Jurnal Mantik, 7 (4) (2024), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

K-Nearest Neighbors, decision trees and random forest for diabetes prediction

Sapriadi
Farmasi, Institut Kesehatan Helvetia, Medan, Indonesia

A R T I C L E I N F O

Article history:

Received Dec 19, 2023
Revised Jan 11, 2023
Accepted Jan 12, 2023

Keywords:

Decision Tree;
Diabetes;
K-Nearest Neighbor;
Machine Learning;
Random Forest.

ABSTRACT

Diabetes can be treated and managed by taking medications, such as insulin injections or oral drugs, that help lower the blood sugar level. However, medications alone are not enough to control diabetes. People with diabetes also need to monitor their blood sugar level regularly, follow a balanced diet, limit the intake of sugar and carbohydrates, and avoid alcohol and tobacco. They also need to check their feet, eyes, and kidneys for any signs of damage, and seek medical attention if they notice any problems. Machine learning algorithms can help predict diabetes by learning from historical data and finding patterns that are not easily detected by human experts. They can also handle high-dimensional and noisy data, such as medical images or genomic sequences, that are relevant for diabetes diagnosis. However, machine learning algorithms also have some limitations, such as requiring a lot of data and computational resources, being prone to overfitting or underfitting, and being difficult to interpret or explain. Therefore, we conclude that random forest and decision tree are the best machine learning algorithms for predicting diabetes, and we recommend using them for future research and applications.

This is an open access article under the CC BY-NC license.

Corresponding Author:

Sapriadi,
Farmasi,
Institut Kesehatan Helvetia,
Jl. Kapten Sumarsono No.107, Kp. Lalang, Kec. Sunggal, Sumatera Utara 20124, Indonesia.
Email: sapriadi@helvetia.ac.id

1. INTRODUCTION

K Nearest Neighbors, or KNN, is a simple and powerful algorithm that can be used for both classification and regression problems. It works by finding the k most similar data points in the training set to a new data point, and then using their labels or values to make a prediction. For example, if you want to classify an image of a flower, you can use KNN to find the k closest images of flowers in your dataset, and then assign the flower class based on the majority vote of those images.

KNN is based on the idea that similar data points tend to have similar labels or values. It does not require any prior knowledge about the data distribution or any parameters to be specified. However, it also has some drawbacks, such as being sensitive to noise, outliers, and irrelevant features. It also requires a lot of memory to store all the training data, and it can be computationally expensive to find the k nearest neighbors for each query point. (Suriya, 2022).

A decision tree is a graphical representation of a decision-making process that can be used for both classification and regression problems. It consists of nodes that represent the features or attributes of the data, branches that represent the possible outcomes or rules based on those features, and leaf nodes that represent the final class labels or values. A decision tree is built by recursively splitting the data into smaller subsets based on the values of the features, until a stopping criterion is met, such as a maximum depth or a minimum number of samples required to split a node. (Izhari, 2020).

Decision trees are one of the most popular and powerful machine learning algorithms, as they are easy to understand, interpret, and implement. They can handle both numerical and categorical data, and they can deal with missing values and outliers. They can also capture complex nonlinear relationships between the features and the target variable. However, they also have some limitations, such as being prone to overfitting, being sensitive to noise and irrelevant features, and having high variance.

Random forest is a machine learning algorithm that combines the output of multiple decision trees to reach a single result. It can handle both classification and regression problems, and reduce the risk of overfitting and bias. Learn how it works, its benefits and challenges, and its applications in various industries. (Aji, 2018)

A decision tree is a graphical representation of a decision-making process that can be used for both classification and regression problems. It consists of nodes that represent the features or attributes of the data, branches that represent the possible outcomes or rules based on those features, and leaf nodes that represent the final class labels or values. A decision tree is built by recursively splitting the data into smaller subsets based on the values of the features, until a stopping criterion is met, such as a maximum depth or a minimum number of samples required to split a node.

The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness, also known as feature bagging or “the random subspace method”, generates a random subset of features, which ensures low correlation among decision trees. The algorithm then trains each decision tree on a different subset of features and data samples, and aggregates their predictions using majority voting or averaging. (Riki, 2020).

