Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

Diabetes Disease Prediction Based on Symptoms Using Machine Learning Algorithms

K. Sujatha[1*], K.V. Krishna Kishore[2], B. Srinivasa Rao[3],Rajkumar Rajasekaran[4]

[1*]Vignan's Foundation for Science Technology and Research, Vadlamudi, Guntur,AP, India.
E-mail: Sujatha101012@gmail.com

[2]Vignan's Foundation for Science Technology and Research, Vadlamudi, Guntur, AP, India.
E-mail: kishorekvk_01@yahoo.com

[3]Vignan's Foundation for Science Technology and Research, Vadlamudi, Guntur, AP, India.
E-mail: drbsraoprofessormmt10672@gmail.com

[4]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.
E-mail: vitrajkumar@gmail.com

ABSTRACT

People's busy lifestyle and food habits throw them into the pool of many lifestyle diseases; diabetes is one of those. Overall, 422 million people live with diabetes as per World Health Organization (WHO) 2014 statistics. This is particularly severe in the middle- and low-income countries irrespective of gender and age. Much research has been done in predicting diabetes disease but primarily focused on symptoms and diagnostic test reports. Although there are many existing algorithms available, no single classifier is providing an optimal solution. Therefore, in this work, we primarily focus on diabetes occurrence risk prediction by ensembling various machine learning techniques. We also aimed at predicting the risk of diabetes among men and women based on various symptoms such as Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, and Irritability. Various combinations of algorithms are taken in an ensemble approach during experimentation. Finally, the combination of Naïve Bayes Classifier, SVM Classifier, J48, Optimized Parametric Multilayer Perceptron are performing better when compared with other combinations. All the experiments were carried out on Kaggle data science datasets. The experiments show that the ensemble classifier predicts the disease with almost 100% accuracy on the benchmark dataset. Further, the experimentation can be extended to predict the level of occurrence and the occurrence of other health complications because of this disease.

KEYWORDS

Diabetes disease,Machine learning models,Pima Indian Diabetes Dataset**,** SVM Classifier, Decision tree Classifier, Naïve Bayes, J48 Pruned tree, Multilayer Perceptron.

Introduction

In a competitive society, people spend their lives too busy, not taking care of their health and food habits. As a result, most people are affectedbylifestyle diseases such as diabetes, blood pressure, gastric problems, acidity problems etc. Diabetes diseaseis closely related to our lifestyle as well as food habits also it may be genetic (Tripathi & Kumar, 2020). Diabetes disease is one of the non-transferable diseases on the planet. Prior clinical records show that the forecast and counteraction of diabetes have become a significant challenge. As the quantity of analyzed patients is expanding, the drugs are yet insufficient to control the infection. So, a superior prescient investigation is needed to treat diabetes at a beginning phase which can help explain fewer issues that can assist with treating the patient with fewer meds and affordability (Rout and Kaur, 2020).

As should be obvious in numerous cases, which shows that the conventional ideal models about the event of a particular sort of diabetes age bunches are not, at this point exact, early indications of diabetes like weariness, obscure vision, cerebral pains, and so on must be inspected before it closes to a genuine stage (American Diabetes Association, 2019). Numerous cases are undiscovered because no early registration or mindfulness about primarily the blood glucose level among various age bunches builds the danger of different infections. That is the reason it is suggested for registration at standard stretches if any of such side effects are seen to maintain a strategic distance from additional harms to the body as even diabetes can go unnoticed, dissimilar to different sicknesses notwithstanding being driving a sound lifestyle (Rout and Kaur, 2020).

Examination on diabetes patients exhibits that diabetes among grown-ups (more than 18 years of age) has ascended from 4.7%to 8.5%in 1980 to 2014 individually and quickly experiencing childhood in second and underdeveloped nations (The Emerging Risk Factors Collaboration,2010). According to the statistical facts and outcomes from the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

studies, around 450 million people were living with diabetes all over the world. This number may be incremented to 690 million by the year 2045 (N. H. Cho et al., 2018). Another factual examination in (P. Saeedi et al., 2019) shows the seriousness of diabetes, where they revealed that a large portion of a billion people has diabetes around the world. The number will increment to 25% and 51% separately in 2030 and 2045. Notwithstanding, there is no drawn-out solution for diabetes. However, it very well may be controlled and forestalled if an early expectation is precisely conceivable (Hasan et al., 2020).

Over the most recent twenty years, AI methods have been generally utilized to characterize different sicknesses in the clinical field. The main purpose of grouping in the clinical field is the exactness of the outcomes and expectations(Bandaru & Kamepalli, 2019). Numerous investigations have been attempting to create productive techniques or calculations to analyze the various illnesses dependent on the available datasets. The fundamental point of applying AI in the well-being area is that the quantity of the information is enormous and needs an ideal opportunity to investigate it, so utilizing AI will distinguish the sickness with not so much time but rather more productive(Sujatha & Srinivasa Rao, 2019). Diabetes infection draws in incredible consideration in the AI people group. Since diabetes is an ongoing illness and should be identified at the beginning phase to manage the right drug. The expansion in the sugar proportion in the blood causes diabetes sickness. Since the delay of diabetes will cause passing and a more muddled life for the patients (Al-Zebari and Sengur, 2019). Moreover, the early identification will assist the patients with evading the inconvenience of the drug and entanglements (V.V. Vijayan etal., 2015, and R. Arora., 2012).

Literature Survey

(Al-Zebari & Sengur, 2019)made the performance comparison of the ML algorithms on diabetes disease prediction. They implemented various machine learning algorithms such as logistic regression, k- nearest neighbours, SVM etc., for diabetic disease prediction in MATLAB. The authors used the PIDD diabetes dataset for experimentation. All the patients in PIDD were are women and 21 years old as minimum age, and all are near to Phoenix, Arizona region.

