Journal of Science and Social Research ISSN 2615 – 4307 (Print)
August 2025, VIII (3): 3776 – 3782 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

DIABETES PREDICTION BASED ON MEDICAL RECORDS (PIMA INDIANS DIABETES DATASET) USING K-NN

Fahmi Ruziq[1], M. Rhifky Wayahdi[2], Subhan Hafiz Nanda Ginting[3]
Battuta University, Medan
email: [1]fahmiruziq89@gmail.com, [2]muhammadrhifkywayahdi@gmail.com,
[3]subhanhafiz16@gmail.com

Abstract: *The development of predictive technologies, especially artificial intelligence (AI) and machine learning, has opened up great opportunities in the health sector, including early detection of chronic diseases such as diabetes. This study aims to implement the K-Nearest Neighbors (KNN) algorithm in predicting the likelihood of a person having diabetes based on medical record data from the Pima Indians Diabetes Dataset. The dataset consists of 768 samples with eight key health features. The analysis process includes data cleaning, data distribution exploration, and data preparation for the modelling process. The distance between data is calculated using the Euclidean formula, and normalization is performed so that all features have equal weight. The data was then divided into training and test data with a ratio of 80:20. The analysis results showed an unbalanced class distribution, with more non-diabetic patients than those with diabetes. The age group of 21-30 years dominates in the dataset. The implementation of KNN in this study shows that the method is effective for medical classification based on numerical data. This research demonstrates the potential of KNN as a practical and easy-to-implement early diagnosis tool in data-driven health systems.*

Keyword: *K-Nearest Neighbors, diabetes prediction, machine learning, medical data, classification.*

Abstrak: Perkembangan teknologi prediktif, khususnya kecerdasan buatan (AI) dan pembelajaran mesin (*machine learning*), telah membuka peluang besar dalam bidang kesehatan, termasuk deteksi dini penyakit kronis seperti diabetes. Penelitian ini bertujuan untuk mengimplementasikan algoritma *K-Nearest Neighbors* (KNN) dalam memprediksi kemungkinan seseorang menderita diabetes berdasarkan data rekam medis dari Pima Indians Diabetes Dataset. Dataset terdiri dari 768 sampel dengan delapan fitur kesehatan utama. Proses analisis meliputi pembersihan data, eksplorasi distribusi data, serta persiapan data untuk proses modeling. Jarak antar data dihitung menggunakan rumus Euclidean, dan dilakukan normalisasi agar seluruh fitur memiliki bobot yang seimbang. Data kemudian dibagi menjadi data latih dan uji dengan rasio 80:20. Hasil analisis menunjukkan distribusi kelas yang tidak seimbang, dengan jumlah pasien non-diabetes lebih banyak dibandingkan yang menderita diabetes. Kelompok usia 21–30 tahun mendominasi dalam dataset. Implementasi KNN dalam studi ini menunjukkan bahwa metode ini efektif digunakan untuk klasifikasi medis berbasis data numerik. Penelitian ini mendemonstrasikan potensi KNN sebagai alat bantu diagnosis awal yang praktis dan mudah diimplementasikan dalam sistem kesehatan berbasis data.

Kata kunci: K-Nearest Neighbors, prediksi diabetes, machine learning, data medis, klasifikasi.

INTRODUCTION

The rapid development of technology (Durney & Donnelly, 2015) has brought various conveniences to humans (Mick & Fournier, 1998) in solving problems, including in the health sector (Umamaheswaran et al., 2022).

One of the most significant innovations is the application of predictive technology (Nearing et al., 1990), which enables data processing to support faster and more accurate decision-making (Eisenhardt, 2017) (Ruziq et al., 2022) (Wayahdi & Ruziq, 2024) (Ruziq & Wayahdi, 2024). In the medical field, these technologies play an important role in early diagnosis of diseases (Uddin et al., 2019) (Rhifky Wayahdi et al., 2022), prediction of health risks (Raza et al., 2022), and provision of treatment recommendations (Sesen et al., 2012) based on available data.