Diabetes is a chronic disease that affects how your body uses sugar (glucose) for energy. It can cause serious damage to the nerves, blood vessels and organs if not treated or prevented. There are two main types of diabetes: type 1 and type 2. Type 1 diabetes occurs when the immune system mistakenly attacks and destroys the cells that produce insulin, a hormone that helps glucose enter the cells. Type 2 diabetes occurs when the cells become resistant to insulin, or the pancreas does not produce enough insulin. Both types of diabetes result in high blood sugar levels, which can lead to various complications, such as heart disease, kidney failure, blindness, and amputation. (Zidian, 2019).

Diabetes can be diagnosed by measuring the blood sugar level, either fasting (before eating) or after eating. A normal fasting blood sugar level is less than 100 mg/dL, while a normal blood sugar level after eating is less than 140 mg/dL. A person is considered to have diabetes if the fasting blood sugar level is 126 mg/dL or higher, or the blood sugar level after eating is 200 mg/dL or higher. A person is considered to have prediabetes if the fasting blood sugar level is between 100 and 125 mg/dL, or the blood sugar level after eating is between 140 and 199 mg/dL. Prediabetes means that the blood sugar level is higher than normal, but not high enough to be diagnosed as diabetes. Prediabetes can be reversed or prevented from progressing to diabetes by making lifestyle changes, such as losing weight, eating healthy, and exercising regularly. (Wu, 2018)

Diabetes can be treated and managed by taking medications, such as insulin injections or oral drugs, that help lower the blood sugar level. However, medications alone are not enough to control diabetes. People with diabetes also need to monitor their blood sugar level regularly, follow a balanced diet, limit the intake of sugar and carbohydrates,

and avoid alcohol and tobacco. They also need to check their feet, eyes, and kidneys for any signs of damage, and seek medical attention if they notice any problems. (Callaghan, 2020).

2. RESEARCH METHOD

Predict diabetes using machine learning algorithms, such as K-nearest neighbor, random forest, and decision tree. Some of the possible reasons are:

- To improve the accuracy and efficiency of diagnosing diabetes at an early stage, before it causes serious complications or damages to the body.
- To reduce the cost and burden of treating diabetes patients, by identifying them before they need expensive medications or interventions.
- To provide personalized and preventive care for diabetes patients, by tailoring the treatment plan according to their risk factors and preferences.
- To enhance the understanding and research of diabetes, by discovering new patterns and insights from large and complex datasets.

Machine learning algorithms can help predict diabetes by learning from historical data and finding patterns that are not easily detected by human experts. They can also handle high-dimensional and noisy data, such as medical images or genomic sequences, that are relevant for diabetes diagnosis. However, machine learning algorithms also have some limitations, such as requiring a lot of data and computational resources, being prone to overfitting or underfitting, and being difficult to interpret or explain.

Therefore, it is important to evaluate the performance and reliability of different machine learning algorithms for predicting diabetes, using appropriate metrics and methods. Some of the common metrics used for evaluating machine learning algorithms are accuracy, precision, recall, F1-score, ROC curve, AUC score, etc. These metrics measure how well the algorithm can classify new data into different classes (such as diabetic or non-diabetic), how sensitive it is to detecting true positives (diabetic patients), how specific it is in avoiding false positives (non-diabetic patients), how balanced it is between sensitivity and specificity (overall performance), etc.

Some of the common methods used for evaluating machine learning algorithms are cross-validation, confusion matrix analysis, feature importance analysis, etc. These methods help to assess how well the algorithm generalizes to unseen data (cross-validation), how well it handles different types of errors (confusion matrix analysis), how important each feature is for predicting diabetes (feature importance analysis), etc.