(Sonar & Jaya Malini, 2019) predicted diabetes disease using various ML algorithms such as Decision tree, SVM, Naive Bayes and ANN. The authors implemented machine learning models on the PIDD dataset,consisting of 8 attributes and 768 instances. All the data is collected from a woman only.(Vijiyakumar, Lavanya, Nirmala, & Sofia Caroline, 2019)Used random forest algorithm for predicting diabetes disease, and the authors implemented the algorithm on the same dataset on which the above two authors worked. From the experimentation, it is clear that the random forest algorithm predicts the disease with an accuracy of 78.05%.(Driss, Boulila, Batool, & Ahmad, 2020)The authors presented three different Machine Learning (ML) methods. Out of three algorithms, Naïve Bayes (NB) has the highest accuracy of 76.30%. They also used the same dataset. (Kohli & Arora, 2018)applied different classification algorithms on different datasets for disease prediction. All the datasets were collected from the UCI machine learning repository. From the results obtained, it is clear that diabetes prediction using Support Vector Machine (linear kernel)done with an accuracy of 85.71%.

(Yahyaoui, Jamil, Rasheed, & Yesiltepe, 2019)proposed a decision support system for diabetes prediction based on machine learning algorithms. They also compared machine learning models with deep learning in predicting diabetic disease. From the experimental results,the random forest method performs well, with an accuracy of 83.67%. they also implemented the models on the same dataset.(Saha, Patwary, & Ahmed, 2019)applied Neural Network(NN), SVM, Random Forest(RF), etc.,algorithms for diabetes disease prediction to compare and get the best accuracy. Neural Network was given the best accuracy (80.4%) than any other techniques. They also used the same dataset.(Jacob, Raimond, & Kanmani, 2019)Does a survey on the different kinds of predictions using ML techniques done on diabetes patients. Their survey concluded that manyresearch worksfrequently focusedonthe Pima Indian diabetes dataset, 768 records. (Hasan, Alam, Das, Hossain, & Hasan, 2020)were conducted the same experiments on the Pima Indian Diabetes Dataset. The authors proposed a robust framework for diabetes prediction using different Machine Learning (ML) classifiers (k-nearest Neighbor, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP) were employed. The weighted ensemble of different machine learning models is also proposed.

(Rout & Kaur, 2020) implemented various ML algorithms on the PIDD dataset to predict diabetes disease. (Tripathi & Kumar, 2020) implemented four machine learning algorithms Linear Discriminant Analysis (LDA), K-nearest neighbour (KNN), Support Vector Machine (SVM), and Random Forest (RF), to do predictive analysis of early-stage

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

diabetes. (Pethunachiyar, 2020) implemented support vector machines to predict diabetic disease with different kernel functions. Linear kernel efficiently makes a diabetic prediction.(Mir & Dhage, 2018) built a classifier model using the WEKA tool to predict diabetes disease by employing Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm. (Sarwar, Kamal, Hamid, & Shah, 2018) discussed the predictive analytics in healthcare by implementing various machine learning algorithms on the PIDD diabetes dataset. (Abbas, Alic, Rios, Abdul-Ghani, & Qaraqe, 2019)Used support vector machines to predict the future development of type-2 diabetes on the San Antonio Heart Study data.

Problem Statement

In all these existing models, various machine learning algorithms are used to predict diabetes disease. But the majority of the authors considered Pima Indian Diabetes Dataset for experimentation. This dataset is prepared based on the symptoms and test reports of woman patients. Not only a woman but men can also be affected by diabetes disease. So, it is better to develop a decision support system that predicts diabetes disease irrespective of gender. This paper aimed at predicting the occurrence of diabetes disease in men and women based on various symptoms such as Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, Obesity.

Objectives

The main objectives of this paper are.

- Studying the state of the art machine learning models that existed in diabetes prediction.
- Analyzing various symptoms that indicates the occurrence of diabetes.
- Implementing machine learning algorithms to predict diabetes disease based on various symptoms.
- Comparing the results of machine learning models and finding the best suitable model for predicting the diabetes disease based on symptoms.

Methodology

The following diagram depicts the flowchart of the proposed work. The methodology includes collecting the dataset, preprocessing the data, feature selection, implementing various machine learning algorithms for predicting diabetes disease, comparing results obtained from various models and finally finding the best suitable method for predicting the diabetes dataset based on symptoms.

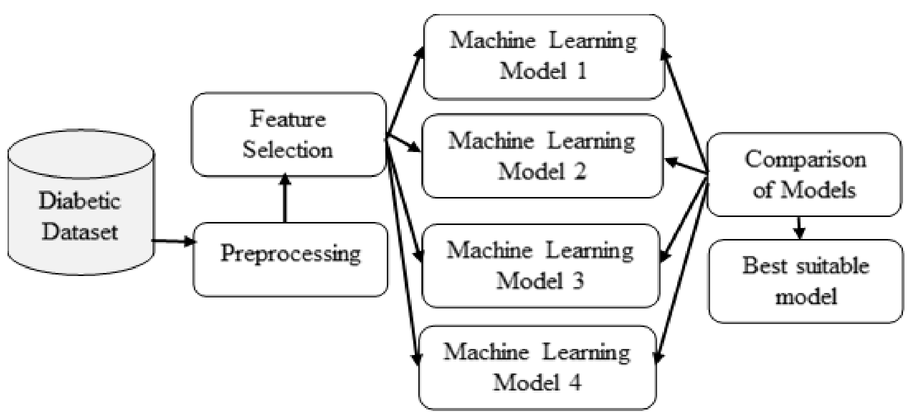

Fig. 1. Proposed Flow of Work

Source: Proposed Methodology

Experimentation

The experiments were conducted by considering four ML algorithms on the diabetes dataset. All the experiments were conducted in the same experimental environment.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

Naive Bayes Classifier: It is a probabilistic ML classifier model that works based on the Bayes theorem. Bayes theorem can be defined as

$$P\left(\frac{A}{B}\right) = \frac{P(B/A)P(A)}{P(B)}$$

Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred. Here, **B** is the evidence, and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is,the presence of one feature does not affect the other. Hence it is called naive.

SMO:SMO alludes to the effective enhancement calculation utilized inside the SVM execution, representing Sequential Minimal Optimization. This carries out John Platt's consecutive negligible advancement calculation for preparing a help vector classifier. "Backing Vector Machine" (SVM) is a directed AI calculation that can be utilized for both grouping or relapse difficulties. In any case, it is generally utilized in order issues. In the SVM calculation, we plot every information thing as a point in n-dimensional space (where n is several highlights you have), with the worth of each component being the worth of a specific arrangement. At that point, we perform order by tracking down the hyper-plane that separates the two classes well overall (take a gander at the underneath depiction).

