IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 5, May 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205275 |

Diabetes Prediction Using Machine Learning

S.P.Nivveydhaa[1], K.Nivanjitha[2], Pragati Shantaram Jadhav[3], Dr P.C.D.Kalaivaani[4]

UG Scholars, Department of Computer Science and Engineering, Kongu Engineering College, Tamil Nadu, India [123]

Assistant Professor (SLG), Department of Computer Science and Engineering, Kongu Engineering College, Tamil Nadu, India [4]

ABSTRACT: - Diabetes is regarded as the deadliest chronic condition that raises blood glucose levels. According to the International Polygenic Disease Federation, polygenic disease is a condition where the exocrine gland does not produce hypoglycemic agents. The side effects include damage to the excretory organs, which typically results in chemical analysis, damage to the eyes, which could cause vision loss, or an increased risk of cardiopathy or stroke. The five main factors that our study identifies as being predictive of type-2 diabetes are glucose, pregnancy, body mass index (BMI), the function of the diabetic pedigree, and age. With the help of the Random Forest algorithm and machine learning techniques, the goal of this article is to create a system that can accurately forecast a patient's risk of developing diabetes early on. Random Forest algorithms are a sort of ensemble learning technique that are frequently employed for both classification and regression applications. Comparing the accuracy level to other algorithms, it is higher and the accuracy is 89.17%. We contend that our model can be used to forecast type 2 diabetes reasonably and that it may be utilized as a supplement to current preventive strategies to lower the incidence of the disease and its costs.

KEYWORDS: Machine Learning, Attacks, DDoS, Metrices

I. INTRODUCTION

A sub field of artificial intelligence (AI) and computer science called machine learning focuses on using data and algorithms to simulate how humans learn, gradually increasing the accuracy of the system. When a machine learning system receives new data, it forecasts the outcome using the prediction models it has built using prior data. The amount of data utilised to forecast output, as a larger data set makes it easier to develop a model that predicts the outcome more precisely. The sample labeled data is given to the machine learning system as training material, and then it uses that information to predict the outcome in supervised learning. In supervised learning, mapping input and output data is the main objective. The foundation of supervised learning is supervision, just like when a pupil is studying under a teacher's supervision. One of the prime examples of supervised learning is spam filtering.

Computer picks up information without any human intervention in unsupervised learning. Unsupervised learning's objective is to reorganize the incoming data into fresh features or a collection of objects with related patterns. A learning agent in a reinforcement learning system receives a reward for each correct action and receives a penalty for each incorrect activity. With the help of this feedback, the agent automatically learns and performs better. The agent explores and engages with the environment during reinforcement learning. The types of machine learning techniques are

- Supervised Learning
- Unsupervised Learning

When there is uncertainty, supervised machine learning creates a model that makes predictions based on data. Using a known set of input data and known reactions to the data, supervised learning trains a model to give accurate predictions for the reaction to incoming data (output). Use supervised learning if you have known data for the outcome you are attempting to estimate. Supervised learning with classification and regression algorithms is used to develop machine learning models. Classification techniques determine discrete replies, such as whether an email is

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205275 |

spam or genuine, or whether a tumor is malignant or benign. The data is classified using classification models. Credit scoring, speech recognition, and medical imaging are all common applications. Continuous reactions, such as those for difficult-to-measure variables, are forecasted using regression algorithms.

Unsupervised learning is a sort of machine learning in which models are trained using unlabeled datasets and are permitted to act on that data unsupervised. Because unlike supervised learning, we have the input data but no corresponding output data, unsupervised learning cannot be used to solve a regression or classification problem directly. Finding the dataset's underlying structure, classifying the data into groups based on similarities, and representing the dataset in a compressed way are the objectives of unsupervised learning. A machine learning technique called unsupervised learning uses training datasets without supervising the models. Instead, models themselves decipher the provided data to reveal hidden patterns and insights. It is comparable to learning that occurs in the human brain while learning new things. Objects are grouped into clusters using the clustering technique so that those with the greatest similarity stay in their own group and share little to no similarity with those in other groups. The data objects are classified based on the existence or lack of commonalities discovered by cluster analysis. An unsupervised learning technique called an association rule is used to uncover the connections among the variables in a sizable database. It establishes the group of elements in the collection that are present together. Marketing strategy is more effective because to the association rule. People who buy X (let's say, bread) also frequently buy Y (let's say, butter or jam).

