FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

BIT 33603 Data Mining

Assignment: PREDICTING DIABETES WITH CLASSIFICATION ALGORITHM

PREPARED FOR:
DR. ROZITA BINTI ABDUL JALIL

PREPARED BY:
NURULAINA NISA BINTI SAHANUAN (AI220179)
NURUL JANNAH BINTI KAMARUL ZAMAN (AI220147)
SERENA NG YEN XIN (AI220061)
TUAN KHALIDAH SYAZWANA BINTI TUAN MOHD KASMAWI (AI220118)

16TH MAY 2025

CHAPTER 1

PREDICTING DIABETES WITH CLASSIFICATION ALGORITHM

Nurul Jannah Kamarul Zaman[1], Nurulaina Nisa Sahanuan[2], Serena Ng Yen Xin[3], Tuan Khalidah Syazwana Tuan Mohd Kasmawi[4]
*Rozita Abdul Jalil[4]**

[1-5] Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia,
Parit Raja 86400, Batu Pahat, Johor, Malaysia
*Corresponding Email: rozita@uthm.edu.my

Abstract.
Diabetes mellitus is a serious health issue that can have long-term consequences if not recognized early. This project focuses on predicting the probability of an individual getting diabetes based on their medical history and demographics using classification algorithms. Based on the attributes, patients are classified into Diabetic, Non-Diabetic, or Pre-Diabetic. The study's goal is to create a prediction model to assist in early diagnosis by utilizing data mining techniques via the Knowledge Discovery in Databases (KDD) process. By using the R language, models such as Logistic Regression, Random Forest, and K-Nearest Neighbors are analyzed. Performance criteria such as accuracy, precision, recall, and F-score are utilized to evaluate the models. The findings show that data mining is successful in medical diagnosis and supports the integration of intelligent systems in healthcare. This study advances proactive diabetes management and lays the groundwork for future improvements in predictive healthcare analytics.

1.0 INTRODUCTION

Diabetes mellitus, namely Type 2 diabetes, is a widespread metabolic disorder affecting millions globally. Most of the time, it is due to the body's resistance to insulin or an inadequate synthesis of insulin, resulting in increased blood glucose. Diabetes, if not properly treated, can cause serious health consequences such as cardiovascular disease, neuropathy, renal failure, and even blindness. The World Health Organization (WHO) states that the worldwide diabetes burden has increased dramatically, with roughly 14% of individuals aged 18 and older living with diabetes by 2022, more than double the rate in 1990. Furthermore, in lower-to-middle-income countries, exceeding half of diabetics go untreated, revealing a serious gap in healthcare access and early intervention [12].

Given the rising number of cases and underdiagnosis of diabetes, computational technologies such as machine learning (ML) have gained traction for their capacity to aid in early identification and risk assessment. The availability of electronic health records and clinical data has aided in the creation of prediction algorithms capable of detecting high-risk patients before symptoms appear clinically. These models can help healthcare practitioners prioritize treatment and manage diabetes more effectively.

The dataset utilized in this study, obtained from Mendeley, consists of anonymised health records covering Number of Patient, Blood Sugar Level, Age, Gender, Creatinine ratio (Cr), Body Mass Index (BMI), Urea Level, Cholesterol (Chol), LDL, VLDL, Triglycerides(TG) and HDL Cholesterol, HbA1c, Class (indicating whether the patient is Diabetic, Non-Diabetic, or Predict-Diabetic). These characteristics are largely accepted as predictive indications for diagnosing Type 2 diabetes. The major purpose of this dataset is to create a classification model capable of reliably predicting an individual's chance of developing diabetes, allowing for prompt intervention.

The main objectives of this study are as follows:

1) To preprocess and analyze the dataset to extract and select the most relevant features for diabetes prediction.
2) To implement multiple classification algorithms using the R programming language and evaluate their performance.
3) To compare the prediction accuracy of each algorithm and identify the most effective model based on standard evaluation metrics.

Several classification techniques are used in this study, including Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN). These strategies were chosen for the ideal combination of accuracy, interpretability, and demonstrated performance in healthcare prediction tasks. Logistic Regression is commonly employed for binary classification issues such as diabetes prediction, but ensemble approaches like Random Forest have proven to be particularly effective in dealing with noisy datasets.