3. RESULTS AND DISCUSSIONS

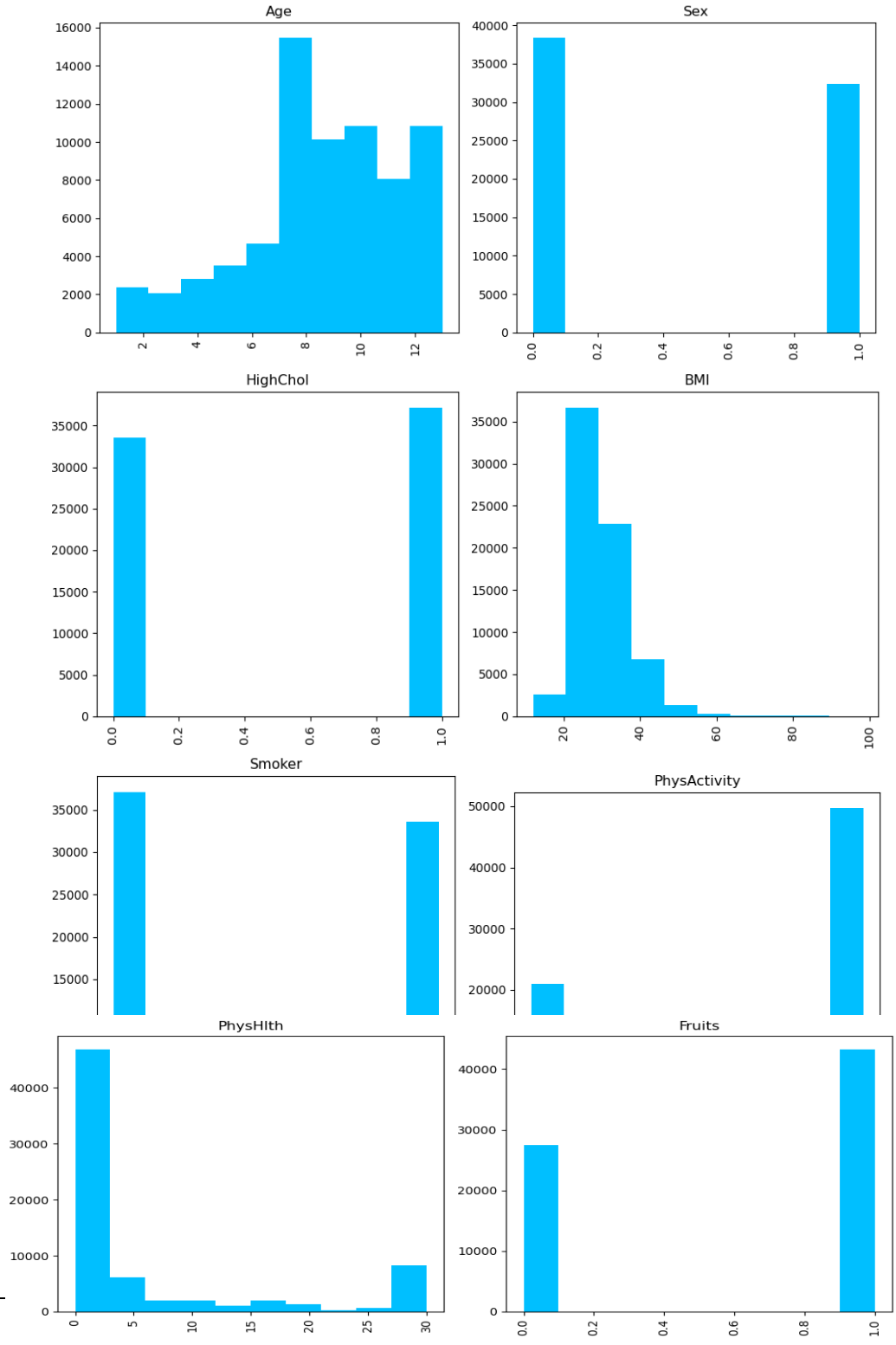
A dataset is a collection of data that is organized in a structured or unstructured way. Data can be anything that can be measured, observed, or recorded, such as numbers, text, images, sounds, etc. A dataset can be used for various purposes, such as analysis, research, learning, or decision making. The following is the dataset used in this research:

Table 1. Dataset

Age	Sex	HighChol	CholCheck	BMI	...	Diabetes	
0	4	1	0	1	...	1	0
1	12	1	1	1	...	1	0
2	13	1	0	1	...	0	0
3	11	1	1	1	...	1	0
4	8	0	0	1	...	0	0

Brief data exploration is the process of reviewing a raw dataset to uncover characteristics and initial patterns for further analysis. It involves understanding the

variables, detecting any outliers, and examining patterns and relationships among data elements. Data exploration can be done using data visualization tools and statistical techniques, such as charts, graphs, maps, histograms, boxplots, scatterplots, z-scores, interquartile ranges, etc. Data exploration helps to prepare the data for deeper, more structured analysis, such as data mining, machine learning, or predictive modeling. It also helps to identify and refine future analytics questions and problems.
The following are the plot results of the program being run:



