IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

DIABETES PREDICTION USING MACHINE LEARNING

Granapoo D Ponmani, Dr. Jai Ruby MCA., M.Phil., Ph.D.

Department of Computer Applications, Sarah Tucker College, Tirunelveli-7.

Abstract

Diabetes is a medical disorder that impacts how well our body uses food as fuel. Most food we eat daily is converted into sugar, commonly known as glucose, and then discharged into the bloodstream. Our pancreas releases insulin when the blood sugar levels rise. Diabetes can cause blood sugar levels to rise if it is not continuously and carefully managed, which raises the chance of severe side effects like heart attack and stroke. Some of the symptoms of diabetes are: feeling more thirsty than usual, urinating often, losing weight, feeling tired and weak, having blurry vision. Due to age, lack of exercise, bad diet, high blood pressure etc., can cause this disease. Diabetes is a disease that has no permanent cure; hence early detection is required. Machine learning (ML) algorithms are used in diabetes prediction in this research. We used the Pima Indian Diabetes (PID) dataset for this research, collected from the Kaggle Machine Learning Repository. This dataset contains information about 768 patients and their corresponding nine unique attributes. We used two machine learning algorithms on the dataset to predict diabetes. We found that the model with K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) works well on diabetes prediction.

I.INTRODUCTION

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. If the blood sugar level is raised then the pancreas releases insulin. Our body, either doesn't make enough insulin or can't effectively use the insulin it makes this condition is called diabetes. Insulin is a hormone that regulates blood glucose. Hyperglycemia is also called raised blood glucose or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves, blood vessels, eyes, kidneys, and other organs. But we educating information about diabetes and taking steps to prevent or manage, it can help to protect our health. In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2019, diabetes was the direct cause of 1.5 million deaths and 48% of all deaths due to diabetes occurred before the age of 70 years. Another 460 000 kidney disease deaths were caused by diabetes, and raised blood glucose causes around 20% of cardiovascular deaths. Between 2000 and 2019, there was a 3% increase in age-standardized mortality rates from diabetes. In lower-middle-income countries, the mortality rate due to diabetes increased 13%. By contrast, the probability of dying from any one of the four main non communicable diseases (cardiovascular diseases, cancer, chronic respiratory diseases or diabetes) between the ages of 30 and 70 decreased by 22% globally between 2000 and 2019.

Machine Learning Algorithm:

Machine Learning algorithms are the programs that can learn the hidden patterns from the data, predict the diabetes, and improve the performance from experiences on their own. Different algorithms can be used in machine learning for different tasks, such as logistic regression that can be used for prediction problems like diabetes prediction. Machine Learning Algorithm can be broadly classified into three types: Supervised Learning Algorithms, Unsupervised Learning Algorithms, Reinforcement Learning algorithm. In this research, we use supervised learning algorithm. Supervised learning is a type of Machine learning in which the machine needs external supervision to learn. The

supervised learning models are trained using the labeled dataset. Once the training and processing are done, the model is tested by providing a sample test data to check whether it predicts the correct output.

Classification:

Classification in machine learning is one of the most common and widely used supervised machine learning processes, where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data. A classification problem is used to identify specific categories of new observations based on one or more independent variables. Also, in this article, we will focus on classification. It helps in categorizing data into different classes and has a broad array of applications, such as diabetes prediction, email spam detection, medical diagnostic test, fraud detection, image classification, and speech recognition among others.

Prediction:

Machine learning prediction, or prediction in machine learning, refers to the output of an algorithm that has been trained on a historical dataset. The algorithm then generates probable values for unknown variables in each record of the new data. The purpose of prediction in machine learning is to project a probable data set that relates back to the original data. This helps organizations predict future customer behaviours and market changes. Essentially, prediction is used to fit a shape as closely to the data as possible. To gain the most success with prediction in machine learning, organizations need to have infrastructure in place to support the solutions, and high-quality data to supply the algorithm.