Previously conducted studies using datasets such as the Pima Indian Diabetes Dataset revealed the efficacy of machine learning in predicting diabetes. Studies have used classification and ensemble approaches to get important insights and identify the most reliable prediction methods [13]. However, choosing the best-performing algorithm remains difficult due to the variability in data properties and performance between scenarios. As a result, this study adds value by applying and evaluating several models to a recent dataset with a variety of properties.

In conclusion, this effort seeks to determine which health markers have the most effect on the onset of diabetes, in addition to achieving good prediction performance. Understanding these contributing characteristics can help to shape public health policy, encourage early interventions, and facilitate tailored treatment planning. With the worldwide diabetes burden constantly expanding, technologies for early and accurate detection are becoming increasingly important.

The remainder of this paper is organized as follows. Section 2 reviews all works related to diabetes prediction using data mining techniques. Section 3 presents the Knowledge Discovery in Database (KDD) methodology utilised in performing the data mining

task, along with the dataset and the evaluation metrics. Section 4 presents the outcomes. Section 5 discusses the research findings, and finally, Section 6 concludes with some directions for future work.

2.0 RELATED WORK

In recent years, several academics have used machine learning and data mining approaches to address the rising issue of diabetes prediction and complication analysis. The following is a detailed evaluation of 10 chosen research that investigated various elements of diabetes prediction utilizing classification algorithms and similar health datasets.

Ibrahim and Khairi (2022) looked at predictive data mining methods for identifying type 2 diabetes. They used the Pima Indian Diabetes dataset to evaluate the performance of Decision Tree, Random Forest, and K-Nearest Neighbors (KNN) models. Their goal was to determine which strategy resulted in the most accurate forecasts. Random Forest outperformed the other models tested, proving its superiority in medical classification challenges [14].

Islam et al. (2020) applied several data mining algorithms to forecast diabetes in the early stages. The researchers analyzed patient data using Decision Tree, Naïve Bayes, and Artificial Neural Networks (ANN), including glucose, insulin, BMI, and physical activity. The ANN model had the best accuracy, at more than 88%, showing its usefulness in detecting early diabetes risk [15].

Khairudin et al. (2020) applied data mining techniques to predict diabetic retinopathy in Type II diabetes patients. They used several classification approaches, including J48 Decision Tree, SVM, and Naïve Bayes. Based on patient parameters such as HbA1c, blood pressure, and age, the J48 Decision Tree offered the best results, with a prediction accuracy of 93%. Their research indicates that machine learning technologies might help anticipate diabetes complications early on [16].

Khan et al. (2021) provided a comprehensive overview of machine learning algorithms for diabetes diagnosis and prediction. They

analyzed over 50 research publications and compared models such as Logistic Regression, Random Forest, and Gradient Boosting Machines. Their findings demonstrated that ensemble approaches performed better and had more reliability than single classifiers. This study emphasizes the promise of sophisticated algorithms for medical prediction tasks [17].

Oon et al. (2024) investigated diabetic retinopathy prediction among Type II diabetes patients with neuropathy. Using a local clinical dataset, they used classification approaches such as Decision Tree, Random Forest, and Naïve Bayes. The Random Forest model had the best accuracy (92%). Their findings demonstrate the increased usefulness of applying sophisticated classifiers to predict complications in individuals with various diabetes-related illnesses [18].

Rastogi and Bansal (2023) created a diabetes prediction model employing several data mining approaches such as Logistic Regression, Support Vector Machine (SVM), and KNN. They tested their models on a cleaned version of the Pima Indian Diabetes dataset. SVM outperformed the other approaches in terms of precision and recall. Their findings underscored the necessity of algorithm selection and sufficient data pretreatment in making accurate predictions [19].

Table 1: Summary Table Of The Related Work

Reference	Purpose of Study	Machine Learning Methods Used	Obtained Results
Ibrahim & Khairi (2022) [14]	To compare predictive models for Type II diabetes diagnosis.	Decision Tree, Random Forest, KNN	Random Forest – Highest accuracy (not explicitly stated)
Islam et al. (2020) [15]	To predict early diabetes likelihood using data mining.	Decision Tree, Naïve Bayes, Artificial Neural Network	ANN – >88% accuracy