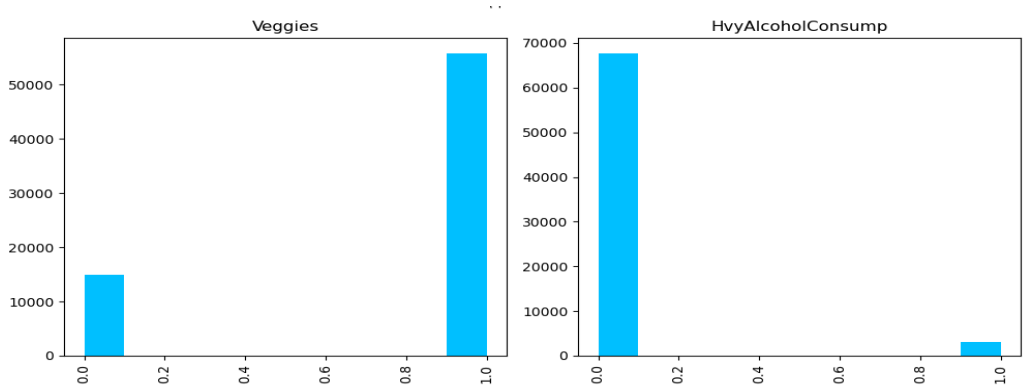
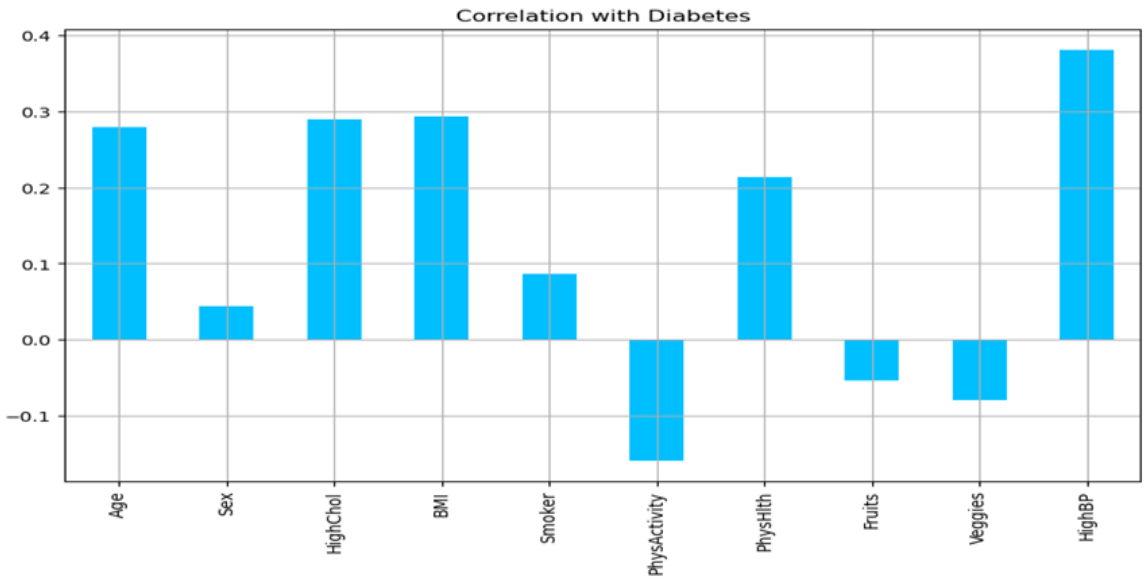
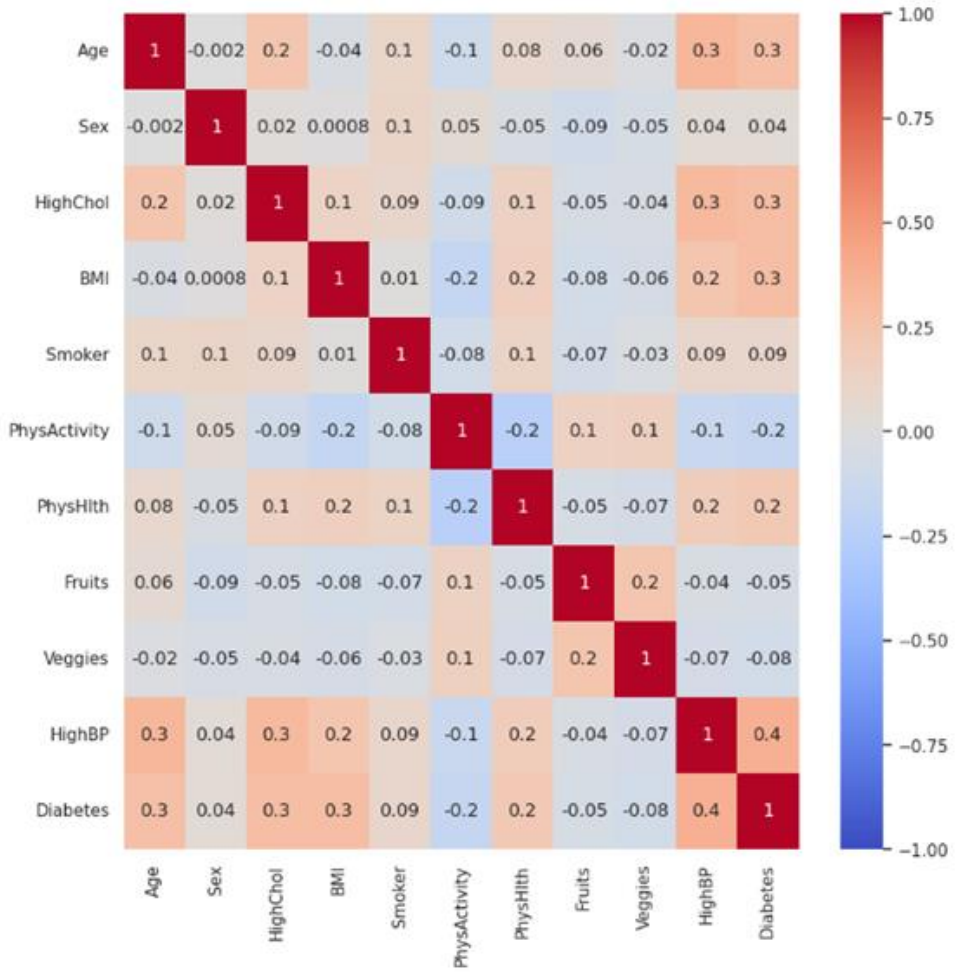


