosis of multiple diseases.

Machine learning's acceptance and significance have also led to several practical applications, which have some significant effects on human society, such as machine learning's based medical image applications are a well-accepted active area of research. The recent development in machine learning further enlarges its use areas directly as-associated with human life, such as accurate and speedy future diabetes predictions, which can reduce the death rates by taking the earlier precautionary measures.

The goal of this research is to plan a system that can fully predict the risk of diabetes in patients. Therefore, the common classification of machine learning algorithms, Support Vector Machine (SVM), Naive Bayes (NB), and Decision Tree (DT) are used as a classifier tool to more accurately diagnose and forecast diabetes at an early stage since early detection is the only solution to stay away from extreme matte health.

1.1 Our Goals

Several statistical strategies have been applied in the last decade to forecast various human body diseases based on the latest medical literature using machine learning methods. In this work, based on certain important features that are ideally adapted based on our data collection that we have gathered, we forecast the incidence of diabetes disease in a patient. In our project, the popular SVM, NB, and DT models are used based on the symptoms, specifically the attributes needed to conduct the prediction. We will quickly figure out which model is the best, using the SVM, NB, and Decision Tree models, which will lead us to a better predictor of the disease.

In summary, the key goal of this work is to use the most popular and reliable machine learning to forecast diabetes disease in the human body, like the SVM, NB, and DT algorithms, to accurately predict disease events in disease-frequent communities. If well predicted well in advance, these types of medical condition forecasts will offer important insights for physicians who can then change their diagnosis, such as disease identification, health care, and community resources and per-patient treatment.

1.2 Paper Structure

The remaining part of this paper is formulated as follows: In Section 2, a detailed description of the use of machine learning models for diabetes disease prediction is presented. Chapter 3 the data summary, number of attributes and definition of each attribute are explained. Chapter 4 displays the measures in data preprocessing included in this analysis. Chapter 5 shows the experiment results and data accuracy results. Chapter 6 concludes this dissertation along with future work.

II. Related Work

This chapter would concisely answer many previous related forms of research based on classifying the dataset of diabetes that was selected as the essential tool of either single or hybrid approaches for computational intelligence..

In [2], Sajida et al. address the part of the machine learning approaches of Adaboost and Bagging ensemble [3], based on risk factors for diabetes, DT is the basis for classifying the patients as diabetic or non-diabetic. Results after the experiment demonstrate that the methodology of the Adaboost machine learning collaborative does well over getting and a DTof J48 in addition. In [4], Orabi et al. developed a method for the estimation of diabetes, the main purpose of which is to forecast the diabetes experienced by a candidate at a given age. By adding a DT, the proposed framework is considered based on the notion of machine learning. The findings obtained were adequate as the method built works fine in forecasting diabetes events at a given age, with a higher age.

In [5], Pradhan et al. applied genetic programming (G.P.) to train and validate the diabetes prediction dataset using the diabetes data collection sourced from the UCI repository. Compared with other applied methods, the outcomes obtained using Genetic Programming [7] has maximum precision. Inaccuracy can be greatly increased by taking less time to classify. A prediction model with two sub-modules was developed by Rashid et al. [8] to forecast diabetes-chronic disease. Compared to multivariate logistics regression (MLR) simulation, Wang et al. [16] suggested an ANN (Artificial Neural Network) as a classifier method for DM 12. The authors verified, on the basis of the analysis, that computer intelligence methods have more detailed outcomes relative to regression methods. Decision tree modeling was suggested by Varma et al. [17] to forecast and identify DM. Present standard DT models are suffering from a crisp boundary crisis. In order to avoid a sharp limit, the authors recommended strengthening the DT with fuzzy computation. Their research used 336 data points that were checked using the MATLAB method, resulting in a 75.8 percent accuracy.

III. MATERIAL AND METHODS

Three common classification algorithms for machine learning are introduced in this section. The basic Waikato Ecosystem for Information Analysis (WEKA) instrument is then defined along-with Minitab software. We would then add the dataset.