One form of predictive technology that is gaining popularity is artificial intelligence (AI). By utilizing machine learning algorithms, AI is able to process medical data to make predictions or classifications (Chang et al., 2019), such as detecting disease types (Ali et al., 2020) or predicting a patient's future condition (Sahoo et al., 2016). Examples of AI implementation in the healthcare field include decision support systems based on medical record data (Zikos & Delellis, 2018) to automated diagnosis applications (Wormanns et al., 2002) that can be used by doctors and other health workers.

Among the various methods used in machine learning, K-Nearest Neighbors (KNN) is one of the most widely used algorithms for prediction and classification tasks (Lubis, Lubis, & Al-Khowarizmi, 2020). KNN works on the principle of finding the closest distance (Halder et al., 2024) (Wormanns et al., 2002) between new data and data with known classification. Its non-parametric nature (Chirici et al., 2012) and ease of implementation (Adeniyi et al., 2016) make KNN a suitable choice for many case studies. In addition, this method can provide competitive results if parameters such as the number of neighbors (k) and distance metric are well optimized.

Several previous studies have shown the success of the KNN method in various case studies, including the prediction of diabetes and other health problems. For example, a study on identifying students at high risk of failure in the early stages of a course used KNN and the results showed that KNN can accurately predict student performance, and even after the first lesson (Tanner & Toivonen, 2010). Other studies implemented KNN for plant species classification based on leaf and stem size data (Ghosh et al., 2022), as well as for heart disease diagnosis based on patient medical records (Riyaz et al., 2022). However, these studies also faced challenges, such as selecting optimal parameters and handling large datasets.

In the context of diabetes prediction, predictive technologies play an important role, given that diabetes is one of the chronic diseases (Liu et al., 2020) that has a significant impact globally. Early detection of diabetes through data-driven prediction can help prevent serious complications and reduce the burden of treatment costs.

Based on this background, this research aims to implement the KNN method in a case study of diabetes prediction based on medical records. This research will also evaluate the performance of the KNN algorithm in this context, so as to provide an overview of the advantages and challenges faced in its use.

METHOD

The KNN algorithm is one of the most popular data mining algorithms. It has been widely and successfully applied to data analysis applications in various research topics in computer science (Zhang, 2022). The K-Nearest Neighbor (KNN) method is used to classify objects based on the closest learning data. KNN finds the K nearest neighbors of the query point based on attributes and training data, then determines the classification (Lubis, Lubis, & Khowarizmi, 2020).

The distance calculation is done using the Euclidean formula, which is simple and effective. However, if the attributes have a large range of values,

this may affect the results. Therefore, the data is normalized in the range of 0 to 1 by dividing the attribute values based on their range (maximum - minimum values). This normalization ensures all attributes have a balanced influence.

This research aims to implement the K-Nearest Neighbors (KNN) method in a diabetes prediction case study based on patient medical record data. The steps in this research are described as follows:

1. Data Collection

The data used in this study is taken from the Pima Indians Diabetes Dataset available on the Kaggle website. This dataset contains medical data of Pima tribal patients, which is used to predict the likelihood of a person having diabetes based on several health parameters.

The dataset includes 768 samples with 8 main features, viz:

- Number of pregnancies,
- Glucose level,
- Blood pressure,
- Skin thickness,
- Insulin level,
- Body mass index (BMI),
- Diabetes Pedigree Function, dan
- Age.

Each data is also labeled with the patient's diabetes status (1 for positive diabetes, 0 for negative diabetes). This dataset is highly relevant for the implementation of the K-Nearest Neighbors (KNN) method, as the features are numeric and suitable for use with distance-based algorithms such as KNN.

