GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 *ISSN: 2583-5955*

# A DATA DRIVEN APPROACH FOR DIAGNOSIS OF TYPE 2 DIABETES USING K-NEAREST NEIGHBOR

Soumya Ranjan Mishra, Dept of CSA, GIET University Gunupur, soumyaranjan@giet.edu
Sachikanta Dash, Dept of CSE, GIET University Gunupur, sachikanta@giet.edu

**Abstract: Long-lasting diabetes alters the manner in which our bodies use glucose and other nutrients. It is a chronic metabolic disorder, which raises many public health threats worldwide. The disease has the symptoms of increase in the sugar level in blood and may also cause a heart attack or diabetes or may lead to stroke or kidney failure or nerve damage. These symptoms worsen morbidity and even death. It is important for early detection and treatment because effective management and prevention of complications depend on such early interventions. However, the present analytic techniques used for diabetes relies heavily on centralized models that are restricted by privacy issues, data heterogeneity, and inability for ample causal relationships to be developed between samples of multi-omics data. The research will focus on the Pima Indians Diabetes Dataset (PIDD) to discover diabetes using the K-Nearest Neighbors (KNN) algorithm. The model presents an effective KNN for prediction of diabetes onset with clinical variables, demonstrated through rigorous experimentation and evaluation. Results contribute to the progress of predictive healthcare analytics and exhibit the need for using machine learning in diabetes management being proactive.**

***Keywords*— predictive analysis, diabetes detection, PIDD, KNN, healthcare system.**

## 1. Introduction:

A metabolic disorder consisting of hyperglycemia, that is a high level of sugar present in the blood as cells fail to absorb or produce a hormone called insulin is known scientifically as Diabetes Mellitus or Diabetes in common parlance. As of the International Diabetes Federation (IDF), the research says that it affects more than 400 million people around the world and is likely to cause an increase in this figure to 700 million before 2045. This epidemic represents a serious health-related problem for the world as a whole because of its devastating effects on individuals, healthcare systems, and economies.

There are many types of Diabetes and among these; Type 1 and Type 2 are the commonest, while gestational diabetes is the most common amongst them. Type 1 is unique because its autoimmune assault on pancreatic beta cells (also known as beta cell production) results in no capacity to produce insulin at all, accommodating complete insulin deficiency, Type 2 evolves usually by inheritance and environmental factors such as poor eating habits or lack of exercise into overweight conditions. Other predisposing factors during pregnancy would be the mother and child being highly predisposed to complications like gestational diabetes. The complications caused by uncontrolled diabetes are severe and debilitating-amputation of lower limbs; renal failure; blindness;

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 *ISSN: 2583-5955*

nerve damage; and cardiovascular diseases. Apart from that, diabetes has a high cost to society at large and thus affects healthcare systems because of the expenses incurred in treatment, disability loss, and lost productivity.

This condition requires a diagnosis and early management so that complications can be reduced and the health condition improved for people suffering from diabetes. Hence, there is an increasing interest in using predictive analysis to improve diabetes screening and risk stratification. This type of patients includes advanced statistical approaches and machine learning models in analyzing huge data sets to discover their possible commonalities, patterns, or associations which lead to prediction of future events or occurrences. However, predictive analysis for diabetes would take a very valuable prospect of being able to predict who is more likely to fall sick with this disease and putting in timely preventive measures or personalized care programs. To underpin this analysis, predictive analytics become widely used in diabetes detection, indicating the opportunities, challenges, and implications for healthcare practice it holds. Data-driven methodologies are integrated with the existing health infrastructures so that we can contribute to the on-going efforts of diabetes combat and make the lives of millions better worldwide.

This predictive analysis is potentially a very strong tool in that aspect called data science because they talk about what it means for the whole area of medicine, and specifically healthcare, in what they might do in terms of predictive applications: creating predictive models for future events or outcomes based on past events or outcomes. This is exactly what empowers their users with those cutting-edge and esoteric algorithms and machines in massive papers laden with statistical or works to help the users empower themselves over health trends, high-risk populations, and intervention optimization.

## 2. Review of Existing Models:

The majority of adults suffer from diabetes, which is a common disease. Several studies have been advanced in trying to forecast the signs and symptoms of diabetes. Different methodologies have been reviewed in many of these studies which include NNs, data mining, ML and genetic algorithms. Lately, there has been an increase in the use of Machine Learning due to its popularity in the construction of models and has thus received much attention from the medical world. There is evidence demonstrating that Machine Learning can better predict compared to analyzing multiple variables simultaneously. Also, Machine Learning has come up with techniques of variable selection that identify complex interactions between them. Past studies have shown that Machine Learning might be helpful in diabetes prediction. The main thing this research deals with is the determination of the categories of diabetes patients using the information

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 *ISSN: 2583-5955*

that the machine learning algorithms used. In this section are discussed some related works which use Machine Learning algorithms.