Khairudin et al. (2020) [16]	To predict diabetic retinopathy among Type II diabetes patients.	J48 Decision Tree, Naïve Bayes, SVM	J48 Decision Tree – 93%
Khan et al. (2021) [17]	Literature review of machine learning techniques for diabetes prediction.	Comparative analysis of >50 studies (including LR, RF, GBM)	Ensemble methods (Random Forest, GBM) most effective
Oon et al. (2024) [18]	To predict diabetic retinopathy in T2DM patients with neuropathy.	Decision Tree, Random Forest, Naïve Bayes	Random Forest – 92% accuracy
Rastogi & Bansal (2023) [19]	To build a reliable model for early diabetes detection.	Logistic Regression, SVM, KNN	SVM – Highest precision & recall

3.0 METHODOLOGY

This data mining task involves predicting diabetes with classification algorithms and is performed following the Knowledge Discovery in Database (KDD) methodology to extract useful information from the diabetes dataset. The flow of the KDD process is as shown in Fig. 1.

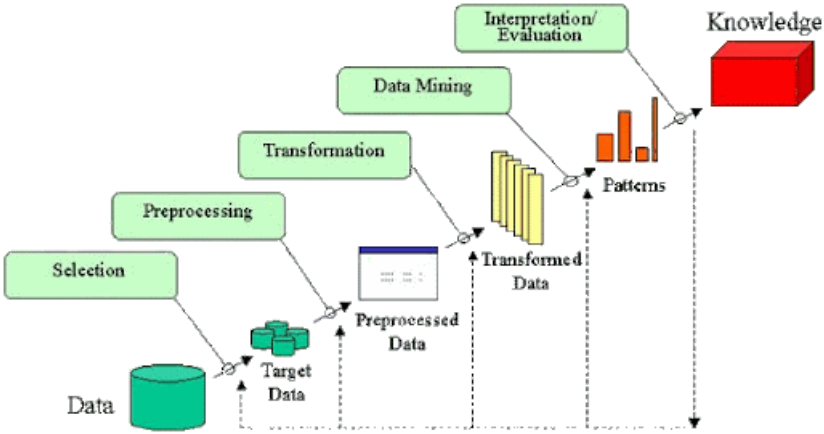

Fig. 1 Knowledge Discovery in Database (KDD) Process

The key stages in the Knowledge Discovery in Database (KDD) process involve data selection, preprocessing, transformation, mining, and interpretation/evaluation, as shown in Fig. 1.

In the data selection phase, relevant data is identified and extracted for analysis. The selected attributes include gender, age, urea, HbA1c, cholesterol, body mass index, creatinine ratio, and fasting lipid profiles (LDL, HDL, VLDL, TG). Irrelevant attributes such as ID and patient number are removed. Data selection ensures that the insights obtained through the KDD process are relevant and true to the intention.

In the data preprocessing phase, data cleaning is performed. Missing values, outliers, and inconsistencies are handled. In the diabetes dataset, no explicit missing value is observed. Categorical variables such as gender are converted to binary, with 1 for Male and 0 for Female, and classes are converted to 0 for non-diabetic, 1 for pre-diabetic, and diabetic. Data preprocessing is crucial in enhancing the effectiveness of data mining.

In the data transformation phase, dimensionality reduction is performed to improve model efficiency. Creatinine ratio and fasting lipid profiles (LDL, HDL, VLDL, TG) are removed from the list of attributes due to their correlation with each other. The dataset

attributes are narrowed to gender, age, urea, HbA1c, cholesterol, and BMI.

In the data mining phase, classification algorithms such as Logistic Regression, Random Forest, and K-Nearest Neighbour (KNN) are used. Meaningful information and patterns are discovered through the process of data mining.

In the data interpretation/evaluation phase, the patterns identified during data mining are assessed to evaluate their relevance. Visualisation is applied to the results to improve the readability of the data, which helps in pattern finding. Through data interpretation/evaluation, we can predict the diabetes risk of future patients.

The experiments were carried out using R Studio with a 5-fold validation method for training and testing. Classification in R follows a structured process to analyse data and make predictions. First, the required packages are loaded to provide the necessary functions. The diabetes dataset is imported and cleaned. A fixed random seed is set to make the results reproducible. The data is then split into two segments, one for training the model and another for testing its accuracy. The relationships between the factors (gender, HbA1c, BMI, etc.) and the outcome (class) are defined. Using this relationship, logistic regression, random forest, and k-nearest neighbours models are trained on the data. Finally, the models are tested to see how well they predict outcomes. Performance is measured using evaluation metrics like accuracy, precision, recall, and F-score.

Classification is performed on this dataset to predict diabetes risk based on clinical and demographic features. Classification of diabetes is crucial in the early detection, prevention, and treatment of diabetes. This helps healthcare providers to make informed decisions for patients.