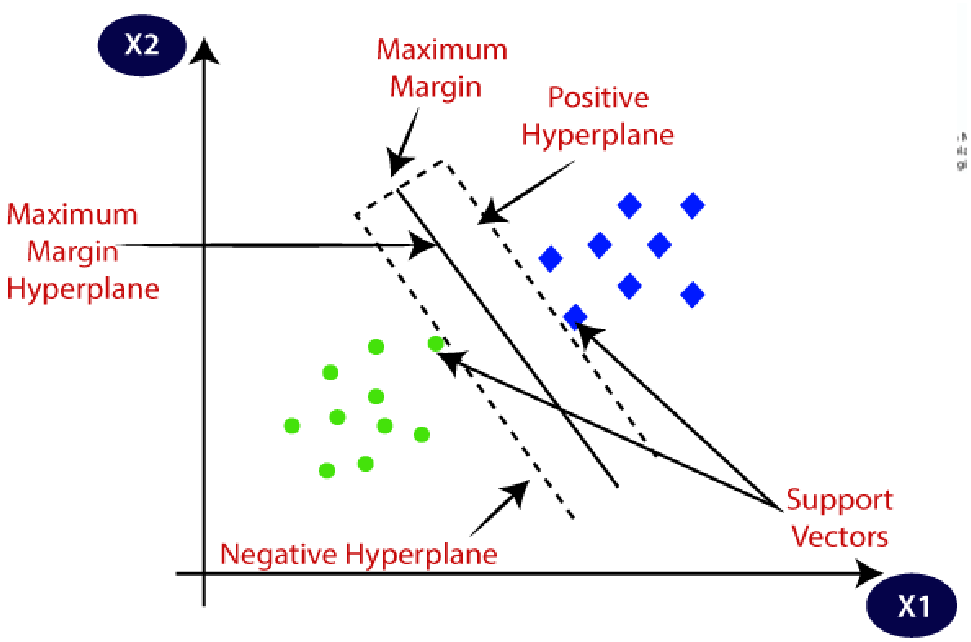

Fig.2. Support Vector Machines
Source: https://static.javatpoint.com/tutorial/machine-learning/images/support-vector-machine-algorithm.png

Support Vectors are simply the coordinates of individual observation. The SVM classifier is a frontier that best segregates the two classes (hyper-plane/ line).

J48 pruned tree:J48 model is extraordinary compared to other AI calculations to analyze the information completely and consistently. C4.5 (J48) is a calculation used to create a choice tree created by Ross Quinlan referenced before. C4.5 is an expansion of Quinlan's previous ID3 calculation. The choice trees produced by C4.5 can be utilized for arrangement, and therefore, C4.5 is regularly alluded to as a measurable classifier.

Multilayer Perceptron:A multi-facet perceptron (MLP) is a class of feedforward counterfeit neural organization. MLP is utilized vaguely, once in a while, freely to any feedforward ANN, at times stringently to allude to networks made out of numerous layers of the perceptron. An MLP comprises of in any event, three layers of hubs: an info layer, a secret layer and a yield layer. Except for the information hubs, every hub is a neuron that utilizes a nonlinear initiation work. MLP uses a regulated learning procedure called backpropagation for preparing.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

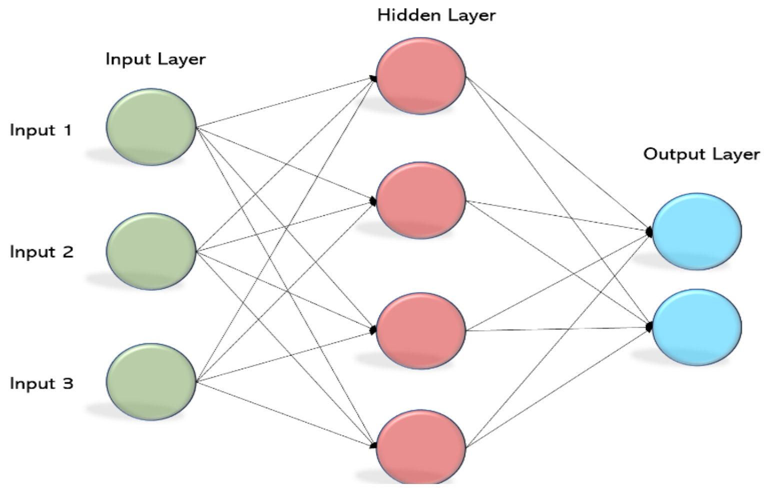

Fig 3: Multi-layer perceptron model
Source: https://miro.medium.com/max/3446/1*-IPQlOd46dlsutIbUq1Zcw.png

Sample Dataset

Dataset was collected from the Kaggle data science community. The dataset consists of 6678 instances and 17 attributes. The following table gives the sample dataset.

Table 1. Sample Dataset

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
55	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Positive
57	Male	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	No	Positive
66	Male	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No	No	Positive
67	Male	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Positive
70	Male	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	Positive
44	Male	Yes	Yes	No	Yes	No	Yes	No	No	Yes	Yes	No	Yes	Yes	No	Positive
38	Male	Yes	Yes	No	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	Positive
35	Male	Yes	No	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	No	Positive
61	Male	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Positive
60	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	No	No	Positive
58	Male	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	No	No	Positive
54	Male	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	No	Yes	No	No	Positive
67	Male	No	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Positive
66	Male	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	Positive

Source: https://www.kaggle.com/shikhnu/diabetes-risk-prediction-dataset

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

Attribute Description: The diabetes dataset collected from**the** Kaggle data science community consists of 17 attributes, including class attribute. These attributes represent various symptoms generally identified in diabetes patients. The following table describes the attributes in the dataset.