In [1], five models using varied ML algorithms have been developed by the authors. These include but are not limited to linear SVM and others herein mentioned, further categories may be found elsewhere within the text concerning among others Boruta wrapper which can choose relevant dataset features automatically. By the experimental results, it can be safely said that all the models seem to have performed well. The two models that excelled at determining whether a patient has or not are KNN and linear SVM.

In [2], the writer in their publication, writers the Journal medicine predicts the future of fighting diabetes through machine learning. A new diabetes model is proposed in the article which delineates improved diabetes classification based on some foreign determinants of diabetes diagnosis, natural factors like blood glucose concentration, BMI, age, and levels of insulin among other things. This new model has higher classification accuracy compared to existing approaches in data science. Additionally, a model for diagnosing diabetes risk has been built up.

In [3], the authors needed efficient models to predict diabetes using machine learning approaches. They trained the data sets using various machine learning algorithms such as KNN, SVM, RF, LR, DT, NB, and gradient boosting (GB). Methods of improving the accuracy rates of prediction models included employing such pre-processing procedures like label encoding, normalization etc. to process raw data before analysis which forms an essential step in any predictive analysis process. Relative to other methods SVM showed better results as stated by authors. One model that can be used employs effective pre-processing strategies like label encoding and normalization to improve the predictive capability of the models. Besides, several risk indicators were identified and then placed in a ranking order through the use of different feature selection techniques according to the authors.

Concerning in [4], a predictive modeling approach for diabetes has been suggested. Recently, they used machine-learning methods in compiling information on PIMA Indians affected by diabetes, where the R tool identifies danger elements movement and shapes. In order to point out diabetes mellitus and non-diabetic cases, five different predictive models were built and investigated with R data managing framework. During that, they resorted to supervised learning algorithms, namely the linear support vector machine (SVM-linear), radius support function kernels (RBF) and multi-dimensional reduction techniques (MDR).

In [5], diabetic patients were searched by the authors by applying ML Techniques. This strategy was first implemented in the PIMA database using bootstrapping

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 ***ISSN: 2583-5955***

resampling to improve accuracy. The authors also used KNN, NB, and DT after increasing the accuracy of their data through a resampling strategy in the PIMA database. It has been shown that using a pre-processing step in data increases the reliability of the accuracy for almost all classifiers; however, decision trees outperformed other methods based on results.

The research article [6] published a novel way of predicting type 2 diabetes. In the experiment, 952 respondents were selected and asked 18 questions concerning their physical activity, way of living, as well as the history of illnesses among family members. PIMA Indian diabetes was also incorporated into the system using the same techniques. Random Forest Classifier output remains the most reliable for both sets of results.

In [7], the hybrid type 2 diabetes diagnosis model was proposed by researchers. To ensure that they are of the same type, we need to clean the data early in our T2ML model. Then we select the features of subset by using RF & XGB classifiers. K means clustering is use for remove any data which was misclassified before using a logistic regression calculator along with cluster membership for imputing missing values.

In [8], authors have suggested an original technique for diagnosing diabetes. This paper's primary subject is the employment of machine learning strategies for diabetes detection, while India's PIMA bears a relation. Cosmetic methods include architectures like deep learning and deep features through frequent approaches: artificial neural networks or decision forest distribution with other sophisticated classifiers as random forests, decision trees and Bayesian classifiers, artificial neural networks; sub tree methods such as random forest, decision stumps or linear discriminants, one rule algorithms with k-nearest neighbours k-NN s or lazy learning (alternative solutions from instances) accessorizing too many control parameters (hidden neurons), single classifiers based on logistic regression models for instance-based learning (e.g. k-nearest neighbour). The research also highlights certain advantages and disadvantages of the findings.

The authors [9] presented a well-defined approach to improve early diagnosis of diabetes and correctly differentiate between its various forms. To achieve this, the researchers experimented with numerous ML techniques including, K-nearest neighbours (KNN), AdaBoost (AB), Gaussian Naïve Bayes (GNB), as well as Gaussian processes (GPC). They measured the reliability of these methods by looking at their respective precision rates and F1 scores as well as recall rates and errors.

The method for diabetes forecast that was proposed by the authors should be made known [10]. Considering the fact that it was specifics of the population that were involved, the four most popular machine learning algorithms used in predicting. The

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 *ISSN: 2583-5955*

inference from the author's experimental work is that of all these ML Methodologies, the decision tree outperforms them all by accurately forecasting same.

In their study, [11] the authors chose four classifiers: RF, NB, AB, and DT, to predict if someone has diabetes or not. Using Logistic Regression (LR), they calculated the risk factors of diabetes considering both the P-value and odds ratio. There were three different methods of partitioning known as K2, K5 and K10 in this case. Classification accuracy ACC along with AUC was measures used in the evaluation process of these classifiers. The researchers discovered that the most important risk factors for diabetes include: age, blood pressure level and diastole blood pressure level, total cholesterol concentration or body mass index. Besides, the combination of LR and RF classifiers improved significantly to aid in predicting diabetes among different individuals.