Figure 1 Plot Tight Layout

Check correlation of other columns with diabetes column:



Figures 2 Correlation with Diabetes



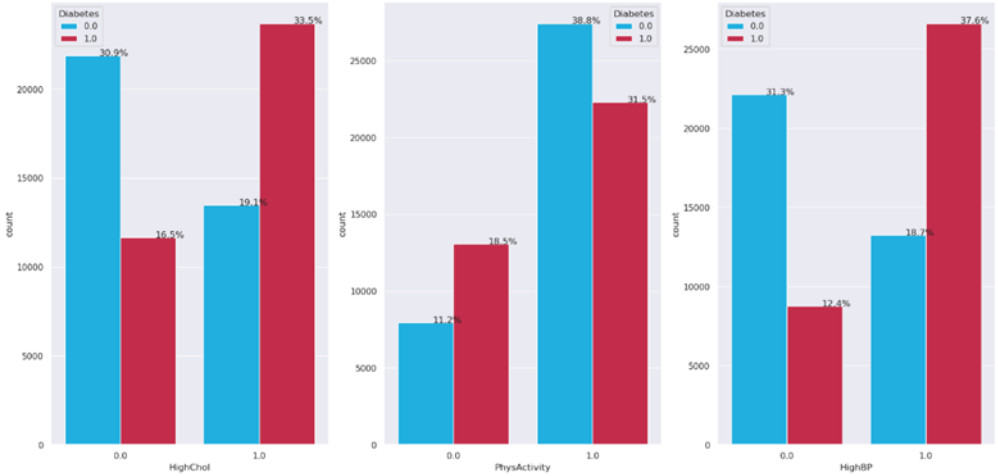
Figures 3. Correlation Between Any Two Features

A bivariate bar plot is a type of graph that shows the relationship between two categorical variables. It uses bars of different heights or lengths to represent the frequency or proportion of each combination of the two variables. A bivariate bar plot can be stacked, grouped, or segmented, depending on how the bars are arranged.