II.DIABETES PREDICTION

Diabetes is a major severe illness. According to the World Health Organization, 422 million people globally, especially in less-income and middle-income countries, have diabetes. This number could rise to 490 billion by the year 2030. However, numerous countries, like Canada, China, and India among others, have higher rates of diabetes. India currently has over 100 million people, however there are around 40 million people with diabetes.

Types of Diabetes

There are two kinds of image segmentation that can resolve problems that are sophisticated for the human brain in a precise and efficient manner.

- Type-I
- Type-II

Type-I

Diabetes type 1 is a chronic illness also referred to as juvenile diabetes or insulin-dependent diabetes. The pancreas produces little or no insulin in this situation. Insulin is a hormone that the body utilizes to let glucose (sugar) into cells where it can be used to make energy. Type 1 diabetes may be brought on by a variety of factors, including genetics and some viruses. While type 1 diabetes typically first manifests in children, it can also strike adults. There is still no cure for type 1 diabetes, despite much research. The goal of treatment is to prevent problems by controlling blood sugar levels with the use of insulin, food, and lifestyle changes.

Type 1

Diabetes symptoms can appear suddenly and may include:

- Feeling more thirsty than usual
- Urinating a lot
- Bed-wetting in children who have never wet the bed during the night
- Feeling very hungry
- Losing weight without trying

Type-II

A disorder in the body's ability to control and utilise sugar (glucose) as fuel is type 2 diabetes. This chronic (long-term) disorder causes the bloodstream to circulate with an excessive amount of sugar. Over time, cardiovascular, neurological, and immune system issues might result from excessive blood sugar levels. Although type 1 and type 2

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205275 |

diabetes can start in childhood and adulthood, respectively, type 2 diabetes used to be classified as adult-onset diabetes. Although type 2 is more prevalent in elderly adults, type 2 instances have increased in younger people as a result of the rise in childhood obesity. Although treatment for type 2-diabetes is not found, you can manage the condition by decreasing weight, eating healthfully, and exercising. When signs and symptoms are present, they may include:

- Dehydration
- Urinating frequently
- Excessive appetite
- Unintended loss of weight

III.EXISTING SYSTEMS

S. Malik, S. Harous, and H. El-Sayed (2020) A comparison of machine learning algorithms for the early detection of diabetes in women. This prediction model used ten different machine learning algorithms, including Naive Bayes,BayesNet,Decision Tree, Random Forest, Adaboost Bagging, K-Nearest Neighbor, Support Vector Machine, Logistic Regression, and Multi-Layer Perceptron, with data from the Frankfurt hospital (Germany), demonstrating that K-Nearest Neighbor, Random Forest, and Decision Tree have the highest accuracy. Nonso Nnamoko, Abir Hussain, and David England reported forecasting diabetes onset: an ensemble supervised learning strategy in which they used five widely used classifiers for the ensembles and a meta-classifier to aggregate their outputs. The findings are reported and compared to previous research that used the same dataset. It is demonstrated that the proposed technique may predict diabetes onset with greater accuracy.

K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline(2019) proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly. This work was proposed by Md. Faisal Faruque, Asaduzzaman, and Iqbal H. Sarkern(2019), who used four famous machine learning methods, namely Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), and C4.5 Decision Tree, on adult population data to predict diabetes mellitus. Our experimental results suggest that the C4.5 decision tree outperformed other machine learning algorithms in terms of accuracy.

