WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Diabetes Prediction Using Data Mining Techniques

Dharadevi S

Sri Krishna Arts and Science College

Abstract:-

Now-a-days, there are endless data mining techniques to predict and analyse the data mining algorithms. In healthcare, the difficult level is ultimately increasing the work to the doctors who are handling the decisions to update the patient's health and treatment on huge data. In this paper, we discuss data mining methods that should be analysed in research work. For experimental approach, the patient medical records and their data are added with different data mining algorithms and their accuracy are measured by the performance analysis. We may use algorithms like K-Nearest Neighbors, Support Vector Machine with Radial Basis Function Classifier to predict whether the patient is having diabetics or not. The aim of this paper to be useful for the doctor to predict the early stages of diabetics using data mining techniques.

Keywords— Predictive Analysis, classification Algorithms, Healthcare, Diabetes, Data Mining, K Nearest Neighbor, Support Vector Machine.

I. INTRODUCTION

With rapid refinement occurring all across the world, innovation is critical. Data Mining is booming in the healthcare industry. Diabetes mellitus is a chronic condition that is fairly frequent these days. It happens for a variety of reasons, but the final result is that the required amount of insulin is not reached for the human body to meet its glucose needs. Normal production of these hormones maintains normal blood sugar levels in the body, ranging from 70 to 180 mg / dl (4.0 to 7.8 mmol / L). One of the symptoms of diabetes is high blood glucose levels and is caused by insulin production. Insulin resistance is a major cause of this type of diabetes. It is a chronic disease, and screening is necessary because it is important to keep the levels within a certain range because it has such a profound effect on the body. In the data provided, we used modelling machine learning models to predict diabetes. Today, Data Mining operates in a variety of fields, including agriculture, health care, security, and banking. Diabetes is a chronic condition marked by elevated blood glucose levels in the body. Diabetes causes damage to the kidneys, eyes, and heart over time. The International Diabetes Federation estimates that diabetes affects 382 million people worldwide. The number will increase to 592 million by the end of the century. Medical specialists find it difficult to predict diabetes early in most cases. Diabetes may be divided into three categories:

Type 1 Diabetic:- Also known as juvenile diabetes or insulin-dependent diabetes, occurs when our body's immune system damages insulin-releasing cells, eventually removing insulin production from the body 11.

Type 2 Diabetic:- Also known as insulin-dependent diabetes, arises when the body becomes resistant to insulin or stops producing it. It can strike anyone at any age11. A type of diabetes that develops during pregnancy is called gestational diabetes.

II. LITERATURE REVIEW

Previous research demonstrates a wide range of conclusions about the examination of healthcare data utilising various methodologies and procedures. Several researchers have explored and developed many predictions and analytical models using various data mining, data management, or a combination of these strategies. Sneha, N, Gangil, T [1] in the paper titled “Analysis of diabetes mellitus for early prediction using optimal features selection”, where they are using five different data mining algorithms to predict diabetes. Their result came up with higher accuracy of decision tree and random forest.

Ambilwade [2] added that by measuring blood glucose levels, we can predict the risk of diabetes with the role of Fuzzy Inference System and Multilayer perceptron (MLP) and type 2 diabetes. measures various aspects of mathematics.

Maniruzzaman, M, Rahman, M.J., Ahammed, B [3]in the paper titled “Classification and prediction of diabetes disease using machine learning paradigm”, the combination of LR and RF-based classifiers performs better. Predicting diabetic patients will be made much easier with this combination.

S. S. Bhat and G. A. Ansari [4] In this study, there are various types of datasets in the Diabetes dataset (DD) and it was mostly noted by the researcher. It was thoroughly verified by the researcher and it is considered as the best and build the model. It should ultimately be ready for the varies medical dataset (MDs). In the classification algorithm evaluation of cross validation is not being used and it is not evaluated.

Raja Priya K. [5] From this study, In the developing soft computing prediction model is used to find the diabetics risks and it is used to analysis the patient records. For this using a real time medical dataset and algorithm technique like genetics algorithm. It can be predicted and analysis the diabetic patient risk level and it is experimented by the result.

M. K. Hasan,, M.A.Alam[6] From this study conveyed that some various skills of data mining and analysis of diabetic prediction Using various technique J48 algorithm, Random Forest, Naive Bayes.

P. Suresh Kumar [7] In this study, for identifying different types of diabetes presented some algorithms like Decision Tree, SVM and data mining techniques.

