Medical Record-Based Prediction of Type 2 Diabetes Mellitus Risk Using the Naïve Bayes Algorithm

Mochammad Choirur Roziqin[1]*, Nabila Ersa Wulandari[1], Bakhtiyar Hadi Prakoso[1], Dony Setiawan Hendyca Putra[1], Muhammad Ifantara Putra[1], Gamasiano Alfiansyah[1]

[1]Health Information Management Study Program, Department of Health, Politeknik Negeri Jember, Indonesia

Email: irul@polije.ac.id*

Submitted: 25 May 2026 Reviewed: 28 July 2026 Accepted: 11 August 2026

ABSTRACT

Background: Type 2 diabetes mellitus is a non-communicable disease with a high prevalence and the potential to cause serious complications if not properly managed. Identifying risk factors is an essential step in the prevention and control of this disease. **Objective:** This study aimed to develop and evaluate a medical record-based prediction model for type 2 diabetes mellitus risk using the Naïve Bayes algorithm and to identify the most influential risk factors among hospitalized patients at Bhayangkara Bondowoso Hospital in 2024. **Methods:** A descriptive quantitative research design was employed, with data processing conducted using the Naïve Bayes algorithm and RapidMiner software. The sampling technique used was total sampling, resulting in 630 medical records, consisting of 315 patients diagnosed with type 2 diabetes mellitus and 315 non-type 2 diabetes mellitus patients. The analyzed variables included age, sex, body mass index (BMI), hypertension, smoking history, cardiovascular disease history, and family history. **Results:** The results indicated that the most influential risk factors for type 2 diabetes mellitus were age ≥45 years, obesity-level BMI, and a history of hypertension. **Conclusion:** The Confusion Matrix evaluation with a 95%:5% split ratio produced an accuracy of 90.62%, a precision of 100%, and a recall of 81.25%. **Suggestion:** Hospitals are encouraged to strengthen health promotion programs regarding type 2 diabetes mellitus risk factors and involve patients' families in educational interventions to help prevent disease-related complications.

Keywords: Machine Learning; Medical Records; Naïve Bayes Algorithm; Risk Factors; Type 2 Diabetes Mellitus

INTRODUCTION

Medical records are documents containing patient identification data, examination results, treatments, medical procedures, and other healthcare services provided to patients (1). Medical records serve as essential documents for documenting patients' health histories and supporting clinical decision-making in healthcare services. Health Information Management (HIM) professionals possess competencies in health statistics, basic epidemiology, biomedical sciences, as well as disease and clinical procedure classification and coding (2). These competencies enable HIM professionals to manage data accurately throughout the processes of data collection, processing, and verification, thereby supporting decision-making in healthcare facilities. Structured medical record data provide comprehensive information regarding patients' conditions, including disease history, examination results, and medical interventions. In cases of Diabetes Mellitus (DM), such data assist healthcare professionals in determining accurate diagnoses and appropriate treatments, thereby improving the quality of healthcare services and the effectiveness of patient management.

Diabetes Mellitus (DM) is a complex chronic disease that requires continuous medical care, including strategies for blood glucose control and the prevention of complications (3). DM may lead to various complications affecting multiple organs, including cardiovascular disease, stroke, kidney failure, and diabetic foot infections that may result in amputation (4–9). According to the World Health Organization (WHO), approximately 422 million people worldwide are living with DM, with a prevalence of around 8.5% among the adult population. In Southeast Asia, the number of individuals with DM has reached approximately 90 million, ranking second only to the Western Pacific region, which has approximately 206 million cases (10).

According to the Indonesian Health Profile 2021, the prevalence of non-communicable diseases, including DM, has shown an increasing trend (11). Data from the 2018 Basic Health Research Survey (Riskesdas) indicated that DM was the third leading cause of death in Indonesia, accounting for 6.7% of total mortality, following stroke

(21.1%) and heart disease (12.9%) (12). Furthermore, the prevalence of DM increased from 1.5% in 2013 to 2.0% in 2018. The International Diabetes Federation (IDF) reported that the prevalence of DM among individuals aged 20–79 years in Indonesia reached 10.6% in 2021, placing Indonesia fifth globally in terms of the number of people living with diabetes (13). Based on data from the East Java Provincial Health Office in 2023, the number of DM cases has shown an increasing trend over the past three years, rising from 785,983 cases in 2020 to 854,454 cases in 2023 (14). Meanwhile, data from the Bondowoso District Health Office indicated that the number of DM cases reached 12,717 in 2022, decreased to 10,158 in 2023, and increased again to 12,717 in 2024. This trend demonstrates that DM remains a significant public health concern requiring special attention in Bondowoso Regency.