3.1 Classification Algorithms

3.1.1 Support Vector Machine
3.1.2 Naïve Bayes
3.1.3 Decision Tree

3.1.1 Support Vector Machine

A supervised classifier model for classification and regression used in machine learning is Support Vector Machine (SVM) [9]. In solving classification problems, it is primarily applied in a multidimensional space, the purpose of the SVM is to categorize data points by an effective hyper-plane. To distinguish data points, a hyper-plane is a decision boundary. The hyper-plane classifies the data points between the groups and the hyper-plane with the highest margin. By optimizing the distance between the two decision limits, the SVM finds the ideal dividing hyper-plane.

Pros

- Memory Efficient.
- Fast Prediction.
- Thrives in High Dimension.
- Kernel Flexibility.
- Both Classification and Regression Skills.

Cons

- Expensive Computation.
- Low Interpretability.
- Suitable for Small Datasets.
- Over-fitting Risk.
- Low Interpretability.

3.1.2 Naive Bayes

Naive Bayes (NB) [10] is a method of grouping which defines all characteristics that are independent and unrelated to each other. It says that in a class, the status of a specific feature does not disturb the position of another feature. It is called a robust algorithm used for classification purposes, as it is based on conditional probability.

Pros

- Easy Implementation.
- Fast and Simple.
- Noise Resilience.
- Easy Training.
- No Over-fitting.
- Computationally Efficient.
- Suitable for Large Datasets.

Cons

- No Regression.
- Biased Nature.
- Vanishing Gradient Problem.
- Limited Applications.

3.1.3 Decision Tree

One of the oldest and most popular machine learning procedures used to solve classification problems is the Decision Tree (DT) [11, 12]. The key goal is to use a DT from preliminary data to predict the target class using a decision law. It utilizes the prediction and grouping of nodes and inter-nodes. The DT selects each node at each point by evaluating the highest. The decision logics are modeled by a DT, i.e. checking and comparing results for categorizing data objects into a tree-like structure. Usually, the nodes of a DT have several levels where the first or top-most node is considered the root node. All internal nodes represent checks on input variables or attributes. The classification algorithm branches towards the required child node, based on the test performance, where the test and branching process repeats before it reaches the leaf node. The nodes of the leaf or terminal match the consequences of the decision. They are a popular component of many medicines.

Pros

- Easy Interpretation.
- No Normalization.
- Easy Data Preparation.
- Handles Missing Values.
- Fast Training.

Cons

- Over-fitting Risk.
- No Regression.
- Inadequate Prediction Powers.
- Imbalance Bias.

3.2 Waikato Environment for Knowledge Analysis (WEKA)

Waikato Environment for Information Analysis (WEKA) [13, 14] is a written Java application used for data classification. WEKA is open-source data mining software that combines machine learning algorithms into a practical graphical interface. Various data preprocessing, sorting, regression, clustering, correlation principles and visualization techniques are used in WEKA. You may add the algorithms to the data cluster either directly or by calling it using Java code. They are also ideal for designing new algorithms for machine learning. Today, in data mining and deep learning, WEKA is regarded as a landmark framework. Within academia and business circles, it has gained general recognition and has become a commonly used method for research into data mining.

3.3 PID Dataset

The Pima Indian Diabetes Dataset (PIDD) [15] of the National Institute of Diabetes and Digestive and Kidney Diseases, is used in the suggested technique (www.niddk.nih.gov). The aim of the dataset, based primarily on diagnostic measures in the dataset, is to forecast whether or not the patient has diabetes.