In [12] the authors have suggested new means for forecasting diabetes with plans to increase its accuracy through use of bootstrapping, resampling technique following by Naive Bayes, Decision Trees, and KNN thereby comparing their efficiency.

In [13] a prior publication, the BRFSS dataset was employed to teach LR and RF models. User information was collected with a chatbot before predicting how common long-term illnesses were; then by visualizing these data sets interactively risk reduction advice could be given. They checked various aspects but found out that RF worked well when looking at diabetes only if nothing else.

The authors proposed a new tool for predicting diabetes in [14]. We further propose the creation of a diabetes diagnosing technique through the application of these characteristics under both five and tenfold cross-validation modes of a neural network. PIMA Learning Machine TOOLS repository servers are; where the PIMA Indian Diabetes (PID) dataset was obtained.

Authors from reference [15] used many popular regression models such as Glmnet, LightGBM, XGB, RF etc. to predict type 2 diabetes mellitus exactly. On starting with one hundred and eleven variables as input into setting up the sample for analysis; only fifty-eight out of these initially introduced had been retained after pre-processing the data. This work aims to investigate the precision in terms of forecasting along with model calibration analysis while conducting comparisons between multi variable regression techniques as well as machine learning based predictive algorithms in terms of their predictive performance. According to research results, one way to improve the prediction models is by updating them with more recent data however what the importance of this variable for any given machine learning method will depend upon varies from one algorithm to another.

The authors suggest a new technique for forecasting type II diabetes in reference to the recommendation made in [16]. The study would include several ML models and

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 ***ISSN: 2583-5955***

Linear Regression (LR). We conducted trials whose purpose was to determine women at PIMA Indians suffering from diabetes. In this case, the recent research concerning PIDD distinguishes an appropriate ML model employing cross-validation strategies.

The authors used DT, RF, KNN, AB, XGB and NB in this work [17]. To achieve a reliable prediction, pre-processing is a must with the PID dataset ensemble model application. According to the outcome, the optimal solution for diabetes prediction is two types of boosting classifiers combined XGB and AB. A better degree of accuracy in predicting diabetes can be achieved by utilizing an optimal combination of AB and XGB dataset when recommended pre-processing techniques are utilized.

A new technique for predicting diabetes in the early stages was proposed by researchers [18]. The prediction is made of diabetes in this paper using key attributes, as well as determining various features' interactions. The data implies that there's a strong connection between these three parameters: body mass index (BMI), glucose amount, and diabetes occurrence.

According to [19] the primary method applied was LR although with ensemble techniques combining LR and other ML algorithms such as DT, NB, SVM and KNN among others thus improving performance. The experiments mainly revolved around two datasets and two different methods for selecting the best features were employed. Initially, they chose the Pima Indians dataset which had nine different attributes for this experiment. The second dataset that was used is the Vanderbilt dataset which has 16 features. Through its conclusions, this research showed that the Logistic Regression (LR) algorithm is among the best algorithms that can be employed in building forecasting models.

In [20], the authors have proposed a novel method for predicting diabetes progression. They can determine beforehand who is at an increased risk of developing pre-diabetes. When someone has diabetes, he could have a personalized treatment plan without taking care of low-risk individuals or preventive interventions. They examined the potential for electronic medical records (EMR), along with a machine learning model, to improve the prediction of incident diabetes rates based on patient information.

The PIDD dataset was employed for this study by [21] which consisted of different characteristics and seven different ML models were constructed. From feature selection, two attributes had been omitted. It was shown that SVM and LR gave rise to the best model for diabetes prediction. In applying various epochs for training NN on the same data set containing multiple layers of hidden neurons than before; however, many more researchers found them converging at different outputs when they were trained using other methods inevitability led to similar results while differences arise due only

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 *ISSN: 2583-5955*

because training sets were chosen differently at random. The authors of this study noted that a NN with two hidden layers outperforms all the other algorithms.

In [22], an alternative means to predicting type 2 diabetes was proposed through an online and offline survey 952 patients were included it involved 18 items about exercise levels, eating habits and tobacco use among others. Aside from PIMA Indian diabetes, these very measures are contained in that database about an individual's health. It comes out that the Random Forest Classifier makes the best predictions for both of the datasets.

In [23], the authors do an extra analysis to check out how effective ML is when it comes to diagnosing diabetes. The outcomes point to the favor of existing strong enough ML algorithms for doctors to identify those likely to get this disease.

