RESEARCH ARTICLE OPEN ACCESS

Diabetes Prediction by K Nearest Neighbour Algorithm

Dr Mohanapriya Jayapal*, Nisha S**, Dhivyadharshini K***, Praveen S****
Department of Biotechnology, KIT
Kalaignarkarunanidhi Institute of Technology, Coimbatore
India

ABSTRACT
Diabetes is a condition that can cause mortality, heart attacks, blindness, and renal failure. In 2019, 463 million people had diabetes, according to the International Diabetes Federation. This figure is expected to increase by 578 million by 2030 and reaches 700 million by 2045 if projections come to pass. As stated in a report released by the Republic of Indonesia's Ministry of Health, In Indonesia is one of the ten nations with the highest incidence of diabetes in 2019. To identify the type of diabetes, specialists are mandatory. Many individuals who are evaluated have a sickness that can be classified as severe because they are slow to identify what illness they have. To avoid serious situations, diabetes detection technology is necessary. Physicians can utilize it to swiftly and precisely diagnose illnesses in the modern medical field. As a result, we may apply machine learning to stop deaths by creating an artificially intelligent model that can forecast diabetes. KNN algorithm is used to train the model and evaluate. In order to predict diabetes based on a number of health characteristics in the dataset using supervised machine learning, the study uses the confusion matrix and ROC curve to evaluate the model. The confusion matrix is used for the classification of true positive, true negative, false positive and false negatives of the provided datasets. ROC curve is used to evaluate the performance of diabetes prediction.

Keywords **—** Diabetes, Machine learning, KNN, Machine learning

I. INTRODUCTION

Computers can become clever like humans thanks to a scientific field called machine learning, which automatically increases their comprehension via experience (Jack Billie Chandra et al.,2023). Systems that can learn on their own, such as making decisions without constant human programming, are the focus of this field of study. Furthermore, the machine is capable of adjusting to a dynamic environment. The four subcategories of machine learning are reinforcement learning, unsupervised learning, semi-supervised learning, and supervised learning (Sarker et al.,2020). One machine learning method for learning with labelled datasets is supervised learning, which finds input labels to generate classifications and prediction (Oza et al.,2022). These machine learning techniques are employed in disease prediction in heathcare. Likewise, Diabetes prediction using machine learning algorithm has various application (Garcia-Carretero et al.,2020). Diabetes mellitus is defined by elevated blood sugar levels. This serious condition impacts how the body manages blood sugar, resulting in persistently high blood sugar levels over time. Timely detection and assessment of diabetes are crucial to prevent or postpone the disease and its associated complications (Panwar et al.,2017). Machine learning algorithms have demonstrated encouraging outcomes in predicting diabetes. One such algorithm is the K Nearest Neighbour (KNN) algorithm (Nirmaladevi et al.,2020). The KNN algorithm is a non-parametric method used for classification and regression tasks (Swathi et al.,2017). In the diabetes prediction, KNN functions as a classification tool to determine whether an individual is at risk for diabetes based on clinical and lifestyle information . The KNN algorithm operates by identifying the closest neighbours of the new data and categorizing the new data points according to the class that most of their nearest neighbours belongs (Uddin et al.,2022). KNN is considered a lazy algorithm, as it does not need any training or learning phases. Instead, it retains all the training data and utilizes it to forecast new instances. To employ the KNN algorithm for diabetes prediction, clinical and lifestyle data from a representative sample are essential (Gupta et al.,2020). This information should encompass significant factors that contribute to the onset of diabetes. Additionally, the dataset must include the target variable, indicating whether the individual has been diagnosed with diabetes (Febrian et al.,2022). Following data collection, it is imperative to prepare the dataset for the KNN algorithm. This preparation involves addressing missing values, normalizing data, and encoding categorical variables (Habibi et al.,2015). The pre-processed dataset will then be divided into training and testing sets (Wu et al.,2018). The KNN algorithm undergoes training during this process, and its efficacy is evaluated through testing by confusion matrix and ROC curve. For each data point in the test set, the KNN algorithm computes the distance between the test point and every point in the training set (Wu et al.,2018). Once the distances are calculated, the algorithm proceeds to classify the test data based on the nearest neighbours.