Table 2. Attribute Description

S. No.	Attribute Name	Description	Values
1	Age	Age (In years)	Ranged from 16 years to 90 years
2	Gender	Gender of the patient	Male or Female
3	Polyuria	Excessive pass of urinals	Yes or No
4	Polydipsia	Excessive thirsty	Yes or No
5	sudden weight loss	Losing weight suddenly without following any diet chart	Yes or No
6	weakness	Weakness is a decrease in the strength	Yes or No
7	Polyphagia	excessive eating	Yes or No
8	Genital thrush	Thrush is a fungal infection caused by Candida yeasts	Yes or No
9	visual blurring	No visual clarity	Yes or No
10	Itching	Skin irritation	Yes or No
11	Irritability	Irritability is a feeling of agitation	Yes or No
12	delayed healing	Delayed wound healing is when it takes longer for a wound to heal than normal.	Yes or No
13	partial paresis	Paresis involves the weakening of a muscle or group of muscles.	Yes or No
14	muscle stiffness	Tightness of muscles	Yes or No
15	Alopecia	Hair falling	Yes or No
16	Obesity	Body fat increase	Yes or No
17	class	A class attribute represents whether the patient is suffering from diabetes or not	Positive or Negative

Source: https://www.kaggle.com/shikhnu/diabetes-risk-prediction-dataset and Internet (Google)

The following diagram depicts the visualization of various attribute values. This clearly shows the distribution of attributes in terms of their distinct values.

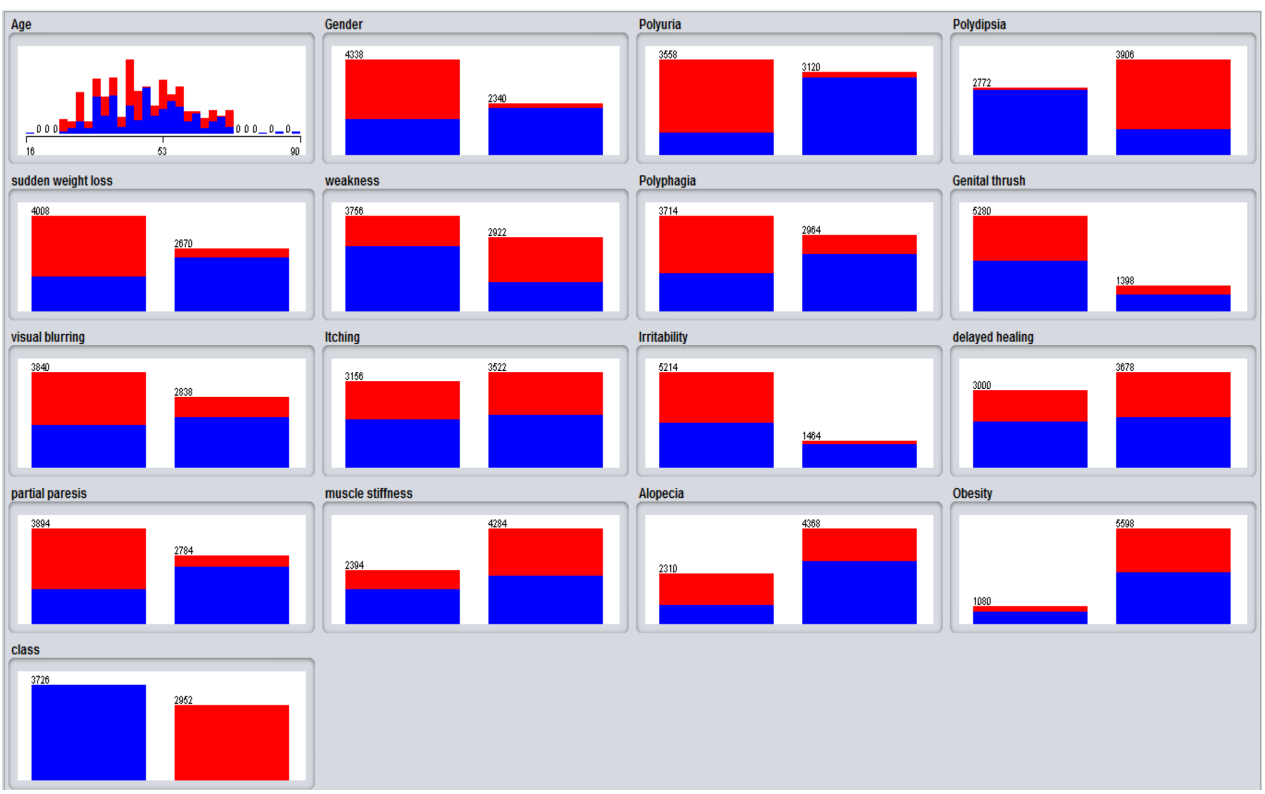

Fig.4. Attribute Visualization

Source: Experimentation

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

Results and Discussions

In this section, the results obtained by implementing various machine learning models on the diabetes dataset were discussed. All four models were implemented in the same experimental environment.

Naive Bayes Classifier: The first machine learning model implementedon the diabetes dataset is the Naïve Bayes classifier model. The following table shows the confusion matrix generated by the Naïve Bayes classifier model.

Table 3. Class matrix of Naïve Bayes classifier model

Naïve Bayes classifier	Positive	Negative
Positive	3202	524
Negative	288	2664

Source: Experimentation

The following diagram depicts the class matrix generated by implementing the Naïve Bayes classifier model on the diabetes dataset.

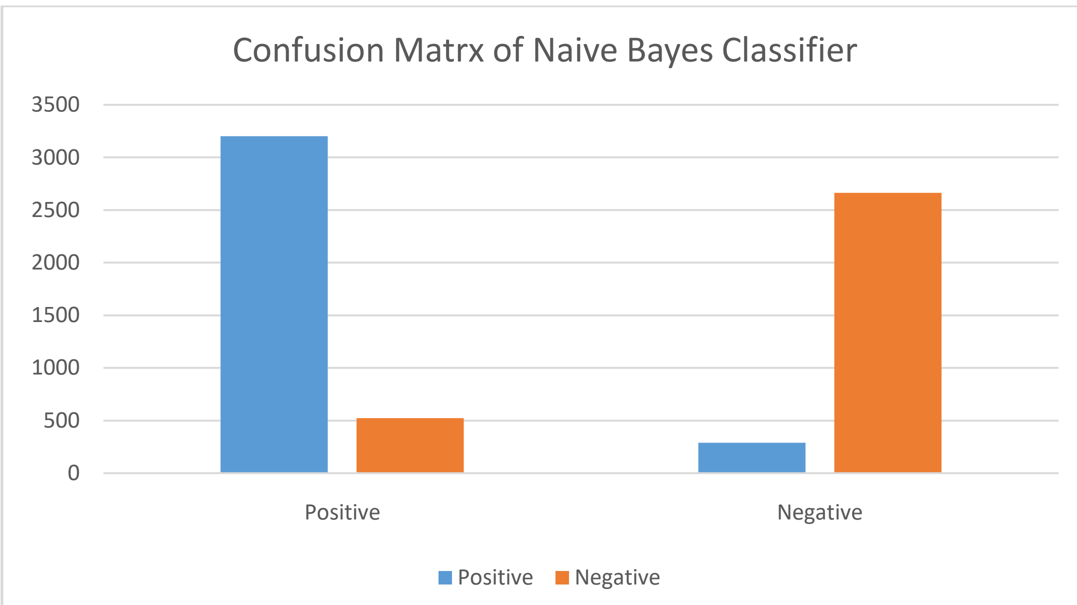

Fig. 6. Graphical representation of a class matrix of Naïve Bayes classifier model
Source: Experimentation