Article [24] utilized machine learning for conducting a meta-analysis on diabetes prognosis methodology. The potentiality of biasedness that might have taken place on these machine learning models was examined using the novel Prediction Model Risk of Bias Assessment Tool PROBAST. Meta disc software package was used to perform meta-analysis and check for heterogeneities. Compared to other tested approaches, machine learning models emerged better in predicting diabetes.

## 3. Methodology:

This dataset was chosen because it contains clear and accurate data, which makes it easier to analyses and act on it accordingly. The main purpose of this project is to analyses and predicts whether the person has Type-2 diabetes (diabetes melilotus) or not. This part of the project is built up to predicting the diabetes and the following model is to be converted to an Application Programming Interface (API) and can be integrated any of the applications and be used as a software to show the status of diabetes i.e. positive or negative. We collected the dataset from Kaggle and after collecting the dataset we went through the dataset by doing exploratory data analysis. The dataset is comprised of 768 entries where 500 negative (denoted by 0) and 268 positive (denoted by 1) cases. The dataset comprises 8 classes or columns which performs as a reference value point and serve to determine the result of diabetes result. The final outcome of the testing data is based on the values of these 8 columns.

## 4. Outcome:

After verifying all of the above, we entered the stage of data pre-processing and cleaning. The steps of data pre-processing include removing undesirable data points and missing values. We visualized the data categories with respect to each other in order to get a better view of data. In this process we got to find out there are some missing values and went to fill those missing values by the Specific columns with their respective mean

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 ***ISSN: 2583-5955***

or median values. We can retain dataset quality and potentially upgrade machine learning efficiency if missing values are substituted by average or median.

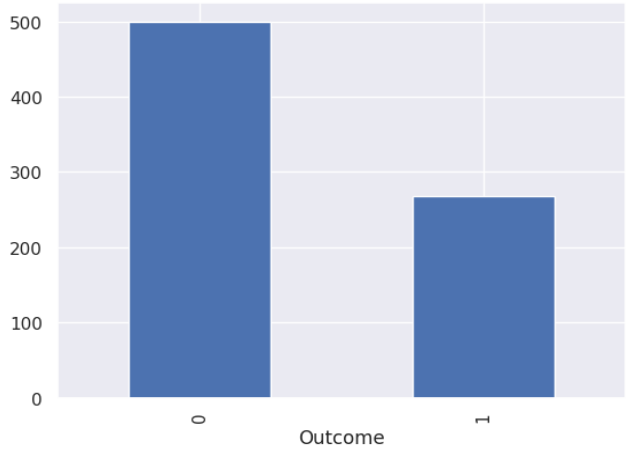


Figure 1: Bar chart for balance of data count

Then we went on to plot a matrix of scatter plots and histograms for dataset related to diabetes.