Prediction can be used to forecast the future and to predict the probability of an outcome. It can also be used to forecast future requirements or run a what-if analysis. Predictive analytics is when data is used to predict future trends or events. With predictive analytics, historical data is used to forecast potential scenarios and use these predictions to drive strategic business aimed decisions.

II.LITREATURE REVIEW

In [1], the authors have provided a comprehensive overview of various diabetic prediction models based on machine learning algorithms. The authors discuss the different types of models, their performance, and the challenges associated with developing and implementing these models.

In [2], the authors have proposed a novel approach to diabetes prediction using machine learning and explainable AI techniques. The authors develop a model that can not only predict the risk of diabetes but also explain the reasons behind its predictions.

In [3], the authors have surveyed various machine learning approaches that have been used for diabetes risk prediction. The authors discuss the strengths and weaknesses of different algorithms and provide recommendations for future research.

In [4], the authors have compared the performance of different machine learning algorithms for diabetes prediction. The authors find that random forests and logistic regression are the most accurate algorithms for this task.

In [5], the authors have discussed the use of machine learning for early diabetes diagnosis and risk stratification. The authors review different machine learning algorithms and their applications in diabetes prediction.

In [6], the authors have proposed a hybrid machine learning approach for diabetes prediction. The authors combine different machine learning algorithms to improve the accuracy of the model.

In [7], the authors have reviewed the use of deep learning for diabetes mellitus prediction. The authors discuss the different deep learning architectures and their applications in diabetes prediction.

In [8], the authors have proposed an interpretable machine learning model for diabetes prediction. The authors use explainable AI techniques to make the model more transparent and easier to understand.

In [9], the authors have discussed the use of machine learning for personalized diabetes management. The authors review different machine learning approaches that can be used to tailor diabetes treatment to individual patients.

In [10], the authors have discussed the future of machine learning in diabetes care. The authors discuss the potential of machine learning to improve diabetes diagnosis, treatment, and prevention.

III.METHODOLOGY

Dataset Description:

The Pima Indian Diabetes Dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information of 768 women from a population near Phoenix, Arizona, USA. The outcome tested was Diabetes, 258 tested positive and 500 tested negative. Therefore, there is one target (dependent) variable and the 8 attributes: pregnancies, OGTT (Oral Glucose Tolerance Test), blood pressure, skin thickness, insulin, BMI (Body Mass Index), age, pedigree diabetes function. The Pima population has been under study by the National Institute of Diabetes and Digestive and Kidney Diseases at intervals of 2 years since 1965. As epidemiological evidence indicates that T2DM results from interaction of genetic and environmental factors, the Pima Indians Diabetes Dataset includes information about attributes that could and should be related to the onset of diabetes and its future complications.

1.Data collection

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. The data collection component of research is common to all fields. Here, in this research we are using Pima Indian Diabetes dataset and it is collected from Kaggle Machine Learning Repository. It consists of several analyzing variables and one target variable which is dependent to that. The main objective of the research is to predict whether the patient having diabetes or not.

2. Data Pre-processing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Data pre-processing is a process of preparing the raw data and making it suitable for machine learning model. It is the first crucial step while creating a machine learning model. The acquired data may suffer from inconsistency in data owing to the presence of missing values. To acquire suitable outcomes on final proposed systems, it is essential for doing the pre-processing stage to acquire the better efficiency for the model. Moreover, it is the essential and primary process of any prediction or classification model. This process is also known as the pre-handling of data. It is simpler process and also makes easier for interpretation. Several processes of eliminating the duplicates or inconsistencies in data to maximize the accuracy of the prediction result. It also eradicates the process of correcting the missing values or incorrect data occurred due to the human errors.

- **Data Standardization**

  This is one of the important steps in data pre-processing. In our dataset had a different range of all these values, it will be difficult for our machine learning model to make some prediction. So, we will try to standardize the data in a particular range.

- **Outliers Removing**

  Secondly, outlier handling is applied, it's the most important for any machine learning prediction model, as outliers affect the prediction model results. A Boxen plot is used here to view the outliers in the way of a graphical representation.