Table 1: Dataset Description

Database	***No. of Attributes***	***No. of Instances***
PIDD	8	769 71.88

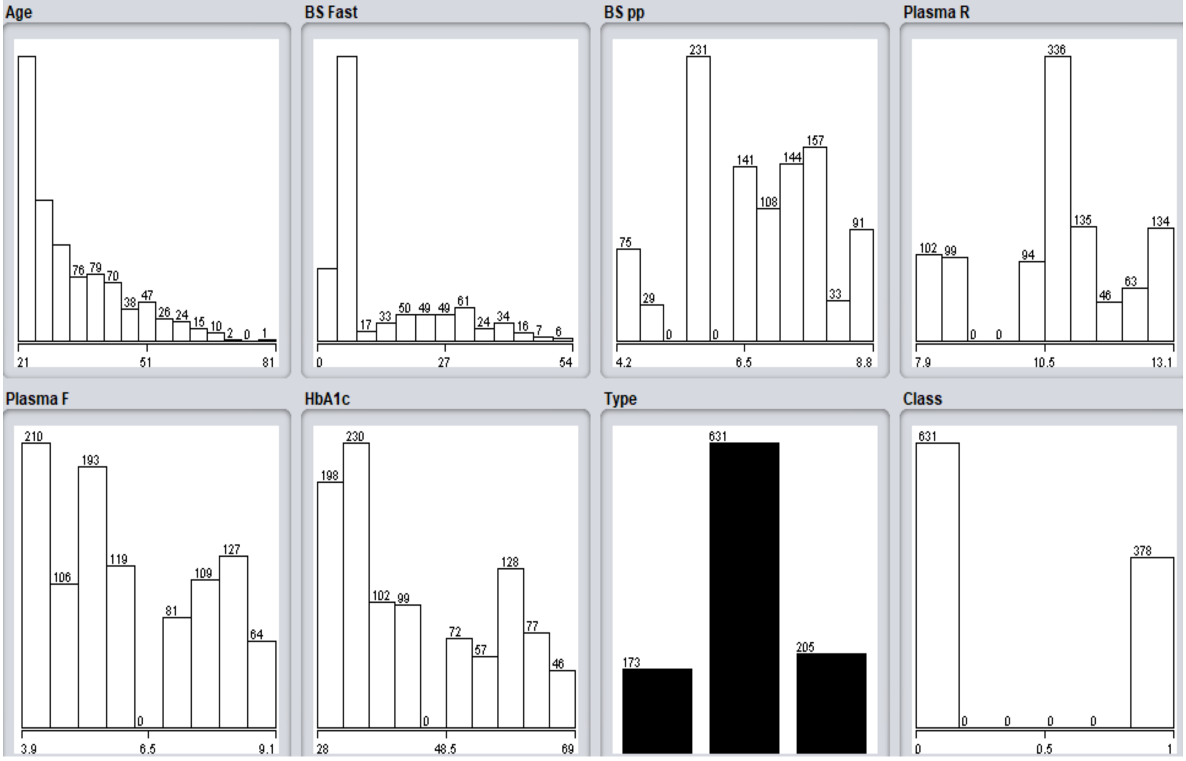

Fig. 2. Internal representation of PID Dataset.

IV. EXPERIMENTS

In this section, the experiment is conducted to assess the performance of three popular classification processes on the practical and straightforward PID dataset. Here, the performance of three popular classification algorithms is compared. The dataset, evaluation metrics, experiment implementation details, experimental results, analysis and comparisons of quantitative results will be described here.

4.1 Evaluation metrics

SVM, NB, and DT algorithms are used in this work. The accuracy of the algorithm in the prediction of instances is measured via different evaluation metrics such as F-Measure (F-M) which is the accuracy and recall weighted average, recall that measure the classifier completeness or sensitivity, precision that measure classifier correctness/ accuracy by precision, and Receiver Operating Curve (ROC) that curves are used to match the usefulness of test, are used for the classification of this work.

Table 2: Accuracy measures.

Measures	***Formula***
Precision	P = TP / (TP+ FP)
Recall	R =TP / (TP+FN)
F-Measure	F=2*(P*R) / (P+R)
Accuracy %	A=(TP+TN) / (Total no of samples)

4.2 Implementation details

All of the algorithms for classification are checked with (WEKA). The WEKA tool is used for experimentation in this job. WEKA is software developed by the University of Waikato (New Zealand) that provides a set of different methods of machine learning for data sorting, clustering, visualization, regression, etc. One of the most important benefits of using WEKA is that it can be configured according to the specifications. The key objective of this analysis is the prediction by the medical database PIDD of the patient affected by diabetes using the WEKA method.