II. RESEARCH PROBLEM

The research topic in this study is that 463 million people have diabetes in 2019 (Kavakiotis et al.,2017). This population is expected to increase by 578 million by 2030 and reach 700 million by 2045 if projections come true According

to a 2020 article by the Republic of Indonesia's Ministry of Health, Indonesia is among the top ten nations in the world for diabetes prevalence in 2019 (Olisah et al.,2022). It needs the expertise of professionals to identify the kind of diabetes. Many individuals who undergo examinations have a sickness that can be classified as severe because they are slow to identify their condition. For the purpose of determining which algorithm is most appropriate for diabetes prediction, the KNN algorithm is used. With eight independent variables and one dependent variable, the Diabetes dataset serves as the study's focus. Pregnancy, blood pressure, glucose, skin thickness, insulin, BMI, diabetes pedigree function, and age are the eight characteristics of diabetes that are employed in this study to assess whether a person has the disease or not. There are a lot of factors to take into account while diagnosing diabetes, thus a fast and effective approach is needed.

- to evaluate different machine learning algorithms' categorization accuracy if comparable data is bought.
- To understand the classification process of some machine learning techniques.
- To use a range of machine learning algorithms to classify diabetes
- Importance of the Research
- Provide suggestions to researchers and others who are considering classifying other diseases using machine learning.
- To increase readers' knowledge of the features of diabetes in order to contribute scientifically to the field of health sciences.

Machine learning algorithms can benefit from the information in this article.

III.MODEL PREPARARTION

A.DATA COLLECTION

This study's data is taken from a publicly available dataset called the Pima Indians Diabetes Database, which is sourced from Kaggle and may be accessed at https://www.kaggle.com/uciml/diabetes-database. The data contains 768 datasets with 8 attributes. There are two classes, class 1 and class 0, and seventy-eight attributes in the diabetes database dataset.

B.DATA PRE-PROCESSING

The first step in converting the incoming data into proper format and processing-ready data is called pre-processing. In order to clean, integrate, reduce, and discretize data, pre-processing may involve a variety of necessary procedures, such as merging, reshaping, or converting data. One process activity or a combination of the aforementioned processes might also make up the pre-processing process. Goals for pre-processing determine the current procedure. The proper technique must be chosen with the understanding that it will enhance classification performance throughout the data pre-processing phase. Therefore, data pre-treatment procedures should be followed in order to enhance the quality of the data that will be evaluated. The author has previously completed this pre-processing to provide high-quality data. Data integration, data cleaning, data reduction, and data transformation are among the procedures involved in this step.

C.DATA CLEANING

In this step, missing data (Missing Value) and incomplete, imprecise, and incorrect data are found. A real human being cannot have values of 0 for any of the parameters (glucose, blood pressure, skin thickness, insulin, and BMI) in this dataset, even though none of the columns have missing values (Table 1). At this stage, data integration is required to convert the original data's measurement scale into a different format so that the diabetic dataset can be read by the analytic tool. Nevertheless, the author does not utilize this stage at this time because the diabetes dataset already has valuable information for the analysis step.

Table 1: Number of missing value in the dataset

CHARACTERISTICS	COLUMN OF 0
Pregnancies	11
Glucose	5
Blood pressure	35
Insulin	374
BMI	11
Skin thickness	227
Diabetes pedigree function	0
Age	0

D.DATA SPILTTING

The data from the dataset is converted into training set data and testing set data. This step provides the efficient model testing and evaluation. About 80% of the data from the dataset is used as training dataset and 20% of the data is used as testing dataset from the diabetes dataset.

IV. MODEL TESTING AND EVALUATION

A.MODEL TESTING

Google Collaborator was utilized to replicate Jupyter Notebook in the cloud for the purpose of testing KNN model.

Thetest model from Google Collaborator.