At Bhayangkara Bondowoso Hospital, Type 2 Diabetes Mellitus has consistently ranked among the ten most common inpatient diseases over the last three years. In 2024, it ranked second, with the number of cases increasing from 43 cases in 2022 to 120 cases in 2023 and 340 cases in 2024. The high incidence of Type 2 DM highlights the need for early identification of risk factors through the utilization of medical record data. Risk factors for Type 2 DM can be categorized into non-modifiable factors, such as age, sex, and family history, and modifiable factors, including body mass index (BMI), hypertension, smoking history, and physical activity. The risk of DM increases among individuals aged ≥45 years due to a decline in pancreatic β-cell function and insulin sensitivity (15). Furthermore, according to the Indonesian Ministry of Health (2019), as cited in Purba and Wahyu (2025), the prevalence of diabetes mellitus is higher among women than among men (16). Elevated BMI is associated with insulin resistance and an increased risk of Type 2 DM, while hypertension is recognized as a contributing factor to insulin resistance (17,18).

Data mining is the process of extracting meaningful patterns and information from large datasets to support faster and more accurate decision-making (19). One of the most widely used classification techniques in data mining is the Naïve Bayes algorithm, a probabilistic classification method that assumes independence among predictor attributes based on the class variable (20). A study conducted by Khasanah (2022) applied the Naïve Bayes Classifier algorithm to classify DM patients at Dirgahayu Hospital, Samarinda, and achieved an accuracy rate of 92% (21). Similarly, Al'Ariq *et al.* (2022) demonstrated that the Naïve Bayes method achieved an accuracy of 85.71% in classifying DM cases at Hasyim Asy'ari Hospital, Jombang (22).

Although previous studies have demonstrated that the Naïve Bayes algorithm provides satisfactory classification performance for diabetes prediction, several limitations remain. Most existing studies primarily focused on classification accuracy without comprehensively identifying the contribution of individual clinical risk factors. In addition, previous research generally evaluated the model using a single training–testing data split and rarely compared multiple partitioning scenarios to determine the optimal model performance. Furthermore, studies utilizing real-world inpatient medical record data from Indonesian hospitals remain limited, particularly in regional hospitals. Therefore, further research is needed to evaluate the performance of the Naïve Bayes algorithm using actual clinical data while simultaneously identifying the most influential risk factors associated with Type 2 Diabetes Mellitus.

Based on the aforementioned background, an analysis of Type 2 DM risk factors using medical record data is needed to support early disease screening and prevention efforts. Therefore, this study aimed to develop and evaluate a medical record-based prediction model for type 2 diabetes mellitus risk using the Naïve Bayes algorithm and to identify the most influential risk factors among hospitalized patients at Bhayangkara Bondowoso Hospital in 2024.

METHODS

This study followed a structured research workflow to develop and evaluate a medical record-based prediction model for Type 2 Diabetes Mellitus (T2DM) risk using the Naïve Bayes algorithm and to identify the most influential risk factors among hospitalized patients at Bhayangkara Bondowoso Hospital in 2024. The workflow comprised problem identification, predictor variable selection, data collection, preprocessing, data transformation, dataset partitioning, model development, validation, performance evaluation, and identification of significant risk factors. The overall research workflow is illustrated in Figure 1.