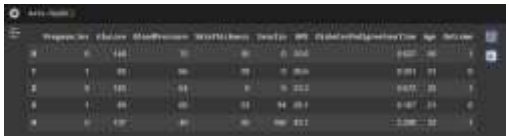

Figure 1 A snapshot of the initial five data points from the Pima Diabetes dataset.

2. Data Preparation and Pre-Processing

The dataset obtained will be further processed through the following steps:

Data Cleaning

At this stage, the data used has been checked to ensure there are no missing values or duplicates. The results showed that there were no missing values in the columns, such as Pregnancies, Glucose, BloodPressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome. In addition, no duplicated data was found in the dataset.

```
data.isnull().sum()
data.duplicated().sum()
```

Figure 2 Check for empty data and duplication in the dataset.

Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to understand the distribution and characteristics of the data. One of the steps taken was to visualize the distribution of the target variable (Outcome) using countplots. In addition, the data was also grouped by age group (Agegroup) to see the age distribution in the dataset.

```
sns.countplot(y='Outcome', data=data, palette='flare')
plt.ylabel('Outcome')
plt.xlabel('Count')
plt.show()
```

Figure 3 Distribution of the number of patients based on diabetes status.

Data Preparation

The data was prepared for the modeling process by separating the independent (X) and dependent (y) variables. The independent variables consist of Pregnancies, Glucose, BloodPressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, and Age, while the dependent variable is Outcome. The data is then divided into training data (train) and test data (test) with a ratio of 80:20.

```
X = data.drop(columns=['Outcome'])
y = data['Outcome']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 4 Division of data into training data and test data.

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
August 2025, VIII (3): 3776 – 3782 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

RESULT AND DISCUSSION

Data Distribution

The results of the exploratory analysis show the distribution of the target variable (Outcome). Visualization using a countplot shows that the number of individuals with a positive outcome of diabetes (Outcome = 1) is less compared to individuals with a negative outcome (Outcome = 0). This indicates class imbalance in the dataset.

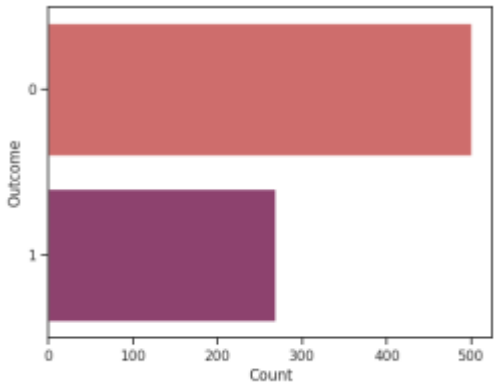

Figure 5 Outcome distribution using countplot.

Age Group Distribution

Data is grouped by age range to see the age distribution in the dataset. The visualization results show that the 21-25 age group has the highest number, followed by the 26-30 age group. This distribution gives an idea of the characteristics of the population in the dataset.

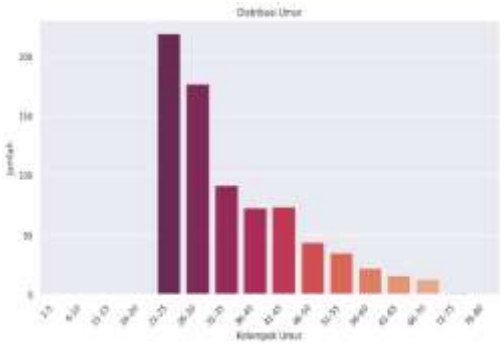

Figure 6 Age group distribution using countplot.

Data Preparation for Modeling

After the data is cleaned and analyzed, it is prepared for the modeling process. The data is divided into training data and test data with a ratio of 80:20. The results of data division show that the training data consists of 614 samples, while the test data consists of 154 samples.