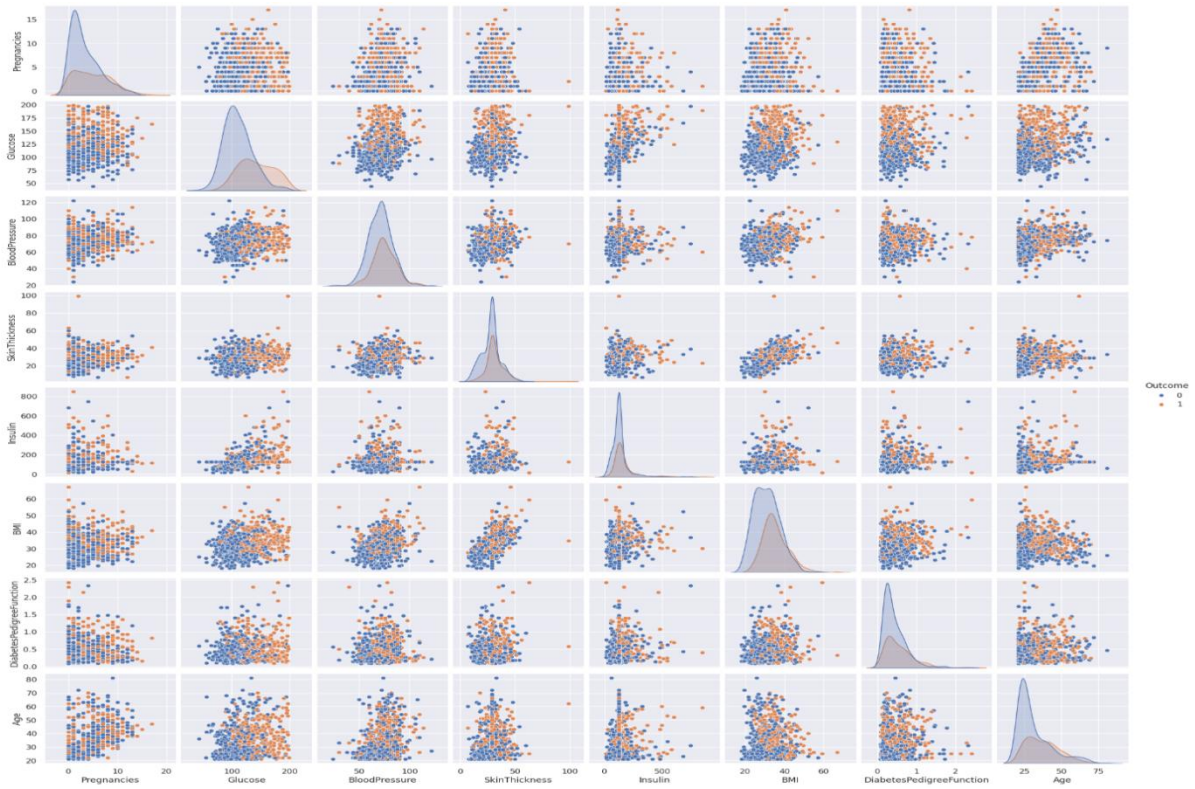


Figure 2: Scatterplot matrix for visualize the relationships

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 *ISSN: 2583-5955*

The diagonal plots display the distribution of individual features through of the Histograms. The points are color-coded based on the target variable, which appears to be a binary outcome (0 or 1, potentially representing the presence or absence of a condition). On the whole, this graphic provides a complete view of the dataset that helps us in getting to know how the individual attributes are distributed, recognizing possible correlations and patterns, and learning more about the connections among features as well as targeting variables.

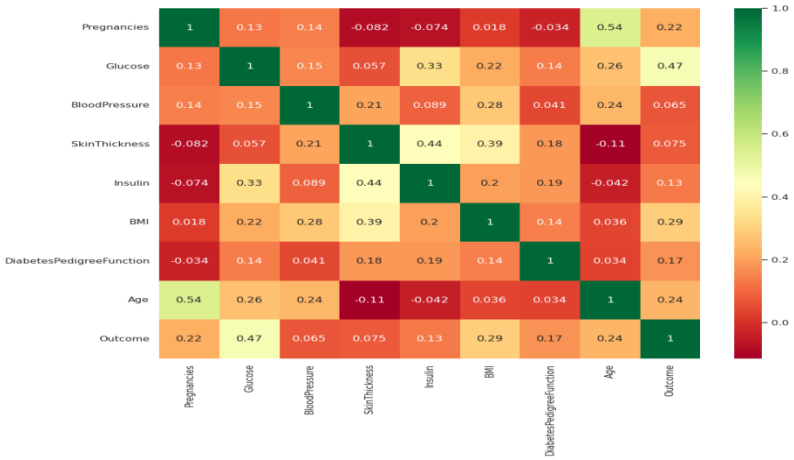


Figure 3: Correlation matrix before cleaning

The image displays a correlation matrix, which shows the pairwise correlations between different variables related to diabetes and health factors. The matrix displays colors green for positive correlations, red for negative correlations, and dark shades for stronger correlations, whether positive or negative. To before working onto the process of the model building, we categorized the dataset into 32 sub-categories for training purposes with a proportion of 8:2 ratio for testing. We have used KNN algorithm in building the model to predict the accurate output with accuracy.

In the development of machine learning solutions especially within sensitive domains, such as healthcare, analytical empirical evaluation and domain knowledge are vital. The identification of the maximum training accuracy and its associated k value(s) sheds light on the efficacy of the KNN algorithm in capturing patterns within the training data.

