SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 1, February 2025, Pages 138-140
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i1.586

Classification of the Relationship Between Fruit Consumption and Diabetes Risk Using the Naïve Bayes Algorithm

Intan Nabila Hilma[1], Aditya Akbar Riadi[2], Ahmad Abdul Chamid[3]

Department of Informatics Engineering, Faculty of Engineering, Universitas Muria Kudus, Kudus, Indonesia

Author Email: nabilahlmaaa@gmail.com[1], aditya.akbar@umk.ac.id[2], abdul.chamid@umk.ac.id[3]

Abstract. Diabetes is a non-communicable disease whose prevalence continues to rise, where dietary patterns, including fruit intake, contribute to diabetes risk. This study classifies the relationship between fruit consumption and diabetes risk using the Naïve Bayes algorithm. The dataset consists of 400 synthetic records with 8 attributes. Data preprocessing included cleaning, normalization, and 5-Fold Cross Validation. Results show accuracy of 96.25%, precision 100%, recall 57%, and F1-score 0.73. The model was implemented into GlucoSense, a web-based system providing real-time diabetes risk predictions. This research proves that Naïve Bayes is effective for classifying diabetes risk based on fruit consumption patterns.

Keywords: classification, diabetes, fruit consumption, naïve bayes, glucosense

1 Introduction

Diabetes is a non-communicable disease that has become a serious health challenge in many countries, including Indonesia. This disease is characterized by high blood glucose levels due to problems in the production or function of the insulin hormone. The increasing prevalence of diabetes indicates that it not only affects quality of life but also imposes a significant economic burden on the national health system [1].

One of the factors affecting the risk of developing diabetes is dietary patterns. Consuming fruit has health benefits as it is rich in fiber, vitamins, and minerals that maintain metabolic balance. However, inappropriate fruit consumption patterns can have different effects on blood sugar levels [2].

With advances in information technology, data mining techniques have become one way to analyze large amounts of health data. Data mining enables the discovery of patterns and hidden information to support decision-making processes. One classification approach often applied in data mining is the Naïve Bayes algorithm, which excels in computational simplicity and efficiency [3].

Based on this issue, this research classifies the relationship between fruit consumption and diabetes risk using the Naïve Bayes algorithm and implements the results in a web-based prediction system named GlucoSense [4].

2 Literature Review

2.1 Related Research

Research on the utilization of Naïve Bayes for diabetes classification using the Pima Indians Diabetes dataset achieved 76.5% accuracy after data preprocessing, handling missing values, and normalization [1].

A comparative study of K-Nearest Neighbor, Naïve Bayes, and Decision Tree classification for predicting diabetes using Kaggle datasets showed Naïve Bayes achieved 77.5% accuracy, Decision Tree 82.4%, and K-NN 80.7% [2].

Classification of diabetes mellitus using the Naïve Bayes Classifier algorithm on 130 patient records from RS Dirgahayu Samarinda (2018–2021) achieved the highest accuracy of 84.6% [3].

Classification of obesity levels using Gradient Boosting Machine (GBM) on 2,111 data records with 16 attributes achieved 95% accuracy [4].

Analysis of early diabetes risk prediction using Naïve Bayes on the Early Stage Diabetes Risk Prediction dataset from the UCI Machine Learning Repository achieved 87.88% accuracy [5].

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 1, February 2025, Pages 138-140
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i1.586

2.2 Theoretical Framework

Classification is a technique in data mining and machine learning that aims to categorize data into certain classes based on its attributes. A model is built from labeled training data, then used to predict the class of new testing data.

Diabetes is a long-term metabolic disease characterized by high blood glucose levels caused by problems in insulin production or function. Risk factors include unhealthy dietary patterns, obesity, and lack of physical activity.

Naïve Bayes is a classification algorithm based on Bayes' Theorem, assuming each attribute is independent of one another. The Bayes' Theorem is expressed as:

$$P(C|X) = P(X|C) \times P(C) / P(X)$$

Where $P(C|X)$ is the posterior probability of class C given data X, $P(X|C)$ is the likelihood, $P(C)$ is the prior probability, and $P(X)$ is the total probability. The class with the highest posterior probability is selected as the classification result.

3 Research Methodology

This study employs a quantitative approach through the experimental method, focusing on numerical data analysis to objectively evaluate the performance of the classification model.