```
print(f"x_train : {x_train.shape}")
print(f"y_train : {y_train.shape}")
print(f"x_test : {x_test.shape}")
print(f"y_test : {y_test.shape}")
```

```
x_train : (614, 8)
y_train : (614,)
x_test : (154, 8)
y_test : (154,)
```

Figure 7 Distribution of training and test data.

CONCLUSSION

This research shows that the K-Nearest Neighbors (KNN) algorithm can be effectively implemented to classify the likelihood of diabetes based on patient medical record data. The process starts from data collection and cleaning, exploratory analysis to understand the data distribution, to the data preparation stage before modeling.

The results of data exploration showed the presence of class imbalance in the target variable (Outcome) and the dominance of certain age groups in the dataset. After the data is separated into training data and test data, the KNN modeling process can be performed optimally by considering data normalization so that attributes have a balanced influence.

By using a distance-based approach such as KNN and applying appropriate pre-processing techniques, the model is able to utilize the numerical features in the dataset to accurately predict diabetes status. This research also confirms the importance of parameter selection and handling class imbalance to improve the performance of predictive models in healthcare.

REFERENCES

Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, *12*(1), 90–108. https://doi.org/10.1016/J.ACI.2014.10.001

Ali, F., El-Sappagh, S., Islam, S. M. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, *63*, 208–222. https://doi.org/10.1016/J.INFFUS.2020.06.008

Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S., & Zhou, S. (2019). A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. *Diagnostics 2019, Vol. 9, Page 178*, *9*(4), 178. https://doi.org/10.3390/DIAGNOSTICS9040178

Chirici, G., Corona, P., Marchetti, M., Mastronardi, A., Maselli, F., Bottai, L., & Travaglini, D. (2012). K-NN FOREST: a software for the non-parametric prediction and mapping of environmental variables by the k-Nearest Neighbors algorithm. *European Journal of Remote Sensing*, *45*(1), 433–442. https://doi.org/10.5721/EUJRS20124536

Durney, C. P., & Donnelly, R. G. (2015). Managing the Effects of Rapid Technological Change on Complex Information Technology Projects. *Journal of the Knowledge Economy*, *6*(4), 641–664. https://doi.org/10.1007/S13132-012-0099-2/METRICS

Eisenhardt, K. M. (2017). Making Fast Strategic Decisions In High-Velocity Environments. *Https://Doi.Org/10.5465/256434*, *32*(3), 543–576. https://doi.org/10.5465/256434

Ghosh, S., Singh, A., Kavita, Jhanjhi, N. Z., Masud, M., & Aljahdali, S. (2022). SVM and KNN Based CNN Architectures for Plant Classification. *Computers, Materials & Continua*, *71*(3), 4257–4274. https://doi.org/10.32604/CMC.2022.023414

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data 2024 11:1*, *11*(1), 1–55. https://doi.org/10.1186/S40537-024-00973-Y

Liu, H., Chen, S., Liu, M., Nie, H., & Lu, H. (2020). Comorbid Chronic Diseases are Strongly Correlated with Disease Severity among COVID-19 Patients: A Systematic Review and Meta-Analysis. *Aging and Disease*, *11*(3), 668. https://doi.org/10.14336/AD.2020.0502

Lubis, A. R., Lubis, M., & Al-Khowarizmi. (2020). Optimization of distance formula in K-Nearest Neighbor method. *Bulletin of Electrical Engineering and Informatics*, *9*(1), 326–338. https://doi.org/10.11591/EEI.V9I1.1464

Lubis, A. R., Lubis, M., & Khowarizmi, A.-. (2020). Optimization of distance formula in K-Nearest Neighbor method. *Bulletin of Electrical Engineering and Informatics*, *9*(1), 326–338. https://doi.org/10.11591/eei.v9i1.1464