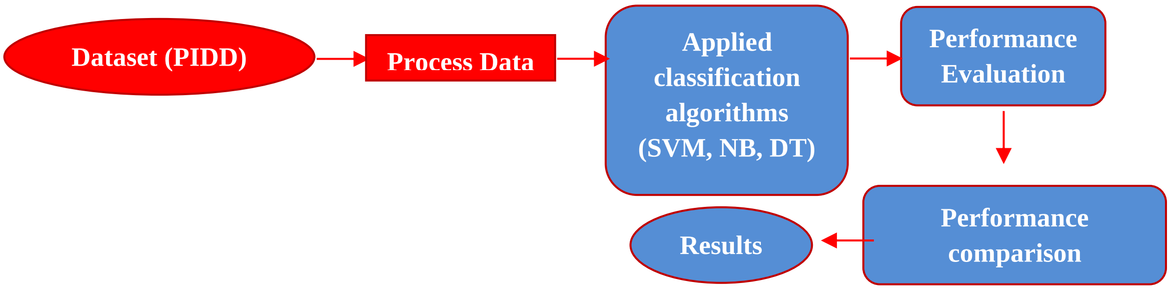

Fig. 3. Proposed Model Diagram

4.3 Experimental Results

The consistency of the grouping of three models in the earlier section is offered in this section to test the efficiency of each model, as seen in **Table 3**.

Table 3: Classification algorithms' comparable performance.

Classification Algorithms	***Pr.***	***Rec.***	***F-M***	***Acc. %***	***ROC***
SVM [9]	0.831	0.818	0.824	75.88	0.773
Naïve Bayes [10]	0.854	0.835	0.845	78.48	0.792
Decision Tree [11,12]	**0.903**	**0.873**	**0.866**	**80.56**	**0.821**

Table 3 will presents diverse performance values of all classification procedures discussed in this paper calculated on numerous measures. From **Table 3**, it can be analyzed which classification algorithms show the maximum accuracy and we may accomplish that which classification process outperforms reasonably other algorithms. As we can see in **Table 3**, a better quantitative score (higher values for Precision = 0.903, Recall = 0.873, F-Measure =0.866, Accuracy % = 80.56, and ROC = 0.821 indicate that DT is the most effective classification algorithms. The experimental results show that DT is the most effective classification algorithms than all other methods on the PIDD dataset discussed in Section 3.1.