- Importing the model that we must test in order to predict diabetes based on many health variables in the dataset.
- The study involved k-Nearest Neighbours algorithms employing Data pre-processing includes reading the dataset, viewing its contents, and filling in the blanks.
- Data scaling, also known as data normalization, is the process of converting the dataset's numerical values

to a common scale without distorting discrepancies in the range of values. Machine learning will learn more quickly if data is normalized and training and testing data are kept apart.

- The KNN model is being tested.
- Cross-validation is carried out to test the experiment and evaluate the KNN model using a confusion matrix to determine the accuracy, precision, and recall values.
- Using the percentage distribution of datasets, the accuracy, precision, and recall values from eight studies are visualized.

B.MODEL EVALUATION

Using a number of health characteristics, we have effectively predicted diabetes using the KNN. Between 80% and 10% of training data is used to calculate the proportion of attribute data. The classification accuracy of the KNN across eight experiments is 82%.

The table below shows that out of the eight tests, the first one using 80% training data had the best accuracy. A good accuracy of 83.2% was obtained by the KNN algorithm. The model is evaluated by its precision, accuracy and confusion matrix. The accuracy of the model in its efficiency of utilizing the complete dataset was found to be 83.2%. The f1 score of the model is found to be 0.87 and 0.74 for the diabetic and non-diabetic condition respectively. The supporting dataset for the diabetic is found to be 500 in number and the supporting dataset for the non-diabetic was found to be 268 in number in the respective database.

Table 2: Model evaluation by diabetic and non-diabetic dataset

Attributes	Diabetic data (0)	Non-diabetic data (1)
Precision	0.85	0.79
F1 score	0.87	0.74
Supported data in number	500	268

C.MODEL EVALUATION PARAMETERS

1.CONFUSION MATRIX

Confusion matrix is a table that summarizes the performance of a classification model by showing the counts of true positive, true negative, false positive, and false negative predictions. It provides a more detailed evaluation of a classifier than just accuracy, allowing us to understand where the model is making mistakes. The confusion matrix for the model trained is given below.

From the confusion matrix the true positive, true negative, false positive, false negative is known,
True positive(TP)- 451
True negative(TN)-188
False positive(FP)-49
False negative(FN)-80

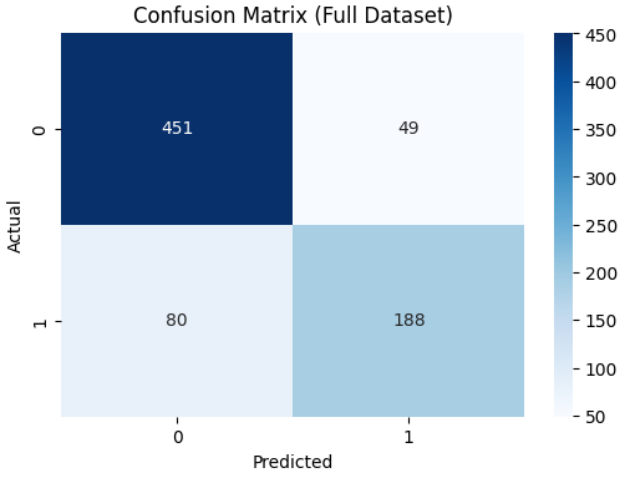

Figure 1: Confusion matrix

2. ROC CURVE

ROC curves (Receiver Operating Characteristic curves) are graphical representations used in machine learning to evaluate the performance of binary classification models. The graph plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds, providing a visual assessment of a model's ability to distinguish between classes. The ROC Curve for the model trained by using K nearest neighbour for Diabetes disease prediction is given below.