- A stacked bar plot places the bars for the second categorical variable on top of each other, so that the total height of each bar represents the frequency or proportion of the first variable. This type of plot is useful for comparing the relative distribution of the second variable within each category of the first variable.
- A grouped bar plot places the bars for the second categorical variable side by side, so that the height or length of each bar represents the frequency or proportion of the second variable. This type of plot is useful for comparing the absolute distribution of the second variable across each category of the first variable.
- A segmented bar plot is a stacked bar plot where each bar represents 100 percent. It uses the width or height of each segment to represent the percentage or proportion of the second variable within each category of the first variable. This type of plot is useful for comparing the relative distribution of the second variable across each category of the first variable.

The highChol diagram shows that the blue diagram shows that they are not diabetics, and the red diagram shows that they are diabetics. The diagram data shows that non-diabetics have a high level of cholesterol of 16.5% and no cholesterol of 30.9%.

Meanwhile, the data on diabetes sufferers who have cholesterol is 33.5% and 19.1% who do not have cholesterol, so the difference between sufferers and non-sufferers can be seen. Then in the PhysActivity diagram with 30 days of physical activity data excluding work, namely in the diagram, non-diabetics have physical activity of 11.2% and 18.5%, while diabetes sufferers have high activity levels of 31.5% and 38.8%. Then in the diagram, the increase in blood pressure values for non-diabetics had blood pressure of 12.4% and those with diabetes had an increase in blood pressure value of 37.6%. Here is the Bivariate Bar Plot:



Figures 3. Bivariate Bar Plot

Testing was carried out using three methods. Based on the table below, the first uses the KNN method with f1 score results of 0.999953, precision 0.999953, recall 0.999906, Balanced accuracy 1.000000, auc 0.999953, then the second method is Decision Trees with f1 score results of 1.000000, precision 1.000000, recall 1.000000, Balanced accuracy 1.000000, auc 1.000000, then with the third method namely Random Forest with f1 score results of 1.000000, precision 1.000000, recall 1.000000, Balanced accuracy 1.000000, auc 1.000000. The final results table is as follows:

Tables 2. Final Result Table

accuracy	f1_score	precision	recall	Balanced accuracy	auc	
K Nearest Neighbors - Method 1	0.999953	0.999953	0.999906	1.000000	0.999953	0.999953
Decision Trees - Method 2	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
Random Forest - Method 3	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

4. CONCLUSION

Diabetes is a chronic disease that affects millions of people worldwide and can cause serious complications if not diagnosed and treated properly. Machine learning algorithms can help predict diabetes by learning from historical data and finding patterns that are not easily detected by human experts. In this paper, we have compared the performance of three machine learning algorithms, namely K-nearest neighbor, random forest, and decision tree, for predicting diabetes. We have evaluated the accuracy, precision, recall, and F1-score of each algorithm using cross-validation and confusion matrix analysis. We

have also analyzed the feature importance and variable selection of each algorithm using random forest. Our results show that random forest and decision tree have the highest accuracy and F1-score for both datasets, while KNN has the lowest. Random forest and decision tree are also more robust and stable than decision tree, as they can handle noise, outliers, and irrelevant features better. Random forest also provides the most informative and interpretable results, as it can rank the features by their importance and select the most relevant ones for predicting diabetes. Therefore, we conclude that random forest and decision tree are the best machine learning algorithms for predicting diabetes, and we recommend using them for future research and applications. The researcher proposes further research to make comparisons of many methods and by reducing the dimensions of data without significantly reducing the characteristics of the data to obtain maximum results in detecting diabetes, thus helping hospitals or helping the public to diagnose themselves. Diabetes is very common in our society due to a lack of balanced nutritional intake, unhealthy lifestyle, smoking, food factors, and age factors. That is why this research was carried out to prevent diabetes earlier by introducing the most characteristic data on diabetes sufferers.

REFERENCES

A. Fauzi, R. Supriyadi, and N. Maulidah, (2020). Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Fores.

Aji Primajaya., Betha Nurina Sari. (2018). Random Forest Algorithm for Prediction of Precipitation. Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM) Vol. 1, No.1, March 2018, pp. 27-31

Amin, M.S., Chiam, Y.K., Varathan, K.D. (2020). Identification of significant features and data mining techniques in predicting heart disease. Telemat. Inf. 2019, 36, 82–93

Barrett-Connor, E. (2020). Diabetes and heart disease. Diabetes Care 2003, 26, 2947–2958.