III. PROPOSED METHODOLOGY

The purpose of the paper is to look into models that can more accurately forecast diabetes. To forecast diabetes, we tested various classification and ensemble methods. In the following, we briefly discuss the phase.

1). Dataset Description- These dataset is collected from the Pima Indian Diabetes dataset. Once the training phase has been completed. In this database we have used seventeen attributes and variations of one category. These class variants contain numerical values. In the data set there are a few missing outputs. Data sets are splitted into two, of these 80% are used for training and 20% are used for testing.

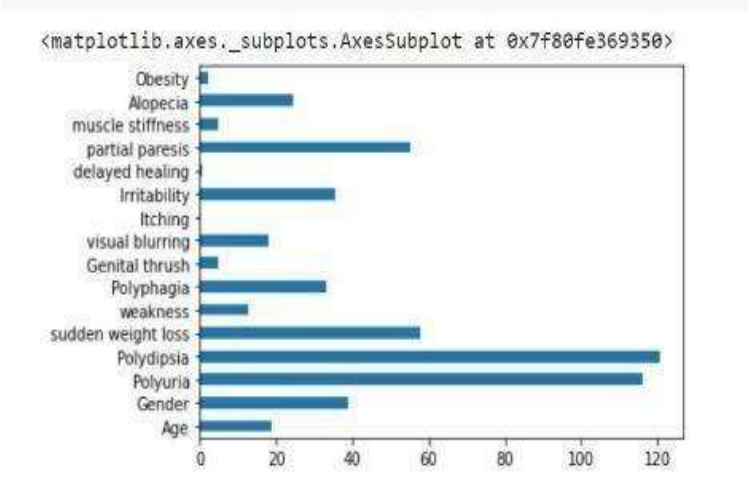
Other features of the database are:

Table 1: Dataset Description

S No.	Attributes
1	Polyuria
2	Polydipsia
3	Sudden Weight Loss
4	Polyphagia
5	Genital Thrush
6	Obesity
7	Delayed Healing
8	Muscle Stiffness
9	Alopecia
10	Partial Paresis

The sum of the importance of each feature playing major role for diabetes have been plotted, where X-axis represents the importance of each feature and Y-Axis the names of the features.

Figure 1: Feature Importance Plot



2). Data Pre-processing- The most crucial phase is data preprocessing. Most data pertaining to healthcare has missing values and other contaminants that can affect how useful the data is. Data preparation is done to enhance the quality and effectiveness of the results produced through mining. This procedure is crucial for accurate results and good prediction when applying machine learning techniques to the dataset. For the Pima Indian diabetes dataset, pre-processing must be done twice.

Missing Values removal- Eliminate all occurrences with a value of 0 (zero). Zero as a value is not conceivable. As a result, this instance is stopped. This procedure, known as features subset selection, decreases the dimensionality of the data by eliminating unnecessary features and instances. It also makes work go more quickly.

Splitting of data- Data is then utilized for model training and testing after being cleaned. After the data is spilt, we train the algorithm on the training data set while putting the test data aside. Based on the logic, methods, and values of the feature in the training data, this training process will create the training model. The primary goal of normalization is to put all qualities on the same scale.

3). Apply Data Mining- When the data is ready, machine learning techniques are used. To predict diabetes, we employ a variety of classification and ensemble algorithms. The procedures used on the diabetes dataset for Pima Indians. The main goal is to use machine learning techniques to examine the effectiveness of these methods, determine their accuracy, and identify the responsible/important characteristic that plays a significant part in prediction. The Techniques are follows-

a. Support Vector Machine- SVM is one of the most extensively used Supervised Learning techniques for Classification and Regression challenges. The majority of the time, it is utilized in machine learning to address categorization issues. The goal of the SVM method is to discover the best line or decision boundary for categorizing n-dimensional space into classes so that subsequent data points can be easily placed in the appropriate category. The Machine Learning uses a function called the radial basis kernel to non- linear regression lines. When m is much higher than n, the dot product can be found in higher dimensionally. The main idea is to use the Linux operating system. A regression curve in higher dimensions becomes a non- linear curve in lower dimensions.

Algorithm-

STEP 1: Select the hyper plane which divides the class better.

STEP 2: To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.