IV.PROPOSED SYSTEM

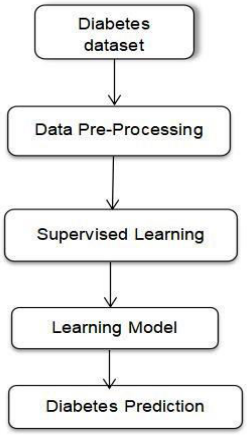

Figure 1. Architecture of proposed system

Figure 2 depicts Architecture of the proposed work. In this paper, the dataset is selected of these cases from a wider database was subject to several criteria. All patients here, in particular, are Pima Indian women over the age of 21.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205275 |

Pre-Processing:

It is the process of encoding the input data into a more stable state before feeding it to the algorithm.
Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model.Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

A. Random Forest Algorithm:

A well-known machine learning algorithm Random Forest is a technique used in the supervised learning process. It can be used to solve ML problems involving both classification and regression. It is based on the concept of ensemble learning, which is a method of combining several classifiers to address tough difficulties and improve model performance. Random Forest is a classifier that employs numerous decision trees on distinct subsets of the input dataset and averages the results to improve the projected accuracy of the dataset. The increasing number of trees in the forest prevents higher accuracy and over-fitting.

B. Support Vector Machine:

The Support Vector Machine (SVM) approach of supervised machine learning is utilised for both classification and regression. Even when we account for regression issues, classification is the best fit. The SVM method's goal is to find a hyperplane in an N-dimensional space that clearly categorises the data points. The number of features influences the size of the hyperplane. Support Vector Machine works by mapping the data to a high dimensional feature space in order to classify the data points even when they are not otherwise linearly separable. After identifying a separator between the categories, the data is changed to allow for the separator's hyperplane representation**.**

A.Performance Metrix:

Various parameters are employed during the evaluation, including

- Accuracy
- Precision
- Recall
- F1 Score

Some basic terms associated with performance evaluation are,

1. True positive: A state where both the expected and actual values are positive.
2. True negative: A state where the expected and actual values are opposite to each other.
3. False Positive: A state in which the value expected is positive, but the value actually obtained is negative.
4. False Negative: A state where the actual value is positive and expected value is negative.

B. Accuracy:

The percentage of correctly predicted outcomes used to measure the classification model's accuracy. Accuracy is a useful metric to use when classes are balanced or when the proportion of instances of each class is nearly equal. However, it should be observed that accuracy is not a valid statistic for datasets with class imbalance.

$$Accuracy = \frac{\text{True positive + True negative}}{\textit{All samples}}$$

C.Precision:

Out of all the positive predictions, precision shows which one are actually positive. It is defined as a ratio of correctly positive predictions to all of the positive predictions in the dataset.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True positive + False positive}}$$

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205275 |

D. Recall:

Recall shows what proportion of actual positive values are positive. The ratio of precise positive predictions to all positive predictions in the dataset is measured.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

E. F1 *Score:*

The harmonic mean of both precision and recall is the F1 Score. It achieves its maximum value of 1(perfect precision and recall) and its minimum value of 0 respectively.

$$\text{Precision} = 2* \frac{\text{Precision * recall}}{\text{Precision +recall}}$$

V. RESULT AND DISCUSSION

A. Dataset Description:

The selection of these cases from a wider database was subject to several criteria. All patients here, in particular, are Pima Indian women over the age of 21.

Accuracy:

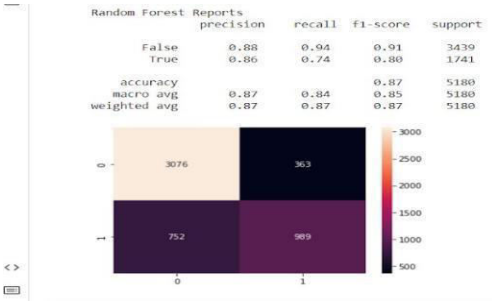

Figure 2. Random Frest Accuracy

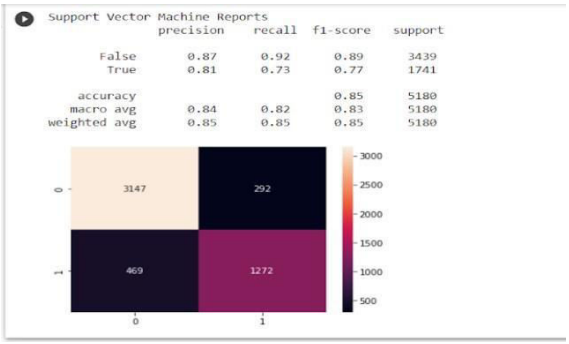

Figure 3. SVM Accuracy

Figure 3 & Figure 4 displays the accuracy comparison of the ML approaches. Of these algorithms, it has been demonstrated that the Random Forest Algorithm is the most accurate.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205275 |