**4.1 Maximum Testing Accuracy:** The maximum testing accuracy attained during hyperparameter tuning is [92.56] %.

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 ***ISSN: 2583-5955***

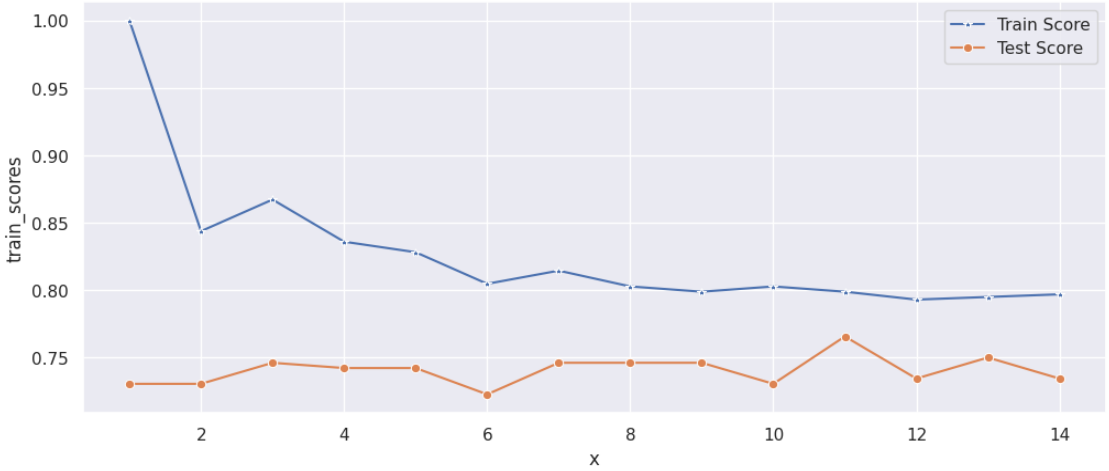


Figure 4: Learning Curve

The table is color-coded. Darker shades indicate stronger correlations. The stronger the correlation, the stronger the colour. The intensity of the colour is used to show how strong the relationship is. Lighter shades indicate weaker correlations. The correlation strength is represented by lighter shades.

**4.2 Model Initialization:** The KNN model was initialized with the optimal *k*k value of 11, which was identified through rigorous hyperparameter tuning. This value represents the number of nearest neighbours considered during classification, striking a balance between model complexity and generalization performance.

**4.3 Model Training and Evaluation:** Pairwise correlations A variety of physiological features that are important in detecting diabetes were included when the KNN model was trained using the training dataset. The model's accuracy to predict if a person has diabetes or not was verified using the testing dataset afterwards.

**4.4 Testing Accuracy:** The testing accuracy achieved by the KNN model with the optimal is [92.56] %. This metric measures how well the model can correctly separate people into two categories: diabetics and non-diabetics, according to their physiological characteristics. The function plot decision regions produced a map of the decisions taken by the KNN model. Through this function, classification models' decision boundaries can become visible, and it gives a representation of how the model splits itself into distinct areas given inputs in feature space. To enhance the visualization, filler feature values and ranges were specified for features 2 to 7 corresponding to BMI, Insulin level, Glucose level, Blood pressure, Age and Skin thickness. By setting uniform values and ranges for these features, we ensure consistency in the visualization and focus solely on the decision boundary.

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 *ISSN: 2583-5955*

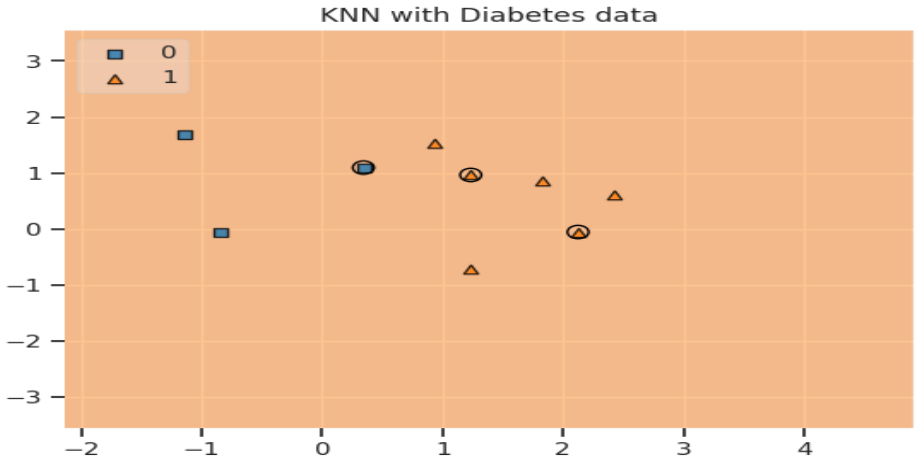


Figure 5: Scatter plot visual representation

The certainty chart is an impactful tool that aids in testing the efficiency of classifier models as it summarizes the right and wrong predictions made. When we chart the certainty chart, we get better understanding of how well the model can properly label its diabetic and non-diabetic instances hence enabling a full evaluation of model functioning. The confusion matrix was created when the actual labels (y_test) were compared against those predicted (y_pred) by the KNN model. The confusion matrix function available in metrics mod was used to calculate the confusion matrix that captures model's four types of predictions.

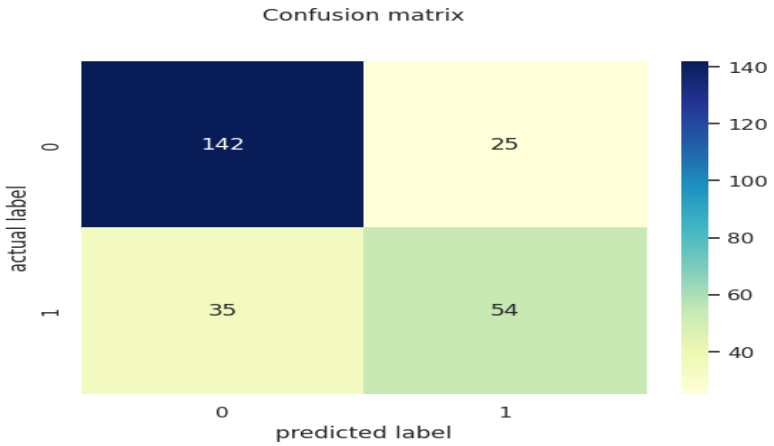


Figure 6: Confusion Matrix

This figure above indicates that the table offers a summary regarding the performance of a type classification model from specific set of test data sets among other types of matrices. What's more, it depicts expected labels across one side (x-axis) against real labels along its other edge (y-axis). In this confusion matrix, there are two classes: 0 and 1. The actual labels are shown in the rows and the predicted ones in the columns.