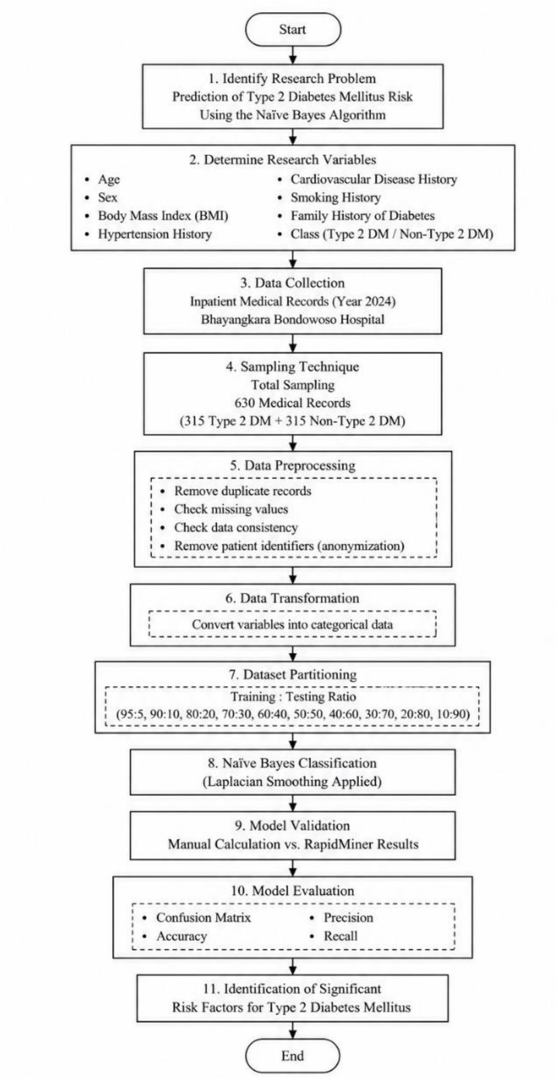

Figure 1. Research workflow for predicting Type 2 Diabetes Mellitus risk using the Naïve Bayes algorithm

Figure 1 illustrates the overall research workflow used to predict the risk of Type 2 Diabetes Mellitus (T2DM) based on inpatient medical records using the Naïve Bayes algorithm. The study began with the identification of the research problem, followed by the selection of seven predictor variables: age, sex, body mass index (BMI), hypertension, cardiovascular disease history, smoking history, and family history of diabetes. Subsequently, inpatient medical record data from Bhayangkara Bondowoso Hospital. The collected data then underwent preprocessing to remove duplicate records, verify data completeness and consistency, and anonymize patient identifiers to ensure data quality and confidentiality.

The cleaned data were transformed into categorical variables before being partitioned into training and testing datasets using several split ratios. The classification model was developed using the Naïve Bayes algorithm with Laplacian smoothing to prevent zero-probability values. The implementation results were validated by comparing manual calculations with the RapidMiner output. Finally, the model performance was evaluated using a confusion matrix based on accuracy, precision, and recall values to determine the optimal classification model and identify the most significant risk factors associated with Type 2 Diabetes Mellitus.

This study employed a quantitative approach using data obtained from inpatient medical records. The data were analyzed using the Naïve Bayes method with the assistance of RapidMiner software, and model performance was evaluated based on accuracy, precision, and recall values derived from the Confusion Matrix. The study population consisted of all hospitalized patients diagnosed with Type 2 Diabetes Mellitus (T2DM) and non-T2DM patients at Bhayangkara Bondowoso Hospital in 2024.

A total sampling technique was applied, resulting in 630 medical records that met the inclusion criteria, comprising 315 records of T2DM patients and 315 records of non-T2DM patients. The inclusion criteria consisted of inpatient medical records of patients diagnosed with T2DM or non-T2DM in 2024 with complete data for all study variables. The exclusion criteria included incomplete medical records, records with primary diagnoses other than T2DM or non-T2DM, and records containing illegible information.

The study variables included sex (male/female), age (<45 years or ≥45 years), body mass index (BMI), hypertension status, cardiovascular disease history, smoking history, and family history of diabetes mellitus. BMI classification followed the Indonesian Society of Endocrinology (PERKENI, 2021) criteria: underweight (<18.5 kg/m²), normal (18.5–22.9 kg/m²), at risk (23.0–24.9 kg/m²), Obesity Class I (25.0–29.9 kg/m²), and Obesity Class II (≥30.0 kg/m²). Hypertension was defined as blood pressure ≥140/90 mmHg, whereas blood pressure <140/90 mmHg was categorized as normal. Cardiovascular disease history, smoking history, and family history of diabetes were categorized as either present or absent.