SMO: The second machine learning model implementedon the diabetes dataset is the SMO classifier model. The following table shows the confusion matrix generated by the SMO classifier model.

Table 4. Class matrix of SMO classifier model

SMO	Positive	Negative
Positive	3525	201
Negative	108	2844

Source: Experimentation

The following diagram depicts the class matrix generated by implementing the SMO classifier model on the diabetes dataset.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

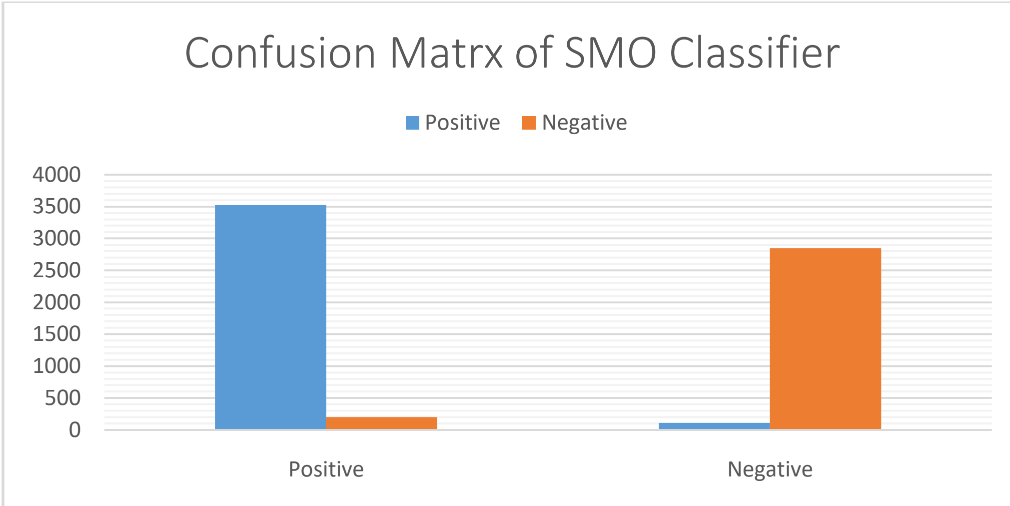

Fig. 7. Graphical representation of a class matrix of SMO classifier model
Source: Experimentation

J48 pruned tree: The third machine learning model implementedon the diabetes dataset is J48 pruned tree classifier model. The following table shows the confusion matrix generated by the J48 pruned tree classifier model.

Table 5. Class matrix of J48 classifier model

J48	**Positive**	**Negative**
Positive	3726	0
Negative	0	2952

Source: Experimentation

The following diagram depicts the class matrix generated by implementing the J48 pruned tree classifier model on the diabetes dataset.

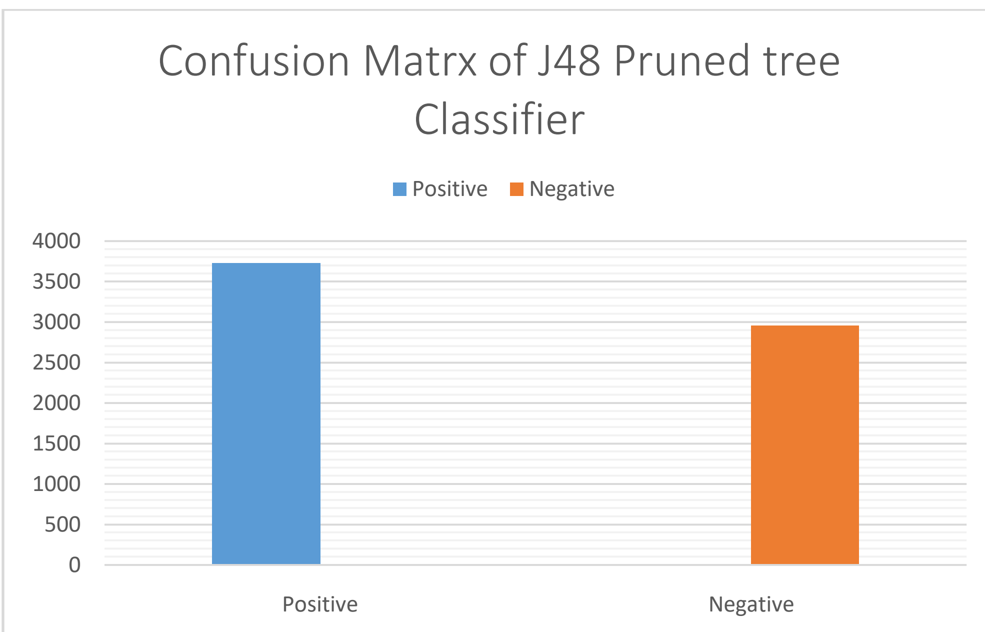

Fig. 8. Graphical representation of a class matrix of J48 pruned tree classifier model
Source: Experimentation

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

The following diagram depicts the decision tree generated by implementing the J48 pruned tree classifier model on the diabetes dataset.