3.1 Dataset

This dataset contains medical and laboratory data focusing on diabetes. Table 1 provides the specific details of the dataset, including its source, publishing year, owner, sample size, number of attributes, usage, missing values, and data types.

Table 1. Specific Description of the Dataset

Description	
Source	Mendeley Data
Publishing Year	18 July 2020
Owner	Ahlam Rashid University of Information Technology
Sample Size	1000
Number of Attributes	14
Usage	Classification for diabetes prediction Applied sciences Health sciences
Missing Values	No explicit missing values
Data Types	Numerical: Age, Urea, BMI Categorical: Gender, Class

Table 2 shows the initial diabetes dataset sample retrieved from Mendeley Data, and Table 3 shows the transformed diabetes dataset sample for data mining purposes.

Table 2. Initial Diabetes Dataset Sample

ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
670	34229	M	43	2.6	67	4	3.8	0.9	2.4	3.7	1	21	N
630	34275	F	48	2.8	55	5	3.6	0.6	2.1	1.2	0.2	23.5	N
705	87668	M	33	4.8	64	5.8	4.8	1.1	1.7	2.6	0.5	25	P
245	24053	M	61	5.7	92	6.2	2.6	1.1	0.9	1.6	0.7	38	Y
299	24071	F	56	5.5	48	9.3	5.8	2.4	0.9	3.9	1.1	36.6	Y

Table 3. Transformed Diabetes Dataset Sample

Gender	AGE	Urea	HbA1c	Chol	BMI	CLASS
1	43	2.6	4	3.8	21	0
0	48	2.8	5	3.6	23.5	0
1	33	4.8	5.8	4.8	25	1
1	61	5.7	6.2	2.6	38	1
0	56	5.5	9.3	5.8	36.6	1

Table 4 shows the features in the diabetes dataset. The features include gender, age, urea, HbA1c, cholesterol, and body mass index.

Table 4. Transformed Diabetes Dataset Sample

Feature	Data Type	Description
Gender	Categorical	Patient's sex (M/F).
Age	Numerical (Discrete)	Patient's age in years.
Urea	Numerical (Continuous)	Blood urea level (mmol/L).
HbA1c	Numerical (Continuous)	Haemoglobin A1c (%), measures the average level of blood sugar over 3 months.
Chol	Numerical (Continuous)	Total cholesterol (mmol/L).
BMI	Numerical (Continuous)	Body mass index (kg/m^2), measures body fat.

Table 5 shows the class labels for the diabetes dataset. The classes are divided into 0 for non-diabetic (N), 1 for pre-diabetic (P) and diabetic (Y).

Table 5. Transformed Diabetes Dataset Sample

Class	Label
Non-diabetic	N (0)
Pre-diabetic	P (1)
Diabetic	Y (1)

3.2 Algorithm

- Logistic Regression

  Logistic regression is a statistical analysis technique used to describe the relationship between one or more independent predictor factors and a binary outcome, or a dependent variable with two categories, yes or no. [2].

  The formula of logistic regression is shown in Eq. 1.

$$p(x) = \frac{e^{a+bx}}{1+e^{a+bx}} \quad (1)$$

  where p(x) = predicted output, a = intercept term, and b = coefficient of single input value (x).

  The implementation of logistic regression in this experiment using R is shown in Fig. 2.

```
#Logistic Regression
log_model<- glm(CLASS ~ ., data = train_data, family = binomial)
log_pred_prob <- predict(log_model, test_data, type = "response")
log_pred <- ifelse(log_pred_prob > 0.5,1,0)

#evaluate all models
log_metrics <- evaluate_model(log_pred, test_y)
```

Fig. 2 Implementation of Logistic Regression in R

- Random Forest

  Random forest is an ensemble learning technique that uses a defined ensemble strategy to merge many decision trees into a single model [3].

  In classification, the formula for the final prediction is based on majority voting among all decision trees, as shown in Eq. 2.

$$\widehat{y} = mode(\{h_1(x), h_2(x), \ ..., h_T(x)\}) \tag{2}$$

  where hi(x) = prediction of i-th tree, T = total number of trees, and mode = prediction of most frequent class.