IV. CONCLUSION AND FUTURE WORK

The diagnosis of diseases is aided by the use of machine learning and data mining techniques. Early diabetes detection is essential for determining the patient's appropriate course of therapy. This study compares the precision of a few widely used classification methods for the medical diagnosis of diabetic patients. A classification problem exists with the correctness expressions. The diabetes dataset was used to train and evaluate two machine learning algorithms on a test dataset. The outcomes of our model implementations show that RF outperforms the other models. The outcomes of association rule mining show that diabetes. This study has a limitation because it has employed a structured dataset. With the use of cutting-edge computational techniques and the availability of a sizable number of epidemiological and genetic diabetes risk datasets, machine learning has the potential to completely transform the ability to forecast the risk of developing diabetes. The major goal is to develop and apply a machine learning system for predicting diabetes and to evaluate the method's effectiveness is substantially correlated with BMI and glucose levels.

REFERENCES

[1] Meng, Xue-Hui, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu. The Kaohsiung journal of medical sciences 29, no. 2 (2013): 93-99.

[2] Kandhasamy, J. Pradeep, and S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus." Procedia Computer Science 47 (2015): 45-51

[3] Malik S., Harous S., El-Sayed H. Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women.Proceedings of the International Symposium on Modelling and Implementation of Complex Systems; October 2020; Algeria. Springer; pp. 95–106.

[4] Mohebbi A., Aradóttir T. B., Johansen A. R., Bengtsson H., Fraccaro M., Mørup M. A Deep Learning Approach to Adherence Detection for Type 2 Diabetics. Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 2017; Korea. pp. 2896–2899.

[5] D. Çalişir and E. Doğantekin, “An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier,” Expert Syst. Appl., vol. 38, no. 7, pp. 8311–8315, 2011.

[6] Mohebbi A., Bengtsson H., Fraccaro M., Mørup M.. Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 2017; Korea. pp. 2896-2866

[7] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13

[8] A. M. Rajeswari, M. Sumaiya Sidhika, M. Kalaivani and C. Deisy, "Prediction of Pre-Diabetes using Fuzzy Logic Baesd Association Classification", India Proceedings of the (ICICCT 2018).

[9] Deepti Sisodiaa and Dilip Singh Sisodiab, "Prediction Diabetes and Classification Algorithm", International Conference on Computational Intelligence and Data Science.

[10] D. Menon, K. Schwab, D. W. Wright and A. I. Maas, Position statement: definition of traumatic brain injury Arch. Phys. Med. Rehabil, vol. 91, pp. 1637-40, Nov 2010.

[11] Aishwarya Mujumdar and V Vaidehi, “Diabetes Prediction using Machine Learning Algorithms”, Procedia Computer Science, vol. 165, pp. 292-299,2019.

[12] Fikirte Girma Wolde Michael and Sumitra Menaria, “Prediction of Diabetes using Data Mining Techniques Dept of Computer Science and Engineering", Proceedings of the 2nd International conference on Trends in Electronics and Informatics (ICOEI 2018).

[13] Terry Jacob Mathew and Elizabeth Sherly, Analysis Supersived Learning Techniques for Cost Effective Diaease Prediction using Non-Clinical Parameters, Trivndrum, July 2018.

[14] J. Vijayashree and J. Jayashree, “An Expert System for the Diagnosis of Diabetic Patients using Deep Neural Networks and Recursive Feature Elimination”, International Journal of Civil Engineering and Technology, vol. 8, pp. 633-641, Dec.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details