3. Feature Selection

Feature selection is one of the important concepts of machine learning, which highly impacts the performance of the model. As machine learning works on the concept of "Garbage In Garbage Out", so feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features. While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant. If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model. Hence it is very important to identify and select the most appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning. Feature selection is performed by either including the important features or excluding the irrelevant features in the dataset without changing them. Supervised Feature selection techniques consider the target variable and can be used for the labelled dataset.

Pearson's Correlation Coefficient:

The Pearson's correlation coefficient(r) is the most common way of measuring a linear correlation. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

- r = Pearson Coefficient
- n= number of pairs of the stock
- $\sum xy$ = sum of products of the paired stocks
- $\sum x$ = sum of the x scores
- $\sum y$= sum of the y scores
- $\sum x^2$ = sum of the squared x scores
- $\sum y^2$ = sum of the squared y scores

4. Data Splitting

Once we preprocess the data, we will split the data into two type training and testing data. So, we train our machine learning model with training data and then we try to find the accuracy score of our model with the help of test data. Dataset for the train and test are typically splitting in the ratio of 80:20.

5. Model Evaluation

Model evaluation is the crucial process that uses some metrics which help us to analyze the performance of the model. As we all know that model development is a multi-step process and a check should be kept on how well the model generalizes future predictions. This is called as evaluation. First of all, we need to understand the accuracy of model how well our model is performing. Only the accuracy is high we can use that particular model for our prediction. This step is called evaluation. To evaluate machine- learning model performance, consider the most common performance metrics such as accuracy, precision, recall, F1 score, support are the performance metrics help to evaluate the model.

Confusion Matrix:

The confusion matrix evaluates the performance of machine learning classification models. Utilizing the confusion matrix, each model was examined. The confusion matrix shows how frequently our models make correct and incorrect predictions.

- **T**rue **P**ositive (TP): when the predicted value and the actual value are both positive.
- **T**rue Negative (TN): when the predicted value and the actual value are both negative.

- **F**alse **P**ositive (FP): when the actual value was negative and the predicted value was positive
- **F**alse **N**egative (FN): when the actual value was positive and the predicted value was negative.

The different performance metrics are used to evaluate the performance of the ML classifiers are listed below:

Precision

Precision is defined as the *ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples* (either correctly or incorrectly).

$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$

Recall

The recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples. The ***recall measures the model's ability to detect positive samples***. The higher the recall, the more positive samples detected.

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

F1-Score

The **F1-score** combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is primarily used to compare the performance of two classifiers. Suppose that classifier A has a higher recall, and classifier B has higher precision. In this case, the F1-scores for both the classifiers can be used to determine which one produces better results.

$$\boldsymbol{F1 = 2.\frac{precision.recall}{precision + recall}}$$

Accuracy

Accuracy is defined as the ratio of the number of correct predictions to the total number of predictions. This is the most fundamental metric used to evaluate the model. The formula is given by

$$\textbf{Accuracy} = \frac{(\boldsymbol{TP+TN})}{(\boldsymbol{TP+TN+FP+FN})}$$

6. Classification

In this work, four different machine learning algorithms such as support vector machine, k nearest neighbor are used for classification of diabetes prediction.

- **Support Vector Machine:** Support Vector Machine is a supervised classification algorithm where we draw a line between two different categories to differentiate between them. SVM is also known as the support vector network. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. The hyperplane dimension needs to be changed from 1 dimension to the Nth dimension. This is called Kernel.
- **Types of SVM:**
  - **Linear SVM:**

    Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

- **Non-linear SVM:**

  Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

❖ **K-Nearest Neighbors:**

The k-nearest neighbor algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another. For these classification problems, a class label is assigned on the basis of a majority vote.

7. Data visualization

Data visualization is a crucial aspect of machine learning that enables analysts to understand and make sense of data patterns, relationships, and trends. Through data visualization, insights and patterns in data can be easily interpreted and communicated to a wider audience, making it a critical component of machine learning.