  The implementation of random forest in this experiment using R is shown in Fig. 3.

```
#Random Forest
rf_model <- randomForest(as.factor(CLASS) ~., data = train_data, ntree =100)
rf_pred <- predict(rf_model, test_data)

#Evaluate model
rf_metrics <- evaluate_model(rf_pred, test_y)
```

Fig. 3 Implementation of Random Forest in R

- K-Nearest Neighbour (KNN)

  By identifying the data points that are closest to a given query and classifying the query based on those nearby instances, the K-nearest neighbour machine learning technique is utilised for classification. [4].

  The formula of Euclidean distance, used to find the nearest neighbours is shown in Eq. 3.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

The implementation of K-nearest neighbour in this experiment using R is shown in Fig. 4.

```
#k-Nearest Neighbors (k=5)
knn_pred <- knn(train = train_x, test = test_x, cl = train_y, k=5)

#Evaluate Model
knn_metrics <- evaluate_model(knn_pred, test_y)
```

Fig. 4 Implementation of K-Nearest Neighbour (KNN) in R

3.3 Evaluation Metrics

The evaluation metrics used in the experiments are accuracy, precision, recall, and F-score values.

- Accuracy. Accuracy is the evaluation of how effectively a model classifies instances into their correct categories [5]. The formula for calculating accuracy is shown in Eq. 4.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

The metrics evaluation of accuracy in this experiment using R is shown in Fig. 5.

```
#Evaluate Function and metrics
evaluate_model <- function(pred, actual) {
  cm <- confusionMatrix(as.factor(pred), as.factor(actual), positive = "1")
  precision <- cm$byClass["Precision"]
  recall <- cm$byClass["Recall"]
  f1 <- 2 * (precision * recall) / (precision + recall)
  accuracy <- cm$overall["Accuracy"]
  error_rate <- 1 - accuracy
  metrics <- c(
    Accuracy = accuracy,
    Error_Rate = error_rate,
    Precision = precision,
    Recall = recall,
    F1_Score = f1
  )
  return(metrics)
}
```

Fig. 5 Implementation of Accuracy Evaluation Metrics in R

- Precision. Precision computes the ratio of true positive predictions (TP) to the total positive predictions (TP+FP) [6]. The formula for calculating precision is shown in Eq. 5.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

The metrics evaluation of precision in this experiment using R is shown in Fig. 6.

```
precision <- cm$byClass["Precision"]
```

Fig. 6 Implementation of Precision Evaluation Metrics in R

- Recall. Recall is also referred to as the true-positive rate or sensitivity, and calculates the probability of the model correctly

identifying true positive cases [7]. The formula to calculate recall is shown in Eq. 6.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

The metrics evaluation of recall in this experiment using R is shown in Fig. 7.

```
recall <- cm$byClass["Recall"]
```

Fig. 7 Implementation of Recall Evaluation Metrics in R

- F1-score. F1-score represents the harmonic mean of precision and recall, offering a balanced evaluation score [8]. The formula to calculate F1-score is shown in Eq. 7.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

The metrics evaluation of recall in this experiment using R is shown in Fig. 8.

```
f1 <- 2 * (precision * recall) / (precision + recall)
```

Fig. 8 Implementation of F1-Score Evaluation Metrics in R

4.0 RESULTS

Classification in R involves several key steps to obtain accurate results. The implemented models start with installing and loading the necessary libraries and packages like caret, randomForest, class, and e1071, using install.packages() and library(). The diabetes dataset is then loaded and preprocessed. To ensure reproducibility, a seed is set using set.seed(). The data is divided into training and testing sets by creating an index. From the split dataset, the training dataset is used to build the model, while the testing dataset evaluates the model's performance. A formula is defined to specify the relationships between predictors and the target variable. Then the classification models of logistic regression, random forest, and k-nearest neighbours are built using glm() or train(). The models' performance is evaluated using accuracy, precision, recall, and F-score. Evaluation metrics are generated using the predict() function.

The purpose of the experiments is to make a comparison between the performance of Logistic Regression, Random Forest, and K-Nearest Neighbour algorithms in classifying the dataset for diabetes. The results are shown in Table 6.