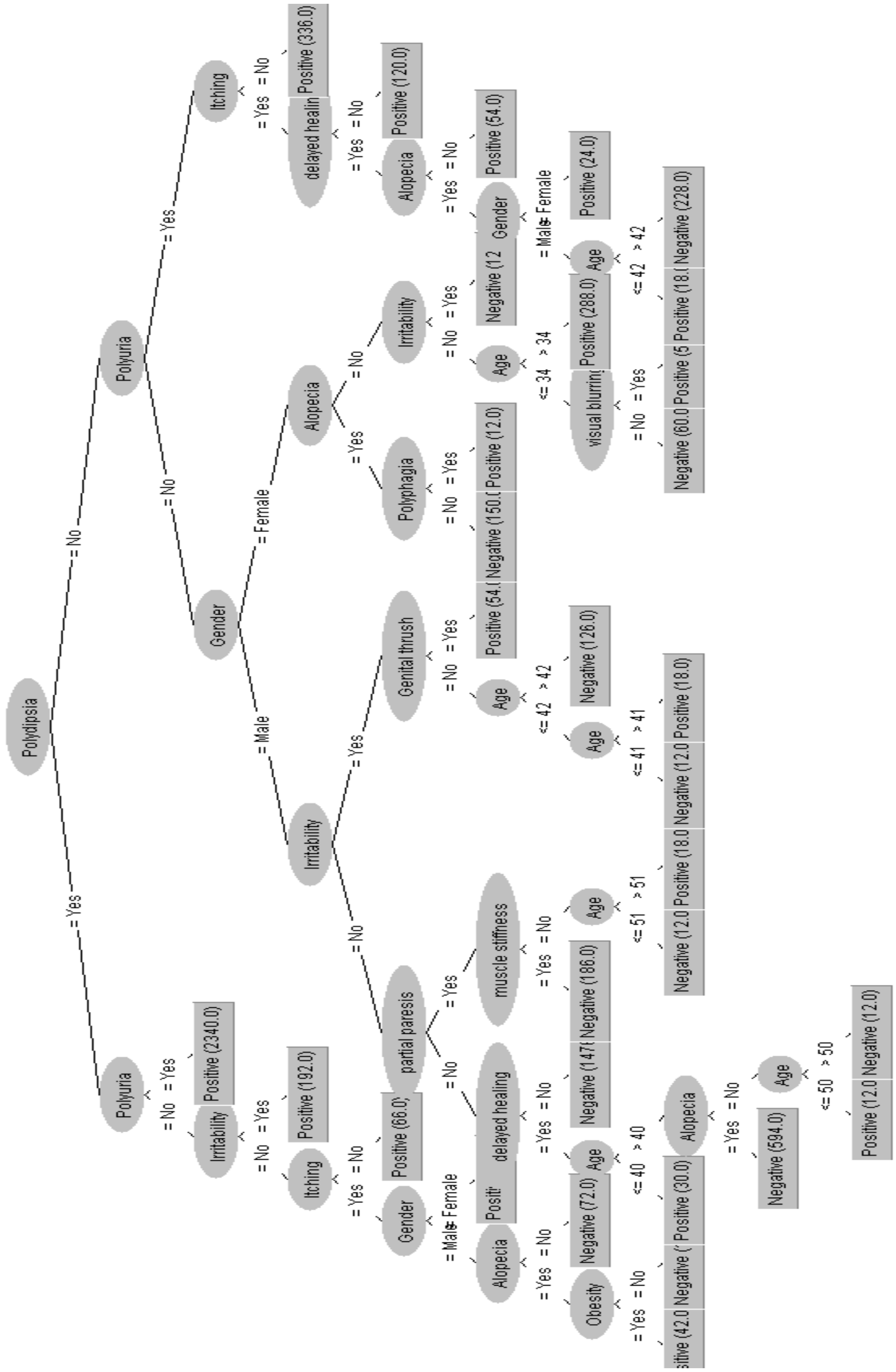

Fig.9. Decision tree constructed from J48 decision tree algorithm

Source: Experimentation

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

Multilayer Perceptron: The fourth machine learning model implementedon the diabetes dataset is the Multilayer Perceptron classifier model. The following table shows the confusion matrix generated by the Multilayer Perceptron classifier model.

Table 6. Class matrix of Multilayer Perceptron classifier model

Multilayer Perceptron	Positive	Negative
Positive	3711	15
Negative	14	2938

Source: Experimentation

The following diagram depicts the class matrix generated by implementing the Multilayer Perceptron classifier model on the diabetes dataset.

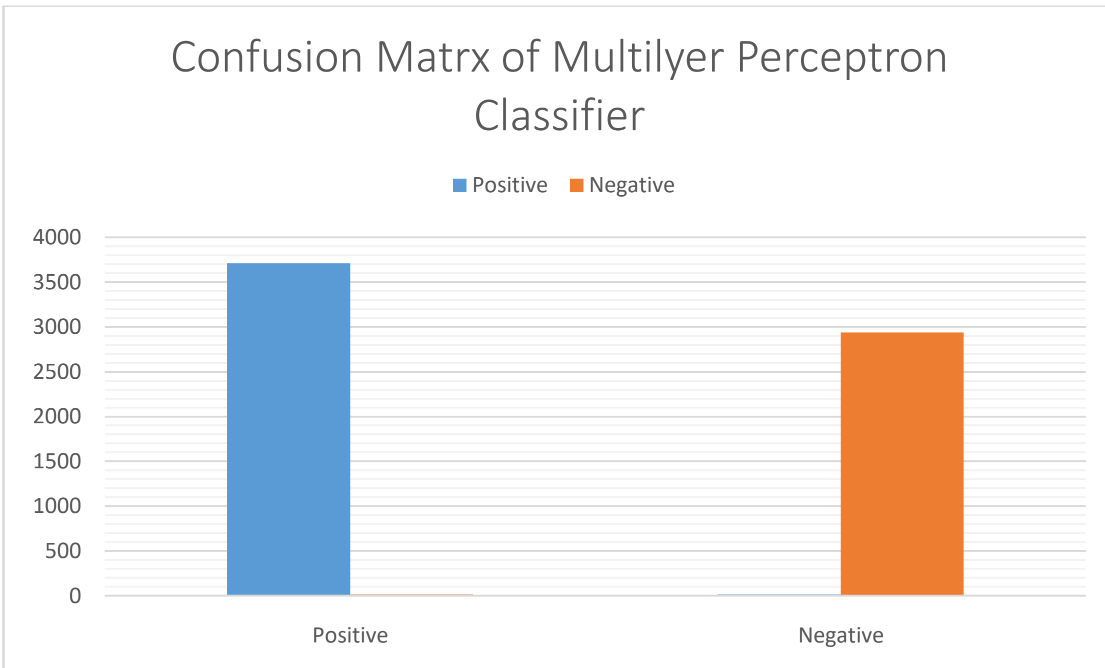

Fig. 10. Graphical representation of a class matrix of Multilayer Perceptron classifier model
Source: Experimentation

Here we compared the four machine learning models implemented on the diabetes dataset in terms of the correctness of the classification.