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 ***ISSN: 2583-5955***

The Receiver Operating Characteristic (ROC) curve was used to assess KNN model for diabetes detection. When the ROC curve equals one minus specificity, we are below all thresholds.

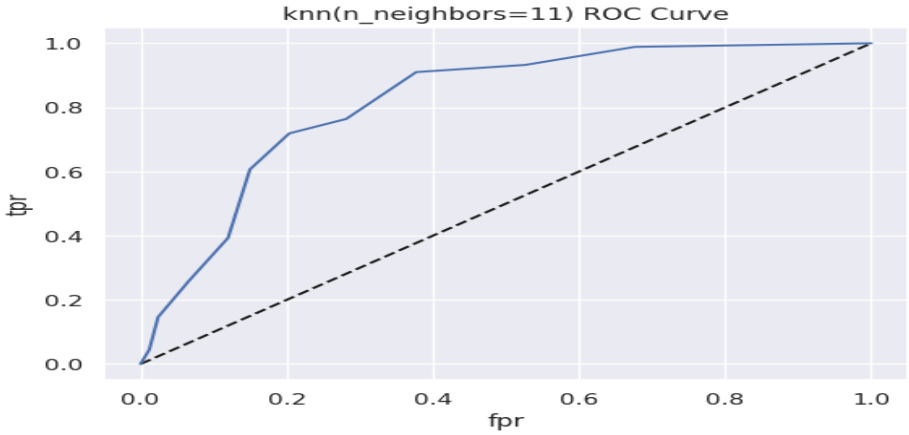


Figure 7: ROC Curve

The model's problem types are shown by the confusion matrix and could even be used to show where to enhance or even where the data's class imbalances are probable. The ROC Curve to store K Nearest Neighbours classification model is displayed in the image with the number of neighbours set to 11.

The ROC Curve provides a graphical outline of how the differences in misclassified items as well as correctly recognized items vary across various thresholds put in place for classification. The x-axis here is used to represent the false positive rate (FPR) whereas the y-axis reveals the true positive rate (TPR).

Blue curve represents the performance of KNN model with 11 neighbours on the given dataset. The dotted diagonal line represents the ROC curve of a random classifier, which has no predictive power. For a model to be said as genuinely good in classification an ROC curve that clings near the upper left corner of the graph would mean high true positives and low false positives out of all possible negatives. In this case, KNN model with 11 neighbours performs reasonably well, as its ROC curve is above the diagonal line, indicating better performance than a random classifier.

## 5. Result Analysis:

The best cross-validated score obtained during grid search was 92.56, achieved with the following hyperparameters. The hyperparameter tuning process using Grid Search Cross-Validation has enabled the identification of the optimal configuration for the KNN model in diabetes detection. By systematically exploring a range of hyperparameters, we ensure the robustness and reliability of the model, enhancing its predictive performance and real-world applicability. To optimize the performance of the

GIET UNIVERSITY JOURNAL, 2025, VOL.4, NO.1, PP. – 29-43

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 *ISSN: 2583-5955*

K-Nearest Neighbours model for diabetes detection and enhance its real-world applicability. Our journey in developing a robust predictive model for diabetes detection has been marked by meticulous experimentation and rigorous evaluation. Through the utilization of machine learning techniques, we aimed to leverage the power of data-driven approaches to improve healthcare outcomes and enhance patient care.

## 6. Discussion:

This study emphasizes the significance of predictive analytics in healthcare, particularly in the quit onset and control of diabetes. By leveraging machine learning algorithms such as KNN, healthcare providers can develop proactive interventions and personalized treatment plans for individuals at risk of diabetes. In clinical practice, there is a need for more research on ways to improve the practicality of forecast models which would address factors like data quality, scalability, and interpretability. Additionally, interdisciplinary collaboration between healthcare professionals, data scientists, and policymakers is essential for translating research findings into actionable insights and improving patient outcomes.

## 7. Conclusion

To sum up, this study offers a thorough examination of predictive analysis for diabetes detection utilizing the KNN algorithm and the Pima Indian Diabetes Dataset. This study advances predictive healthcare analytics by proving that KNN is effective in correctly forecasting the onset of diabetes based on clinical characteristics. By leveraging machine learning algorithms and large-scale healthcare datasets, we can enhance early detection, intervention and management of diabetes can be improved hence making the patient outcomes better while reducing the burden of the chronic disease.