Table 6. Result of Evaluation Metrics for each Algorithm

Data split (%)	Algorithm	Error	Accuracy	Precision	Recall	F1-score
30-70	Logistic Regression	0.0414	0.9586	0.9684	0.9855	0.9769
	Random Forest	0.0257	0.9743	0.9778	0.9936	0.9856
	KNN	0.0643	0.9357	0.9558	0.9727	0.9641
40-60	Logistic Regression	0.0317	0.9683	0.9795	0.9850	0.9822
	Random Forest	0.0267	0.9733	0.9796	0.9906	0.9851
	KNN	0.0533	0.9467	0.9596	0.9812	0.9703
50-50	Logistic Regression	0.0340	0.9660	0.9776	0.9842	0.9809
	Random Forest	0.0200	0.9800	0.9822	0.9955	0.9888
	KNN	0.0520	0.9480	0.9604	0.9820	0.9710
60-40	Logistic Regression	0.0375	0.9625	0.9746	0.9829	0.9787
	Random Forest	0.0175	0.9825	0.9831	0.9972	0.9901
	KNN	0.0600	0.9400	0.9529	0.9400	0.9663
70-30	Logistic Regression	0.0367	0.9633	0.9769	0.9807	0.9788
	Random Forest	0.0200	0.9800	0.9847	0.9923	0.9885
	KNN	0.0600	0.9400	0.9547	0.9768	0.9656

5.0 DISCUSSION

Random Forest performs most accurately in classifying the diabetes dataset, according to the results of the experiment. In general, across all of the data splits, it consistently has the highest accuracy, precision, recall, and F1-score. Its highest F1-score at a 60-40 train-test split is 0.9901. The reason for Random Forest's outstanding performance lies in its capability to process both linear as well as non-linear data relationships, manage complicated feature interactions, and minimize overfitting with ensemble learning. In medical classification tasks where the prediction accuracy is of utmost importance, its flexibility and stability make it ideal.

In addition, logistic regression is consistent and robust with high performance, especially when there are linear interactions between features and outcomes. It showed stability with various data splits and had a high F1-score, which is 0.9822 at best. However, when there are non-linear trends, its performance can be restricted. Moreover, K-Nearest Neighbour (KNN) fared poorly in this test. Higher error rates and less precise results are achieved because it is susceptible to feature scaling and poor performance under noisy or redundant features.

This experiment's data set includes a categorical variable which is gender and diabetes class, as well as a number of continuous variables, including age, blood pressure, BMI, and cholesterol level. Random Forest is well-suited to deal with this combination because it does not make strict assumptions about data distribution or variable type. Being a linear model, logistic regression will overlook complicated relationships, yet performs well for continuous features. Feature scaling and irrelevant features have a significant impact on KNN since it's based on distance measures. Random Forest's better performance can therefore be explained by how it handles continuous and categorical data with less preprocessing.

A study confirmed that Random Forest outperformed KNN and Logistic Regression on accuracy and F1-score when run on the Pima Indian Diabetes dataset in an experiment [9]. Seeing how Random Forest was able to cope better with variances and imbalances prevalent in medical data, the researchers highlighted how strong models are important in clinical applications. Their findings support the findings of the experiment, which illustrated how Random Forest consistently led on all metrics.

Furthermore, in a comparison of diabetes prediction machine learning algorithms, an analysis concluded that Random Forest was more precise and recallful than others [10]. The analysis noted that even though it was processing mixed types of data, its ensemble structure allowed it to avoid overfitting and make better predictions. This consistency in this experiment is in line with their research, where Random Forest had good predicted accuracy even for limited data.

Lastly, another recent paper demonstrates the efficacy of Random Forest in medical diagnosis applications [11]. According to their study, Random Forest outperformed KNN, Decision Trees, and Logistic Regression on diabetic data. They emphasized its high sensitivity and specificity, which are crucial in accurately diagnosing diabetic patients without generating false negative results. This validates the deduction that Random Forest is the most accurate and best model to predict diabetes, especially when using data sets with continuous and categorical values.

6.0 CONCLUSION

This study was able to effectively show how data mining methods can be utilized to predict diabetes using three classification models: K-Nearest Neighbour (KNN), Random Forest, and Logistic Regression. The research process was in accordance with the Knowledge Discovery in Database (KDD) process's well-defined steps that involve model selection, data transformation, and preprocessing. The data set, which was accessed from Mendeley Data consisted of 1,000 cases with categorical features such as gender and class as well as continuous features such as age, BMI, urea, HbA1c, and cholesterol. R programming was used to develop the models after data cleaning and transformation, which were then validated using conventional performance metrics such as accuracy, precision, recall, and F1-score. With the best F1-score of 0.9901 in a 60-40 train-test split, Random Forest outperformed the other two models consistently in all ratios of data split. Logistic regression with the best F1-score of 0.9822 was also extremely good, especially with linear feature correlations. KNN was behind mostly because of its inability to deal with irrelevant features and noise, and sensitivity to feature scaling. These results are also consistent with more recent research, which again emphasizes the way well ensemble methods like Random Forest perform on medical prediction tasks. This gives the model credibility in actual clinical decision systems and ensures that it is well-suited for datasets with varying types of data.