Secondary data were obtained from inpatient medical records for the year 2024. Data collection was conducted through document observation using a structured checklist form. Data analysis was performed following the Knowledge Discovery in Databases (KDD) framework, which consisted of six stages: (1) data collection through medical record observation; (2) data preprocessing and cleaning using RapidMiner to address missing values, inconsistencies, and duplicate records; (3) data transformation; (4) dataset partitioning into training and testing sets with split ratios ranging from 95:5 to 10:90; (5) data mining using the Naïve Bayes algorithm; and (6) model evaluation using the Confusion Matrix.

Naïve Bayes calculations were initially performed manually and subsequently validated using RapidMiner software. Laplacian smoothing was applied to prevent zero-probability values during the classification process. Model performance was evaluated using the following metrics:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \times 100\%$$

$$\text{Precision} = TP / (TP + FP) \times 100\%$$

$$\text{Recall} = TP / (TP + FN) \times 100\%$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

RESULTS

Distribution of the Sex Variable

Table 1. Distribution of Participants by Gender and Type 2 Diabetes Mellitus Status

Gender	Type 2 Diabetes Mellitus, n (%)	Non-Type 2 Diabetes Mellitus, n (%)
Male	123 (39.05)	150 (47.62)
Female	192 (60.95)	165 (52.38)
Total	**315 (100.00)**	**315 (100.00)**

Based on Table 1, Type 2 Diabetes Mellitus (T2DM) cases were more frequently observed among female patients, accounting for 60.95% of the total cases. This finding may be attributed to the higher tendency of women to experience increased body mass index (BMI), which is influenced by biological factors such as premenstrual syndrome and fat accumulation during the postmenopausal period. These hormonal changes can accelerate fat deposition, thereby increasing the risk of developing T2DM (23). This finding is consistent with previous studies demonstrating that the prevalence of diabetes mellitus is higher among women than among men (24–26), suggesting that female sex may be associated with an increased susceptibility to Type 2 Diabetes Mellitus.

Distribution of the Age Variable

Table 2. Distribution of Participants by Age Group and Type 2 Diabetes Mellitus Status

Age Group	Type 2 Diabetes Mellitus, n (%)	Non-Type 2 Diabetes Mellitus, n (%)
< 45 years	134 (42.54)	167 (53.02)
≥ 45 years	181 (57.46)	148 (46.98)
Total	**315 (100.00)**	**315 (100.00)**

Based on Table 2, patients aged ≥45 years accounted for 57.46% of all Type 2 Diabetes Mellitus (T2DM) cases. Individuals aged ≥45 years have been reported to be approximately nine times more likely to develop T2DM than those younger than 45 years. This increased risk is associated with age-related declines in the body's ability to metabolize glucose effectively, which may lead to glucose intolerance and a higher likelihood of developing diabetes (27,28). These findings suggest that advancing age remains an important risk factor for T2DM and should be considered in early screening and prevention strategies.

Distribution of the Hypertension Variable

Table 3. Distribution of Participants by Hypertension Status and Type 2 Diabetes Mellitus Status

Hypertension Status	Type 2 Diabetes Mellitus, n (%)	Non-Type 2 Diabetes Mellitus, n (%)
Yes	74 (23.49)	1 (0.32)
No	241 (76.51)	314 (99.68)
Total	**315 (100.00)**	**315 (100.00)**

Based on Table 3, 23.49% of patients with Type 2 Diabetes Mellitus (T2DM) had hypertension, which was substantially higher than the proportion observed among non-T2DM patients (0.32%). Hypertension is recognized as a major risk factor for T2DM because elevated blood pressure may contribute to insulin resistance, a condition in which body cells respond inadequately to insulin. This impairment disrupts glucose uptake and may lead to increased blood glucose levels (18,29). Furthermore, hypertension has been reported to occur approximately 1.5-2 times more frequently among patients with T2DM than among individuals without diabetes (27). These findings highlight the close relationship between hypertension and T2DM and emphasize the importance of blood pressure management as part of diabetes prevention and control strategies.