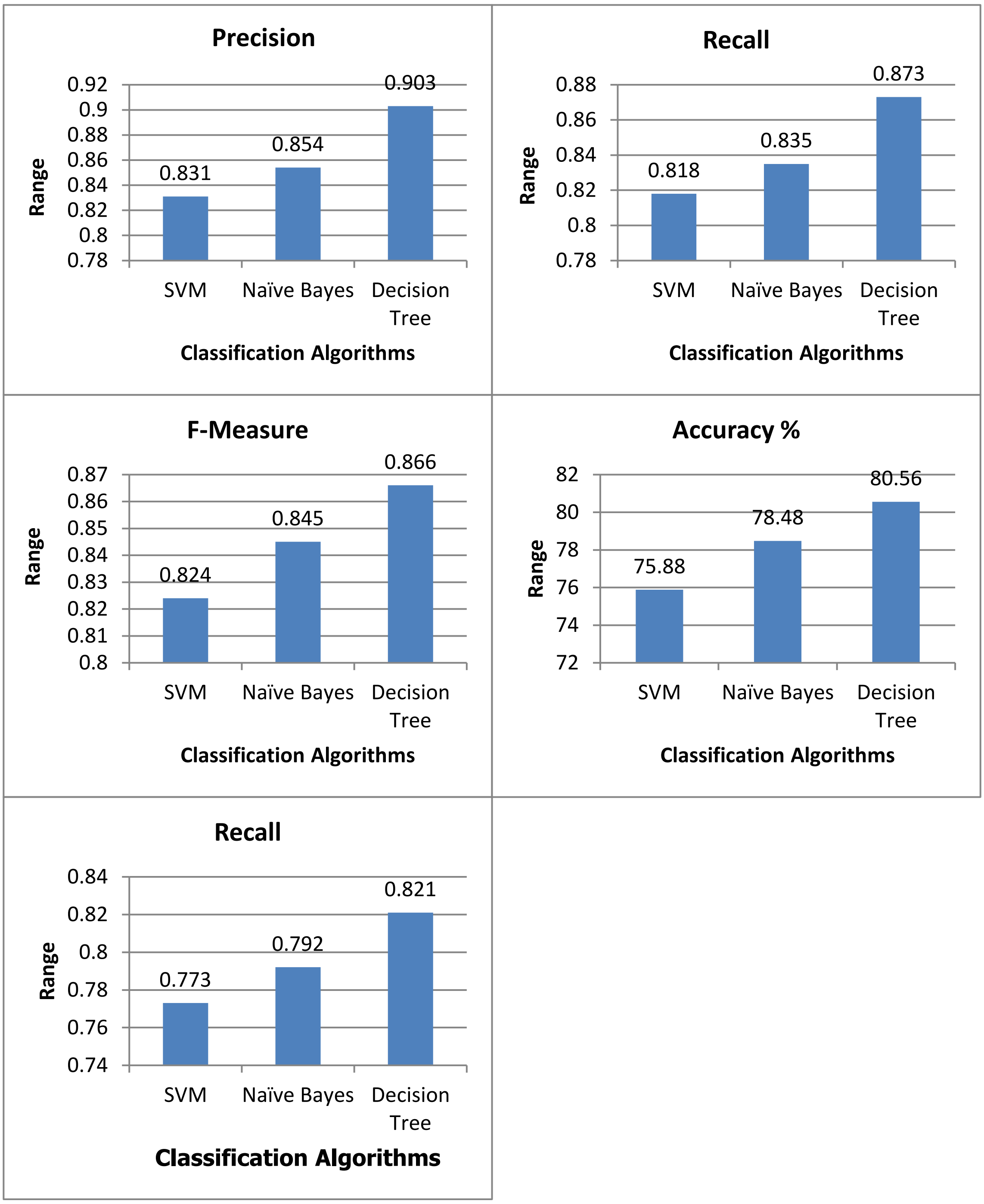

Fig. 4. Visualization of the performance of different classification algorithms on PID Dataset.

4.4 Statistical Analysis

Statistical research is one of the key methods used in epidemiology, and is mainly anxious with studying population health and illness. Popular descriptive statistical measurements are those of central tendency and dispersion.

4.4.1 Measures of Central Tendency

A series of measurements that represent a dataset's center point or value, around which most values would fall in a distribution. Measuring of central tendency includes the terms as shown in **Table 4**.

Table 4: Measures of central tendency.

Central Tendency Algorithms	***Formula***	***Description***
Mean	$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$	The mean is calculated by dividing the total number of rows in a column by the sum of the values in that column.
Median	Median $(x) = x_{(n+1)/2}$	A median is the value of an attribute that is in the middle.
Mode		In a list, the mode is the data point with the most counts.

Let's calculate the mean, median, and mode of the dataset (PIDD) attributes using WEKA tool.

Table 5: Measures of central tendency.

Central Tendency	***PIDD Attributes***					
	Age	BS Fast	BS pp	Plasma P	Plasma F	HbA1c
Mean	33.40	12.57	6.66	10.73	6.14	43.48
Median	29	6.7	6.8	10.9	5.6	40

4.4.2 Measure of Dispersion

Dispersion metrics explain how information is transmitted. Measuring of dispersion includes the terms as shown in **Table 6**.

Table 6: Measures of dispersion.

Dispersion	***Formula***	***Description***
Range		The spectrum is the difference between the maximum and minimum values in a column.
Variance	$Var_n (x) = E_n[x- E_n (x)]^2$	The square of standard deviation is variance.
Skewness	$Sk = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^{\sigma}$	The data distribution should have a Gaussian form (bell curve).
Standard Deviation	$\sigma = \sqrt{Var\ (n)}$	The standard deviation value indicates how far all data points differ from the mean.

Let's calculate the range, variance, skewness and standard deviation of the dataset (PIDD) attributes using WEKA tool.

Table 7: Measures of dispersion.

Dispersion	***PIDD Attributes***					
	Age	BS Fast	BS pp	Plasma P	Plasma F	HbA1c
Range	60	54	4.6	5.2	5.2	41
Variance	135.34	152.13	1.45	2.06	2.63	145.62
Skewness	1.05	1.21	-0.25	-0.53	0.26	0.54
Standard Deviation	11.63	12.33	1.21	1.44	1.62	12.07

V. CONCLUSION

In the current lifestyle, Diabetes disease is an important cause of dreariness and death. Three famous classification models: Support vector machine, Bayes classifiers, and Decision Tree were first studied. Then, WEKA methods were investigated for improving the strength of such models. Findings the best classification algorithm for disease measurement risk on the basis of the performance. This system statically analyses the more accurate model to recognize diabetes diseases using a machine learning approach. The obtained results will confirm that which machine learning algorithm contests powerfully against the detection of diabetes diseases according to the data set presented.