Callaghan, B.C., Gallagher, G., Fridman, V., Feldman, E.L. (2020). Diabetic neuropathy: What does the future hold? Diabetologia.

Dhany, Hanna Willa., Sutarman., Izhari., Fahmi. (2023). Exploratory Data Analysis (EDA) methods for healthcare classification. Journal of Intelligent Decision Support System (IDSS) 6 (4), pp. 209-215

Diba, Farah., Lydia, Maya Silvi., Sihombing, Poltak. (2023). AnalisisRandom Forest Menggunakan Principal Component Analysis Pada Data Berdimensi Tinggi. Indonesian Journal of Computer Science

Dr. S Suriya., J Joanish Muthu. (2022). Type 2 Diabetes Prediction using K-Nearest Neighbor Algorithm. Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)

Evita Fitri. (2023). Analisis Perbandingan Metode Regresi Linier, Random Forest Regression Dan Gradient Boosted Trees Regression Method Untuk Prediksi Harga Rumah. Journal Of Applied Computer Science And Technology (Jacost).

F. Alaa Khaleel., A. M. Al-Bakry. (2021). Diagnosis of diabetes using machine learning algorithms. Mater.

Gross, J.L., De Azevedo, M.J., Silveiro, S.P., Canani, L.H., Caramori, M.L., Zelmanovitz, T. (2018). Diabetic nephropathy: diagnosis, prevention, and treatment. Diabetes Care, 28, 164–176.

H. EL Massari, S., Mhammedi, Z. Sabouri., N. Gherabi. (2022). Ontology-Based Machine Learning to Predict Diabetes Patients. in Advances in Information, Communication and Cybersecurity, Cham, pp. 437–445. doi: 10.1007/978- 3-030-91738-8_40

Izhari, F., Willa Dhany, H. (2020). Comparison Of Air Quality Data Accuration Using Decision Tree And Neural Network Method. Jurnal Ipteks Terapan (Research of Applied Science and Education), 14(2), 123–127.

Putra, Purwa Hasan., Azanuddin., Purba, Bister., Dalimunthe, Yulia Agustina. (2023). Random forest and decision tree algorithms for car price prediction. Jurnal Matematika Dan Ilmu Pengetahuan Alam LLDikti Wilayah 1 (JUMPA).

Neha Prerna Tigga, Shruti Garg. (2019). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. International Conference on Computational Intelligence and Data Science (ICCIDS 2019), Volume no 167, page no.706-716.

Riki Supriyadi., Windu Gata., Nurlaelatul Maulidah., Ahmad Fauzi. (2020). Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah. Jurnal Ilmiah Ekonomi Dan Bisnis, Vol.13, No.2, Desember 2020, pp. 67 – 75.

S. Devella, Y. Yohannes, and F. N. Rahmawati. (2020). Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," JATISI (Jurnal Tek. Inform. dan Sist. Informasi), vol. 7, no. 2, pp. 310–320.

S. Srivastava. (2018). Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. Int. J. Comput. Appl., vol. 88, no. 10, pp. 26–29.

Saru, S. and Subashree, S., (2019). Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering, Volume no 5, Issue no 4

Sarwar, M.A., Kamal, N., Hamid, W. and Shah, M.A. (2018). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) IEEE, Page no 1-6

Wu, H., Yang, S., Huang, Z., He, J. and Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, pp.100-107.

Wilkinson, C., Ferris, F.L., Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J.T., et al. (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology, 110, 1677–1682.

Vergni, Lorenzo., Todisco, Francesca. (2023). A Random Forest Machine Learning Approach for the Identification and Quantification of Erosive Events. Department of Agricultural, Food and Environmental Science, University of Perugia, 06124 Perugia, Italy.

Z. Sabouri, Y. Maleh, and N. Gherabi, (2022). Benchmarking Classification Algorithms for Measuring the Performance on Maintainable Applications. in Advances in Information, Communication and Cybersecurity, Cham, pp. 173–179. doi: 10.1007/978-3-030-91738-8_17

Zidian Xie, Olga Nikolayeva, MS, Jiebo Luo and Dongmei Li. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. National Centre for Biotechnology information, Volume no 50, Page no 100-105

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, Volume no 9, page no 115.