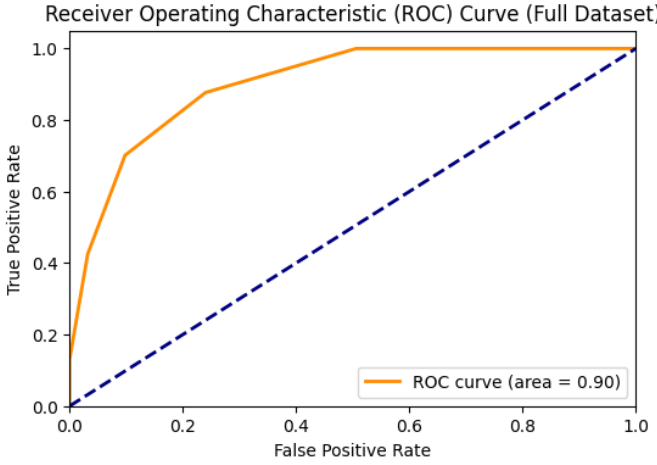

Figure 2: ROC Curve

V.CONCLUSION:

The model trained is ended by providing a better precision and accuracy for the k-Nearest Neighbours algorithms for supervised machine learning-based diabetes prediction based on a number of health characteristics in the dataset. The results of our experiments and the evaluation of the algorithm using the Confusion Matrix by 83.2% accuracy. Consequently, it can be said that when it comes to diabetes prediction using the Diabetes dataset, the KNN approach is found to predict diabetes with higher accuracy and precision. Additional algorithms, such as neural networks, and other techniques can be included for future research to improve precision and accuracy. Particle Swarm Optimization can also be used to refine the results and construct application programs.

REFERENCES

[1]. Jack Billie Chandra, Dewi Nasien. Application Of Machine Learning K-Nearest Neighbour Algorithm To Predict Diabetes. International Journal of Electrical, Energy and Power System Engineering. 2023;6:134–9.

[2]. Sarker IH, Faruque F, Alqahtani H, Kalim A. K-Nearest Neighbor Learning based Diabetes Mellitus Prediction and Analysis for eHealth Services. EAI Endorsed Transactions on Scalable Information Systems. 2020;7:1–9.

[3]. Oza A, Bokhare A. Diabetes Prediction Using Logistic Regression and K-Nearest Neighbor. Lecture Notes on Data Engineering and Communications Technologies. Springer Science and Business Media Deutschland GmbH; 2022. p. 407–18.

[4]. Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, Soguero-Ruiz C, Barquero-Perez O, Ramos-Lopez J. Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. Med Biol Eng Comput. 2020;58:991–1002.

[5]. Panwar M, Acharyya A, Shafik RA, Biswas D. K-nearest neighbor based methodology for accurate diagnosis of diabetes mellitus. Proceedings - 2016 6th International Symposium on Embedded Computing and System Design, ISED 2016. Institute of Electrical and Electronics Engineers Inc.; 2017. p. 132–6.

[6]. Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, Soguero-Ruiz C, Barquero-Perez O, Ramos-Lopez J. Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. Med Biol Eng Comput. 2020;58:991–1002.

[7]. Nirmaladevi M, Alias Balamurugan SA, Swathi U V. An amalgam KNN to predict diabetes mellitus. 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology, ICE-CCN 2013. 2013. p. 691–5.

[8]. Uddin S, Haque I, Lu H, Moni MA, Gide E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep. 2022;12.

[9]. Gupta SC, Goel N. Enhancement of Performance of K-Nearest Neighbors Classifiers for the Prediction of Diabetes Using Feature Selection Method. 2020 IEEE 5th International Conference on Computing Communication and Automation, ICCCA 2020. Institute of Electrical and Electronics Engineers Inc.; 2020. p. 681–6.

[10]. Febrian ME, Ferdinan FX, Sendani GP, Suryanigrum KM, Yunanda R. Diabetes prediction using supervised machine learning. Procedia Comput Sci. Elsevier B.V.; 2022. p. 21–30.

[11]. Habibi S, Ahmadi M, Alizadeh S. Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining. Glob J Health Sci. 2015;7:304–10.

[12]. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked. 2018;10:100–7.

[13]. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. Elsevier B.V.; 2017. p. 104–16.

[14]. Olisah CC, Smith L, Smith M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Comput Methods Programs Biomed. 2022;220.