Mick, D. G., & Fournier, S. (1998). Paradoxes of Technology: Consumer Cognizance, Emotions, and Coping Strategies. *Journal of Consumer Research*, *25*(2), 123–143. https://doi.org/10.1086/209531

Nearing, M. A., Lane, L. J., Alberts, E. E., & Laflen, J. M. (1990). Prediction Technology for Soil Erosion by Water: Status and Research Needs. *Soil Science Society of America Journal*, *54*(6), 1702–1711. https://doi.org/10.2136/SSSAJ1990.03615995005400060033X

Raza, A., Siddiqui, H. U. R., Munir, K., Almutairi, M., Rustam, F., & Ashraf, I. (2022). Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction. *PLOS ONE*, *17*(11), e0276525. https://doi.org/10.1371/JOURNAL.PONE.0276525

Riyaz, L., Butt, M. A., Zaman, M., & Ayob, O. (2022). *Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review*. 81–94. https://doi.org/10.1007/978-981-16-3071-2_8

Ruziq, F. & Wayahdi, M. R. (2022). Sistem Pendukung Keputusan Seleksi Karyawan Baru dengan Simple Additive Weighting pada PT. Technology Laboratories Indonesia. *Jurnal Minfo Polgan*, *11*(2), 153–159. https://doi.org/10.33395/JMP.V11I2.13506

Ruziq, F. & Wayahdi, M. R. (2024). Implementation of SAW Method in Website-Based Application (Case Study: New Employee Recruitment at PT. Technology Laboratories Indonesia). *Jurnal Minfo Polgan*, *13*(1), 1220–1227. https://doi.org/10.33395/JMP.V13I1.13998

Sahoo, P. K., Mohapatra, S. K., & Wu, S. L. (2016). Analyzing Healthcare Big Data with Prediction for Future Health Condition. *IEEE Access*, *4*, 9786–9799. https://doi.org/10.1109/ACCESS.2016.2647619

Sesen, M. B., Kadir, T., Alcantara, R. B., Fox, J., & Brady, M. (2012). Survival Prediction and Treatment Recommendation with Bayesian Techniques in Lung Cancer. *AMIA Annual Symposium Proceedings*, *2012*, 838. https://pmc.ncbi.nlm.nih.gov/articles/PMC3540451/

Tanner, T., & Toivonen, H. (2010). Predicting and preventing student failure – using the k-nearest neighbour method to predict student performance in an online course environment. *International Journal of Learning Technology*, *5*(4), 356. https://doi.org/10.1504/IJLT.2010.038772

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 1–16. https://doi.org/10.1186/S12911-019-1004-8/FIGURES/12

Umamaheswaran, S. K., Kaur, G., Pankajam, A., Firos, A., Vashistha, P., Tripathi, V., & Mohammed, H. S. (2022). [Retracted] Empirical Analysis for Improving Food Quality Using Artificial Intelligence Technology for Enhancing Healthcare Sector. *Journal of Food Quality*, *2022*(1), 1447326. https://doi.org/10.1155/2022/1447326

Wayahdi, M. R. & Ruziq, F. (2022). KNN and XGBoost Algorithms for Lung Cancer Prediction. Journal of Science Technology (JoSTec), 4(1), 179–186. https://doi.org/10.55299/JOSTEC.V4I1.251

Wayahdi, M. R., & Ruziq, F. (2024). Designing an Used Goods Donation System to Reduce Waste Accumulation Using the WASPAS Method. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *8*(4), 2325–2334. https://doi.org/10.33395/SINKRON.V8I4.14115

Wormanns, D., Fiebich, M., Saidi, M., Diederich, S., & Heindel, W. (2002). Automatic detection of pulmonary

nodules at spiral CT: Clinical application of a computer-aided diagnosis system. *European Radiology*, *12*(5), 1052–1057. https://doi.org/10.1007/S003300101126/METRICS

Zhang, S. (2022). Challenges in KNN Classification. *IEEE Transactions on Knowledge and Data Engineering*, *34*(10), 4663–4675. https://doi.org/10.1109/TKDE.2021.304925

Zikos, D., & Delellis, N. (2018). CDSS-RM: A clinical decision support system reference model. *BMC Medical Research Methodology*, *18*(1), 1–14. https://doi.org/10.1186/S12874-018-0587-6/FIGURES/3

\