## 8. References

1. H. Kaur and V. Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach,” Applied Computing and Informatics, vol. 18, no. 1/2, pp. 90–100, 2020.
2. A. Mujumdar and V. Vaidehi, “Diabetes prediction using machine learning algorithms,” Procedia Computer Science, vol. 165, pp. 292–299, 2019.
3. N. Ahmed, R. Ahammed, M. M. Islam et al., “Machine learning based diabetes prediction and development of smart web application,” International Journal of Cognitive Computing in Engineering, vol. 2, pp. 229–241, 2021.

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 ***ISSN: 2583-5955***

4. H. Kaur and V. Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach,” Applied Computing and Informatics, vol. 18, no. 1/2, pp. 90–100, 2022.
5. A. Aada and S. Tiwari, “Predicting diabetes in medical datasets using machine learning techniques,” International Journal of Scientific Research and Engineering Trends, vol. 5, no. 2, pp. 257–267, 2019.
6. N. P. Tigga and S. Garg, “Prediction of type 2 diabetes using machine learning classification methods,” Procedia Computer Science, vol. 167, no. 2019, pp. 706–716, 2020.
7. S. Albahli, “Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection,” Journal of Medical Imaging and Health Informatics, vol. 10, no. 5, pp. 1069–1075, 2020.
8. A. Choudhury and D. Gupta, A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques, Vol. 740, Springer, Singapore, 2019.
9. P. Palimkar, R. N Shaw, and A. Ghosh, “Machine Learning Technique to Prognosis Diabetes Disease: Random forest Classifier Approach,” Advanced Computing and Intelligent Technologies, Springer, Singapore pp. 219–224, 2022.
10. M. F. Faruque, Asaduzzaman, and I. H. Sarker, “Performance analysis of machine learning techniques to predict diabetes mellitus,” in Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) 2019, pp. 1–4, 2019.
11. M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, “Classification and prediction of diabetes disease using machine learning paradigm,” Health Information Science and Systems, vol. 8, pp. 7–14, 2020.
12. Y. Jeevan Nagendra Kumar, N. Kameswari Shalini, P. K. Abhilash, K. Sandeep, and D. Indira, “Prediction of diabetes using machine learning,” International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 7, pp. 2547–2551, 2019.
13. G. Bhola, A. Garg, and M. Kumari, “Comparative study of machine learning techniques for chronic disease prognosis,” Computer Networks and Inventive Communication Technologies, vol. 58, pp. 131–144, 2021.
14. S. Islam Ayon, M. Milon Islam, and M. Milon Islam, “Diabetes prediction: a deep learning approach,” International Journal of Information Engineering and Electronic Business, vol. 11, no. 2, pp. 21–27, 2019.
15. L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, “Early detection of type 2 diabetes mellitus using machine learningbased prediction models,” Scientific Reports, vol. 10, no. 1, Article ID 11981, 2020.
16. G. Battineni, G. G. Sagaro, C. Nalini, F. Amenta, and S. K. Tayebati, “Comparative machine-learning approach: a follow-up study on type 2 diabetes predictions by crossvalidation methods,” Machines, vol. 7, no. 4, pp. 74–11, 2019.

Manuscript received: 5th January 2025, revised: 6th January 2025

Accepted: 7th January 2025 ***ISSN: 2583-5955***

17. M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, “Diabetes prediction using assembling of different machine learning classifiers,” IEEE Access, vol. 8, pp. 76516–76531, 2020.
18. T. Mahboob Alam, M. A. Iqbal, Y. Ali et al., “A model for early prediction of diabetes,” Informatics in Medicine Unlocked, vol. 16, no. January, Article ID 100204, 2019.
19. P. Rajendra and S. Latif, “Prediction of diabetes using logistic regression and ensemble techniques,” Computer Methods and Programs in Biomedicine Update, vol. 1, Article ID 100032, 2021.
20. A. Cahn, A. Shoshan, T. Sagiv et al., “Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model,” Diabetes, vol. 36, no. 2, pp. 1–8, 2020.
21. J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” ICT Express, vol. 7, no. 4, pp. 432–439, 2021.
22. N. P. Tigga and S. Garg, “Prediction of type 2 diabetes using machine learning classification methods,” Procedia Computer Science, vol. 167, no. 2019, pp. 706–716, 2020.
23. S. Kodama, K. Fujihara, C. Horikawa et al., “Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: a meta-analysis,” Journal of Diabetes Investigation, vol. 13, no. 5, pp. 900–908, 2022.
24. Z. Q. Zhang, L. Q. Yang, W. T. Han et al., “Machine learning prediction models for gestational diabetes mellitus: meta-analysis,” Journal of Medical Internet Research, vol. 24, no. 3, Article ID e26634, 2022.