Data visualization helps machine learning analysts to better understand and analyze complex data sets by presenting them in an easily understandable format. Data visualization is an essential step in data preparation and analysis as it helps to identify outliers, trends, and patterns in the data that may be missed by other forms of analysis.

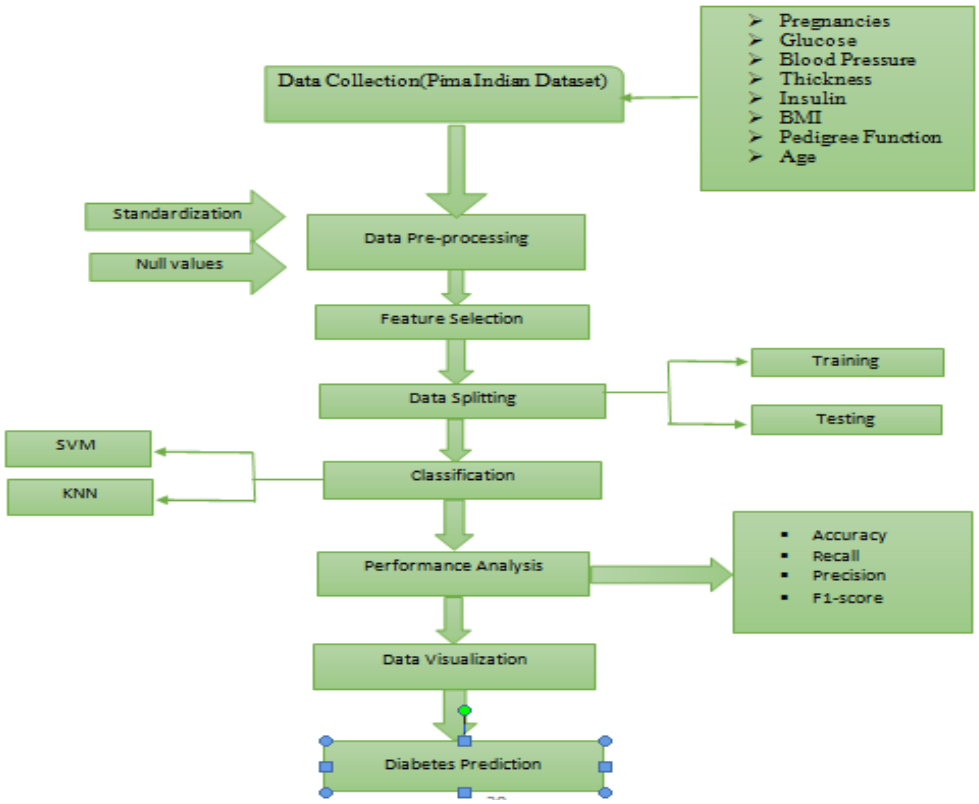
8. Descriptive statistics

In Descriptive statistics, we are describing our data with the help of various representative methods using tables, excel files, etc. In descriptive statistics, we describe our data in some manner and present it in a meaningful way so that it can be easily understood. Most of the time it is performed on small data sets and this analysis helps us a lot to predict some future trends based on the current findings. Some measures that are used to describe a data set are measures of central tendency and measures of variability or dispersion.

9. Diabetes Prediction

It is more beneficial to identify the early symptoms of diabetes than to cure it after being diagnosed. There is several numbers of methods, models and researches existed, it is complicated to analyze which classifier or system is best for getting the final diabetes outcomes. Every system is dependent on a specific attribute dataset and thus, it's is essential for designing a reliable model for adapting to each attribute dataset. Moreover, the classifiers are helpful in predicting the precise results for getting the outcomes.

Thus, this work uses two classifiers like Support Vector Machine and K-Nearest Neighbor and then evaluate the best one among them for further research.