STEP 3: If the distance between the classes is low then the chance of miss conception is high and vice versa. STEP 4: Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.

b. K-Nearest Neighbor -The K-NN technique places the new case in the category that matches the current categories the most, presuming that the new case/data and old instances are similar. The K-NN approach keeps all current data and classifies new data points based on similarity. This means that as new data is generated, the K-NN approach can swiftly categorize it into a suitable category. The K-NN technique can be used for both regression and classification, however classification is the most popular application. K-NN is a nonparametric approach, meaning it does not make any assumptions about the data. The KNN method simply saves the dataset during the training phase and then classifies it into a category that is quite similar to the incoming:

$$d(P,Q) = \sum_{i=1}^{n} (P_i - Q_i)^2$$

Algorithm-

STEP 1: Take a sample dataset of Pima Indian diabetes dataset.

STEP 2: Take a sample of attributes and rows.

STEP 3: Find the Euclidean distance by the help of formula.

STEP 4: Then, pick a random value of K. is the number of nearest neighbors.

STEP 5: Then with the help of these minimum distance and Euclidean distance find out the nth column of each. STEP 6: Find out the same output values.

STEP 7: If the values are same, then the patient is diabetic, otherwise not.

IV. MODEL BUILDING

This is the most crucial phase, which includes developing a model for diabetes prediction. For the purpose of diabetes prediction, we have implemented various machine learning algorithms in this.

Procedure of Proposed Methodology-

STEP 1: Import diabetic dataset and necessary libraries.

STEP 2: Pre-process data to omit missing data in step two.

STEP 3: Divide the dataset by 80% into a training set and 20% into a test set.

STEP 4: Choose the machine learning algorithm (K-Nearest Neighbor, Support Vector Machine, etc.) at.

STEP 5: Create the classifier model based on the training set for the aforementioned data mining algorithm. STEP 6: Use a test set to evaluate the Classifier model for the aforementioned data mining algorithm.

STEP 7: Compare and evaluate the results of each classifier's experimental performance.

STEP 8: Select the highest performing algorithm after analysis based on various metrics.

V. RESULT AND DISCUSSIONS

The performance of the classification models is determined by the confusion matrix. It is possible to determine the true values for test data. The matrix is easy to understand, but the related terminologies may be confusing. An error matrix is when the model performance is shown in the form of a matrix.

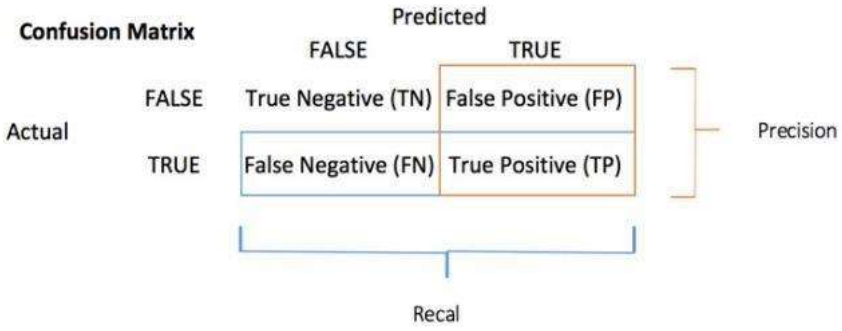


Figure 2: Outcomes of Matrix

STEP 1: First, the expected outcome values need to be tested**.**

STEP 2: Predict all the rows.

STEP 3: The expected predictions and outcomes should be calculated.

- The Correct predictions of each class.
- There were incorrect predictions of each class.