Acknowledgement

Author would like to thank management and staff of National College of Business Administration and Economics, Pakistan, for their kind support in bringing out the above literature and providing lab facilities.

Abbreviations

The following abbreviations are used in this manuscript:

ROC	Reciever Operating Characteristic
F-M	F-Measure
Pr.	Precision
Rec.	Recall
Acc.	Accuracy
SVM	Support Vector Machine
NB	Naïve Bayes
DT	Decision Tree
PIDD	Pima Indian Diabetes Dataset

REFERENCES

1. Global Report on Diabetes by World Health Organization. ISBN 978 92 4 156525 7. (30) 2016.
2. S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. In Procedia Computer Science, Volume 82, Pages 115-121, 2016.
3. N. Nai-Arun, and P. Sittidech. Ensemble Learning Model for Diabetes Classification. In Advanced Materials, Volume 931-932, Pages 1427-1431. 2014.
4. K. M. Orabi, Y. M. Kamal, and T. M Rabah. Early Predictive System for Diabetes Mellitus Disease. In Industrial Conference on Data Mining, Springer, Pages 420–427. 2016.
5. P.M.A. Pradhan, G.R. Bamnote, V. Tribhuvan, K. Jadhav, V. Chabukswar, and V. Dhobale. A Genetic Programming Approach for Detection of Diabetes. In International Journal of Computational Engineering Research, Volume 2, Pages 91–94, 2012.
6. M. Pradha, and G. R. Bamnote. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. In Advances in Intelligent Systems and Computing, Volume 1, Pages 763–770, 2014.
7. A. A. Sharief, and A. Sheta. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. In International Journal of Advanced Research in Artificial Intelligence (IJARAI), Volume 3, Pages 54–59, 2014.
8. T. A. Rashid, S. M. Abdullah, and R. Z. Abdullah. An Intelligent Approach for Diabetes Classification, Prediction and Description. In Advances in Intelligent Systems and Computing, Volume 424, Pages 323–335, 2016.
9. B. Schoslkopf, and A. Smola. Learning with Kernels, Support Vector Machines. London: MIT Press, 2002.
10. I. Rish. An Empirical Study of the Naive Bayes classifier. In Workshop on Empirical Methods in Artificial Intelligence, Pages 41–46, 2001.
11. F. Esposito, D. Malerba, G. Semeraro, J. Kay. A comparative analysis of methods for pruning decision trees. In IEEE Transactions on Pattern Analysis and Machine Intelligence Volume 19, Pages 476–491, 1997.
12. A. Priyam, R. Gupta, A. Rathee, and S, Srivastava. Comparative Analysis of Decision Tree Classification Algorithms. In International Journal of Current Engineering and Technology Volume 3, Pages 334–337, 2013.
13. R. Arora, and Suman. Comparative Analysis of Classification Algorithms on Different Datasets Using WEKA. In International Journal of Computer Applications, Volume 54, Pages 21–25, 2012.

14. S.R. Garner. Weka: The Waikato Environment for Knowledge Analysis. In Proceedings of the New Zealand computer science research students' conference, Citeseer, Pages 57–64. 1995.
15. K. Kayaer, and K. Yildirim. Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks. In Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP), Pages 181–184. 2003.
16. C. Wang, L. Li, L. Wang, Z. Ping, MT. Flory, G. Wang, et al. Evaluating the Risk of Type 2 Diabetes Mellitus Using Artificial Neural Network: An Effective classification Approach. In Diabetes Research and Clinical Practice, Pages 111-118, 2013.
17. K. V. Varma, A.A. Rao, TSM. Lakshmi, and PN. Rao. A computational Intelligence Approach for a Better Diagnosis of Diabetic Patients. In Computers & Electrical Engineering, Pages. 1758-1765, 2014.