IV. RESULT AND DISCUSSION

In this research paper, we used two machine learning classification model were applied on the dataset (Pima Indian Diabetes) to predict diabetes. We found that the model with K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) works well on diabetes prediction. Table1 represent the result of SVM and KNN. In this research, we used SVM classifier with Linear kernel and KNN classifier with 8 number of neighbors. We found accuracy, precision, recall and f1-score.

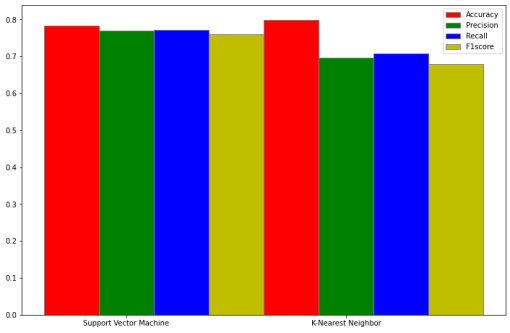
Result:

Table1: Performance metrics of SVM and KNN

S.NO	Classification Model	Accuracy	Precision	Recall	F1-Score
1	SVM	0.7833	0.7704	0.7727	0.7604
2	KNN	0.7996	0.6972	0.70779	0.6800

As seen from the above table, the accuracy of SVM with Linear kernel is 78% and the accuracy of KNN is 80%.

Bar chart:



V.CONCLUSION

Diabetes is the leading death cases in women worldwide. Early prediction is more important to recover the patients or diet and an exercise is also more important for everyone. We have proposed a prediction model, which is specifically designed for prediction of diabetes using machine learning algorithms such as Support Vector Machine and K-Nearest Neighbor algorithms. The model predicts the person had diabetic or the person is non diabetic person. The model uses supervised learning which is a machine learning concept where we provide dependent and independent columns to machine. It uses classification technique which predicts the diabetes. In our research we had 78% for support vector machine and 80% for K-nearest neighbor algorithms. The goal of the research is to classify the patients to predict diabetes by classification technique with higher accuracy.

REFERENCES

1) Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A Comprehensive Review of Various Diabetic Prediction Models: A Literature Survey. Journal of Healthcare Engineering, 2022, 8100697. https://doi.org/10.1155/2022/8100697

2) Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2022). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(12), 12039. doi:10.1049/htl2.12039

3) Birjais, A. M., Sadhu, A., Verma, N., & Singh, A. K. (2022). A survey on diabetes risk prediction using machine learning approaches. Journal of Family Medicine and Primary Care, 11(11), 6929-6934. doi:10.4103/jfmpc.jfmpc_502_22

4) Faraz, S., & Singh, P. (2022). Diabetes Prediction using Machine Learning Algorithms. Journal of Applied Science and Education, 2(2), 1-12.

5) Kwon, S. Y., & Rhee, H. W. (2022). Machine Learning for Early Diabetes Diagnosis and Risk Stratification. Journal of Clinical Medicine, 11(7), 1980. doi:10.3390/jcm11071980

6) Jain, A., & Thakur, N. (2021), A Hybrid Machine Learning Approach for Diabetes Prediction, Procedia Computer Science, 182, 206-211.

7) Zhu, T., Li, K., & Georgiou, P. (2021). Deep Learning for Diabetes Mellitus Prediction: A Systematic Review. IEEE Journal of Biomedical and Health Informatics, 25(7), 2744-2757. doi:10.1109/JBHI.2020.3040225

8) Hussain, T., Khan, H., & Ullah, R. (2022). An Interpretable Machine Learning Model for Diabetes Prediction, Entropy, 24(4), 557. doi:10.1016/j.entropy.2022.04.

9) Han, P., & Wang, J. (2022). Machine Learning for Personalized Diabetes Management. IEEE Journal of Biomedical and Health Informatics, 26(1), 148-157.

10) Huang, Y., & Powers, D. A. (2022). The Future of Machine Learning in Diabetes Care. Diabetes Care, 45(10), 2303-2309. doi:10.2337/dc22-0542.