Table 7. Correctly and incorrectly classified instances

Classifier Model	Correct classifications	Percentage of correct classifications	Incorrect classifications	Percentage of incorrect classifications
Naive Bayes Classifier	5866	87.8407%	812	12.1593%
SMO	6369	95.3729%	309	4.6271%
J48 pruned tree	6678	100%	0	0%
MultilayerPerceptron	6649	99.5657%	29	0.4343%

Source: Experimentation

The following diagram depicts the comparison of four machine learning algorithms in terms of the correctness of classification on the diabetes dataset.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 3805 - 3817
Received 25 April 2021; Accepted 08 May 2021.

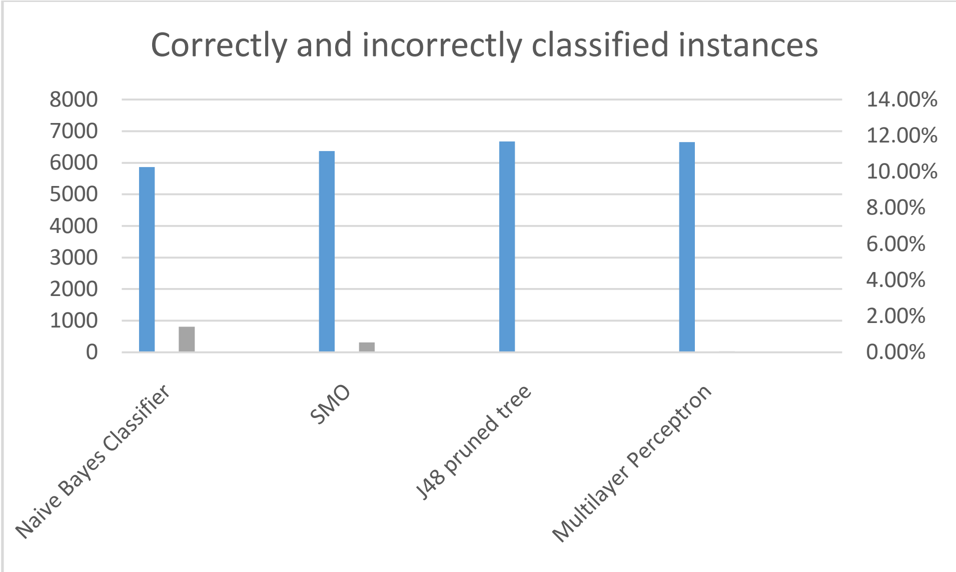

Fig. 11. Graphical representation ofcorrectly and incorrectly classified instances

Here we compared the four machine learning models implemented on the diabetes dataset regarding various parameters such as True Positive rate, False Positive rate, Precision, Recall etc.

Table 8. Comparison of machine learning models implemented in terms of various parameters

	TP	**FP**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Naive Bayes Classifier	0.878	0.117	0.881	0.878	0.879	0.959
SMO	0.954	0.044	0.954	0.954	0.954	0.955
J48 pruned tree	1.000	0.000	1.000	1.000	1.000	1.000
Multilayer Perceptron	0.996	0.004	0.996	0.996	0.996	0.999

Source: Experimentation

The following diagram depicts the comparison of four machine learning algorithms in terms of various parameters such as TP rate, FP rate, Precision, Recall etc., of classification on the diabetes dataset.

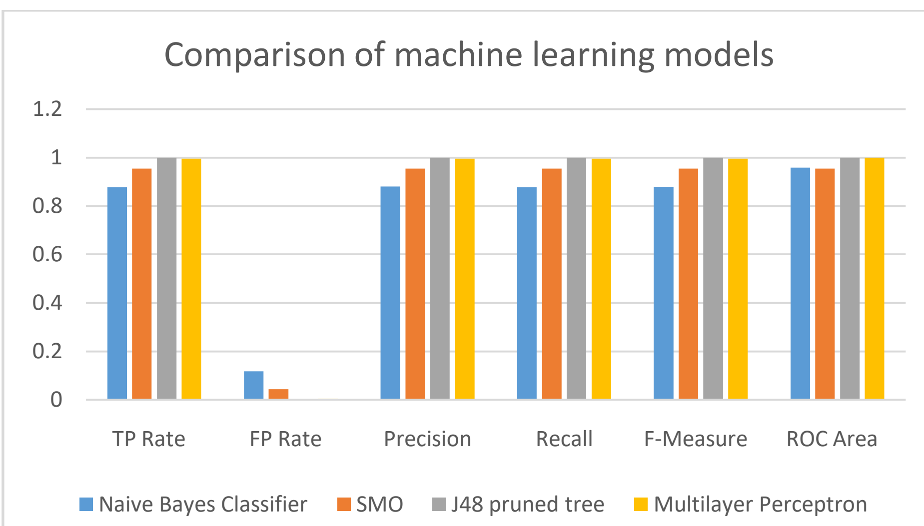

Fig.12. Graphical representation of a comparison of machine learning models implemented in terms of various parameters.

Source: Experimentation

From the results obtained from the experimentation, the four machine learning classifier models Naive Bayes Classifier, SMO, J48 pruned tree, and Multilayer Perceptron predicted with an accuracy of 87.8407%, 95.3729%, 100%, and 99.5657%, respectively.