Future research needs to explore multiple paths of development and improvement in order to benefit from the good findings of this research. To start with, there is conceivably room to further improve classification accuracy with other types of machine learning algorithms like Support Vector Machines (SVM), Gradient Boosting, or even deep-learning techniques for that matter, particularly when working with larger and more complicated datasets. The model's interpretability and accuracy may also be improved through the inclusion of feature selection techniques or dimension reduction algorithms like Principal Component Analysis (PCA). The models can even spot temporal trends and patterns of improvement using time-series health data or electronic health

records (EHRs), which would provide additional clinical utility to the predictions. The potential for enhanced detection of chronic diseases, such as diabetes, through the use of deep learning and EHR data was highlighted in a research [17]. An additional key addition to the model would be making the model more generalisable between demographic groupings by using a more balanced and diverse dataset. Finally, the model's practical application might be evaluated by implementing it in healthcare systems and using explainability tools like SHAP or LIME in order to provide transparency and reliability to medical personnel. In addition to improving the predictive power, these improvements would facilitate the extended use of smart systems in prevention and customized care.

BIBLIOGRAPHY

[1] GeeksforGeeks. (2025, April 5). *Evaluation metrics in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/metrics-for-machine-learning-model/#accuracy.

[2] A. Das, "Logistic regression," in *Encyclopedia of Quality of Life and Well-Being Research*. Cham: Springer International Publishing, 2024, pp. 3985–3986.

[3] Z. Sun et al., "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Systems with Applications*, vol. 237, p. 121549, 2024.

[4] P. Cunningham and S. J. Delany, "K-nearest neighbour classifiers—a tutorial," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–25, 2021.

[5] C. Miller et al., "A review of model evaluation metrics for machine learning in genetics and genomics," *Frontiers in Bioinformatics*, vol. 4, p. 1457619, 2024.

[6] Ž. Vujović, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021.

[7] G. S. Handelman et al., "Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods," *American Journal of Roentgenology*, vol. 212, no. 1, pp. 38–43, 2019.

[8] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, 2024.

[9] M. Goyal and A. Kadam, "Performance evaluation of machine learning algorithms for diabetes prediction," *International Journal of Computer Applications*, vol. 182, no. 5, pp. 1–5, 2018.

[10] A. Kumar and G. Sahoo, "Prediction of diabetes disease using ensemble techniques," *International Journal of Scientific & Technology Research*, vol. 8, no. 11, pp. 3581–3585, 2019.

[11] M. Rashid and N. Aslam, "Comparative study of machine learning algorithms for diabetes prediction," *Journal of Healthcare Engineering*, vol. 2021, Article ID 6632645, 2021.

[12] World Health Organization. (2022). Diabetes. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[13] M. Shaik and A. S. Gowda, "Diabetes Prediction using Machine Learning Techniques," International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 6, 2020. [Online]. Available: https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques

[14] Ibrahim, S., & Khairi, S. S. M. (2022). Predictive Data Mining Approaches for Diabetes Mellitus Type II Disease. International Journal of Global Optimization and Its Application, 1(2), 126-134.

[15] Islam, M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In Computer Vision and Machine Intelligence in Medical Image Analysis: International Symposium, ISCMM 2019 (pp. 113-125). Springer Singapore.

[16] Khairudin, Z., Abdul Razak, N. A., Abd Rahman, H. A., Kamarudin, N., & Abd Aziz, N. A. (2020). Prediction of diabetic retinopathy among type II diabetic patients using data mining techniques. Malaysian Journal of Computing (MJoC), 5(2), 572-586.

[17] 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. IEEE Access, 9, 43711-43735.

[18] Oon, N. B., Khairudin, Z., Abd Rahman, H. A., Kamarudin, N., Abu Bakar, N. S., & Abd Aziz, N. A. (2024). Prediction of diabetic retinopathy among diabetic neuropathy in T2DM patients using data mining algorithm. Malaysian Journal of Computing (MJoC), 9(2), 1916-1929.

[19] Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, 100605.