3.1 Data Collection

The dataset consists of 400 synthetic data records with 8 attributes: age, sex, BMI, blood glucose, blood pressure, family history of diabetes, diabetes diagnosis history, and frequency of apple consumption per week, with two class labels: Low Risk and High Risk.

3.2 Data Preprocessing

Data preprocessing includes cleaning missing values, normalizing attributes to a uniform scale, and checking class distribution. 5-Fold Cross Validation was applied: the data was divided into 5 equal folds, each serving as test data in rotation, improving model evaluation reliability.

3.3 Classification Using Naïve Bayes

The Naïve Bayes algorithm calculates the probability of data belonging to each class (Low Risk or High Risk) based on the prior probability and the likelihood of each attribute. The class with the highest posterior probability is selected as the final prediction.

3.4 Model Evaluation

Model performance was evaluated using metrics derived from the Confusion Matrix: Accuracy, Precision, Recall, and F1-Score.

3.5 System Development

The Prototyping method was used for web-based system development, allowing iterative refinement of the GlucoSense system based on user needs and model testing results.

4 Results and Discussion

4.1 Dataset Distribution

The dataset (N = 400) exhibits class imbalance, with Low Risk samples significantly outnumbering High Risk samples. This imbalance influenced the recall metric for the High Risk class.

Table 1. Confusion Matrix Results of Naïve Bayes Classification

Metric	High Risk	Low Risk
TP / TN	4	73
FP / FN	0	3

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 1, February 2025, Pages 138-140
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i1.586

4.2 Evaluation Results

Table 2. Naïve Bayes Model Evaluation Results

Metric	Value
Accuracy	96.25%
Precision	100%
Recall	57%
F1-Score	0.73

The Accuracy of 96.25% demonstrates strong overall classification performance. Precision of 100% confirms zero false positives for High Risk predictions. The Recall of 57% is affected by class imbalance, and the F1-Score of 0.73 reflects the balance between precision and recall.

4.3 GlucoSense Web System

GlucoSense successfully integrates the Naïve Bayes model into a web-based system, accepting user health input (age, BMI, blood glucose, blood pressure, fruit consumption frequency) and providing real-time diabetes risk predictions with detected risk factors and personalized health recommendations.

5 Conclusion

Naïve Bayes effectively classifies the relationship between fruit consumption and diabetes risk into Low Risk and High Risk categories with 96.25% accuracy. Precision of 100% confirms zero false positives for High Risk predictions, while recall of 57% is affected by class imbalance. The model was successfully integrated into GlucoSense, delivering real-time predictions. The system serves as an early diabetes risk detection tool but is not intended as a substitute for formal medical diagnosis.

6 Suggestions

Future research should use real datasets from health institutions. Techniques such as oversampling, undersampling, or SMOTE can address class imbalance. Comparing Naïve Bayes with Decision Tree, Random Forest, or Support Vector Machine is recommended. GlucoSense can be extended with health history tracking, personalized lifestyle recommendations, and a mobile application.

Acknowledgements

The authors thank Universitas Muria Kudus, particularly the Informatics Engineering Study Program, for supporting this research. Special thanks to the supervisors for their guidance throughout this work.

References

[1] Handayani, D., & Sari, R. (2021). Pemanfaatan Metode Naïve Bayes untuk Klasifikasi Diabetes Berdasarkan Data Medis. *Jurnal Informatika Kesehatan*, *8*(2), 45–53.
[2] Pratiwi, A., & Nugroho, B. (2022). Prediksi Penyakit Diabetes Berdasarkan Perbandingan Klasifikasi K-Nearest Neighbor, Naïve Bayes, dan Decision Tree. *Jurnal Ilmu Komputer dan Informasi*, *15*(1), 12–21.
[3] Rahmawati, S., & Wijaya, T. (2022). Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier. *Jurnal Teknologi Informasi*, *10*(3), 78–89.
[4] Kusuma, F., & Santoso, E. (2023). Klasifikasi Tingkat Obesitas Menggunakan Metode Gradient Boosting Machine (GBM) dan Confusion Matrix. *Jurnal Data Science Indonesia*, *4*(1), 22–31.
[5] Wibowo, A., & Putri, I. (2023). Analisis Prediksi Risiko Diabetes Tahap Awal Menggunakan Algoritma Naïve Bayes. *Jurnal Sistem Informasi Kesehatan*, *6*(2), 55–64.