Conclusion

Diabetes mellitus (DM) is one of the most lethal non-communicable diseases in the world. A lot of research has been done in predicting diabetes disease. In all these existing models, various machine learning algorithms are used to predict diabetes disease. But the majority of the authors considered the PIDD dataset for experimentation. This dataset is prepared based on the symptoms and test reports of woman patients. This paper mainly concentrated on predicting the diabetes disease occurrence risk prediction irrespective of gender. Various combinations of algorithms are taken in an ensemble approach during experimentation. Finally, the combination of Naïve Bayes Classifier, Support Vector Machine, J48, and Optimized Parametric Multilayer Perceptron is performing better when compared with other combinations. All models were implemented on a diabetes dataset collected from the Kaggle data science community. From the results obtained, the four machine learning classifier models Naive Bayes Classifier, SMO, J48 pruned tree, and Optimized Parametric Multilayer Perceptron, did the prediction with an accuracy of 87.8407%, 95.3729%, 100%, and 99.5657%, respectively. Further, the experimentation can be extended to predict the level of occurrence and the occurrence of other health complications because of this disease.

References

[1] Abbas, H., Alic, L., Rios, M., Abdul-Ghani, M., &Qaraqe, K. (2019). Predicting diabetes in healthy population through machine learning. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 2019-June, 567–570. https://doi.org/10.1109/CBMS.2019.00117

[2] Al-Zebari, A., &Sengur, A. (2019). Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection. 1st *International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC* 2019 - Proceedings, 2–5. https://doi.org/10.1109/UBMYK48245.2019.8965542

[3] Bandaru, S. R., &Kamepalli, S. (2019). Artificial intelligence: applications, framework and concerns-BFSI. *Journal of Advanced Research in Dynamical and Control Systems, 11*(9). https://doi.org/10.5373/JARDCS/V11I9/20192766

[4] Driss, K., Boulila, W., Batool, A., & Ahmad, J. (2020). A Novel approach for classifying diabetes' patients based on imputation and machine learning. 2020 *International Conference on UK-China Emerging Technologies, UCET* 2020, 14–17. https://doi.org/10.1109/UCET51115.2020.9205378

[5] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*,*8*, 76516–76531. https://doi.org/10.1109/ACCESS.2020.2989857

[6] Jacob, S.M., Raimond, K., &Kanmani, D. (2019). Associated machine learning techniques based on diabetesbased predictions. 2019 *International Conference on Intelligent Computing and Control Systems, ICCS 2019, (Iciccs)*, 1445–1450. https://doi.org/10.1109/ICCS45141.2019.9065411

[7] Kohli, P.S., & Arora, S. (2018). Application of machine learning in disease prediction. 2018 4th *International Conference on Computing Communication and Automation, ICCCA* 2018, 2018–2021. https://doi.org/10.1109/CCAA.2018.8777449

[8] Mir, A., &Dhage, S.N. (2018). Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare. Proceedings - 2018 4th *International Conference on Computing, Communication Control and Automation, ICCUBEA* 2018. https://doi.org/10.1109/ICCUBEA.2018.8697439

[9] Pethunachiyar, G. A. (2020). Classification of Diabetes Patients Using Kernel Based Support Vector Machines. 2020 *International Conference on Computer Communication and Informatics (ICCCI -2020)*, 22–25. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9104185

[10] Rout, M., & Kaur, A. (2020). Prediction of Diabetes Risk based on Machine Learning Techniques. *Proceedings of International Conference on Intelligent Engineering and Management*, ICIEM 2020, 246–

251. https://doi.org/10.1109/ICIEM48762.2020.9160276

[11] Saha, P.K., Patwary, N.S., & Ahmed, I. (2019). A widespread study of diabetes prediction using several machine learning techniques. 2019 22nd *International Conference on Computer and Information Technology, ICCIT* 2019, (December), 18–20. https://doi.org/10.1109/ICCIT48885.2019.9038559

[12] Sarwar, M.A., Kamal, N., Hamid, W., & Shah, M. A. (2018). Prediction of diabetes using machine learning algorithms in healthcare. ICAC 2018 - 2018 24th *IEEE International Conference on Automation and Computing: Improving Productivity through Automation and Computing*, (September), 6–7. https://doi.org/10.23919/IConAC.2018.8748992

[13] Sonar, P., & Jaya Malini, K. (2019). Diabetes prediction using different machine learning approaches. Proceedings of the 3rd *International Conference on Computing Methodologies and Communication, ICCMC* 2019, (Iccmc), 367–371. https://doi.org/10.1109/ICCMC.2019.8819841

[14] Sujatha, K., & Srinivasa Rao, B. (2019). Recent Applications of Machine Learning : A Survey. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8*(6), 263–267. https://doi.org/Retrieval Number: F10510486C219 /19©BEIESP

[15] Tripathi, G., & Kumar, R. (2020). Early Prediction of Diabetes Mellitus Using Machine Learning. ICRITO 2020 - IEEE 8th *International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, 1009–1014. https://doi.org/10.1109/ICRITO48877.2020.9197832

[16] Vijiyakumar, K., Lavanya, B., Nirmala, I., & Sofia Caroline, S. (2019). Random forest algorithm for the prediction of diabetes. 2019 *IEEE International Conference on System, Computation, Automation and Networking, ICSCAN* 2019. https://doi.org/10.1109/ICSCAN.2019.8878802

[17] Yahyaoui, A., Jamil, A., Rasheed, J., &Yesiltepe, M. (2019). A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. 1st *International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC* 2019 - Proceedings, 1–4. https://doi.org/10.1109/UBMYK48245.2019.8965556.

[18] American Diabetes Association. (2019). 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2019. *Diabetes Care,* 42(Supplement 1), S13-S28 (Crossref).

[19] G. A. Ogunmola, B. Singh, D. K. Sharma, R. Regin, S. S. Rajest and N. Singh, "Involvement of Distance Measure in Assessing and Resolving Efficiency Environmental Obstacles," 2021 *International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 2021, pp. 13-18, doi: 10.1109/ICCIKE51210.2021.9410765.

[20] N.H. Cho, J.E. Shaw, S. Karuranga, Y. Huang, J.D. da Rocha Fernandes, A. W. Ohlrogge, and B. Malanda, "IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,"*Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, Apr. 2018 (Crossref).

[21] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition,"*Diabetes Res. Clin. Pract.*, vol. 157, Nov. 2019, Art. no. 107843 (Crossref).

[22] V. V. Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus—A machine learning approach," presented at the 2015 *IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 2015, pp. 122–127 (Crossref).

[23] R. Arora, "Comparative analysis of classification algorithms on different datasets using WEKA,"*International Journal of Computer Applications*, vol. 54, no. 13, 2012 (Crossref).