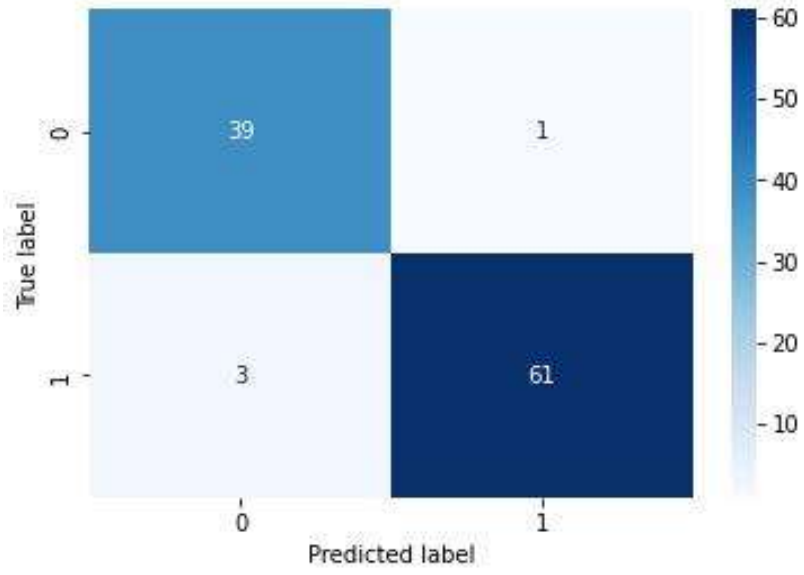


Figure 3: Result of KNN

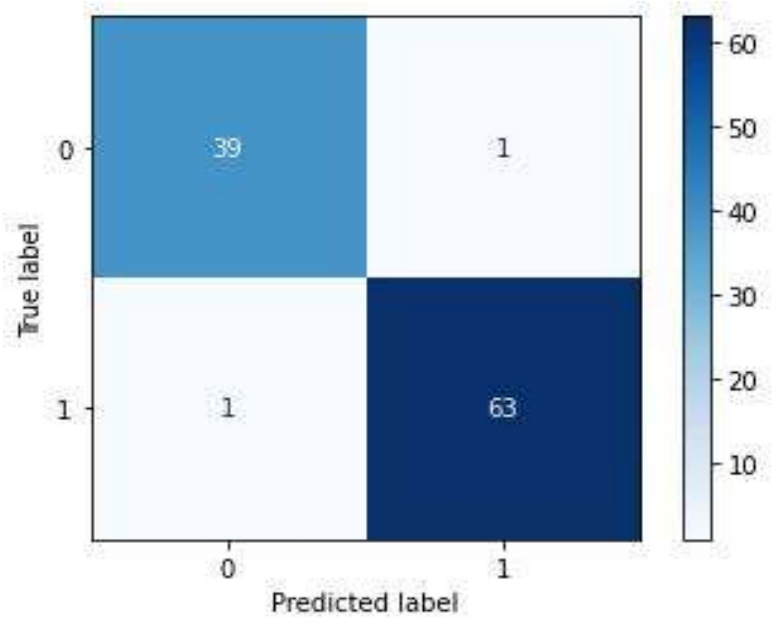


Figure 4: Result of SVM RBF

TABLE 2: OUTPUT OF ALGORITHMS

CLASSIFIER MODEL	ACCURACY
Support Vector Machine RBF	98.07%
KNN	98.08%

This project's primary goal was to build and implement methods for predicting diabetes using data mining, and to assess the effectiveness of those methods. The suggested strategy makes use of SVM, KNN, and other classifiers in an ensemble learning setting. Additionally, 90% categorization accuracy was attained. The experimental findings can help health care providers make early predictions and decisions to treat diabetes and save lives.

VI. CONCLUSION

One of the important real-world medical problems is the detection of diabetes at early stage. In this paper, systematic efforts are made to predict the diabetes with greater accuracy. During this work, two data mining classification algorithms are studied and evaluated on various measures. Experiments are performed on Pima Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 98% using KNN and SVM RBF algorithms.

VII. REFERENCES

1. Khan, F. A., Zeb, K., Al-Rakhami, M., Derhab, A., & Bukhari, S. A. C. (2021). “Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review”. doi:10.1109/access.2021.3059343
2. Aldallal, Ammar; Al-Moosa, Amina Abdul Aziz (2018). “Using Data Mining Techniques to Predict Diabetes and Heart Diseases”, [IEEE 2018 4th International Conference on Frontiers of Signal Processing (ICFSP) -. , (), 150–154. doi:10.1109/ICFSP.2018.8552051
3. Sharvani M S, Siddharth Warad, Fayaz M, Darshan N S. (2020), “Diabetes Prediction using Data Mining”, International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 08.
4. Aditya Saxena, Megha Jain, Prashant Shrivastava. (2021), “Data Mining Techniques Based Diabetes Prediction”, Indian Journal of Artificial Intelligence and Neural Networking (IJAINN) ISSN: 2582-7626 (Online), Volume-1 Issue-2.
5. G. Geetha, K.Mohana Prasad.(2020),”Prediction of Diabetics using Machine Learning”,International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-5.

6. Woldemichael, Fikirte Girma; Menaria, Sumitra (2018). “Prediction of Diabetes Using Data Mining Techniques” [IEEE 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) - Tirunelveli, India (2018.5.11-2018.5.12)]

7. Baiju, B.V.; Aravindhar, D. John (2019).” Disease Influence Measure Based Diabetic Prediction with Medical Data Set Using Data Mining”, [IEEE 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT) - CHENNAI, India (2019.4.25-2019.4.26)

8. Nahla B., Andrew et al. "Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

9. A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.

10. Nonso Nnamoko. Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

11. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabe- tes Disease Prediction Using Data Mining "International Con- ference on Innovations in Information, Embedded and Com- munication Systems (ICIIECS), 2017.