Distribution of the Body Mass Index (BMI) Variable

Table 4. Distribution of Participants by BMI Category and Type 2 Diabetes Mellitus Status

BMI Category	Type 2 Diabetes Mellitus, n (%)	Non-Type 2 Diabetes Mellitus, n (%)
Underweight (<18.5)	0 (0.00)	0 (0.00)
Normal (18.5–22.9)	80 (25.40)	295 (93.65)
Overweight (23–24.9)	73 (23.17)	20 (5.56)
Obese I (25–29.9)	159 (50.48)	0 (0.00)
Obese II (≥30)	3 (0.95)	0 (0.00)
Total	**315 (100.00)**	**315 (100.00)**

Based on Table 4, 50.48% of patients with Type 2 Diabetes Mellitus (T2DM) were classified as having Class I obesity according to their body mass index (BMI). This finding indicates that obesity was a highly prevalent risk factor within the study population. Obesity is strongly associated with insulin resistance due to excessive fat accumulation, which can impair insulin function in regulating blood glucose levels (23). In addition, unhealthy lifestyle behaviors, including the consumption of high-fat and low-fiber diets as well as insufficient physical activity, contribute to increased BMI, thereby elevating the risk of developing T2DM (30–32). These findings underscore the importance of maintaining a healthy body weight through balanced nutrition and regular physical activity as part of effective strategies for preventing and controlling T2DM.

Distribution of Cardiovascular History, Smoking History, and Family History Variables

Table 5. Distribution of Participants by Cardiovascular History, Smoking History, and Family History According to Type 2 Diabetes Mellitus Status

Variable	Category	Type 2 Diabetes Mellitus, n (%)	Non-Type 2 Diabetes Mellitus, n (%)
Cardiovascular History	Yes	5 (1.59)	0 (0.00)
	No	310 (98.41)	315 (100.00)
Smoking History	Yes	38 (12.06)	22 (6.98)
	No	277 (87.94)	293 (93.02)
Family History	Yes	8 (2.54)	0 (0.00)
	No	307 (97.46)	315 (100.00)
Total		**315 (100.00)**	**315 (100.00)**

Based on Table 5, five patients with Type 2 Diabetes Mellitus (T2DM) (1.59%) had a history of cardiovascular disease. Although the proportion was relatively small, a history of cardiovascular disease may still contribute to the risk of developing T2DM. Patients with cardiovascular conditions are often advised to limit strenuous physical activity, which may reduce metabolic activity. Elevated blood glucose levels occur when glucose metabolism is impaired due to insulin resistance or reduced insulin secretion. Under these conditions, glucose is unable to enter cells efficiently and consequently accumulates in the bloodstream, thereby increasing the risk of developing Type 2 Diabetes Mellitus (33,34).

A history of smoking was identified in 12.06% of patients with T2DM. Smoking is associated with vasoconstriction and may contribute to the development of insulin resistance. A study by Sastrawan *et al.* (2023) reported that patients with T2DM who had a history of smoking were 3.33 times more likely to experience diabetes-related complications (35). Among individuals with diabetes, smoking has been shown to significantly increase the risk of kidney disease, diabetic retinopathy, and circulatory disorders.

A family history of diabetes mellitus was identified in 2.54% of patients with Type 2 Diabetes Mellitus (T2DM), indicating that genetic factors contribute to individual susceptibility to the disease. Individuals with a family history of diabetes are at a higher risk of developing T2DM due to genetic predispositions that may affect pancreatic β-cell function and insulin sensitivity. A study conducted by Irayani *et al.* (2024) demonstrated that a family history of diabetes is a significant risk factor associated with the occurrence of T2DM, increasing the likelihood of developing the disease by up to 14.916 times compared with individuals without a family history of diabetes (36). Family history is considered a non-modifiable risk factor.

A family history of diabetes mellitus was identified in 2.54% of patients with Type 2 Diabetes Mellitus (T2DM), indicating that genetic factors contribute to individual susceptibility to the disease. Individuals with a family history of diabetes are known to have a higher risk of developing T2DM compared with those without such a history. Genetic predisposition may influence pancreatic β-cell function, insulin sensitivity, and glucose metabolism, thereby increasing susceptibility to T2DM.

A study conducted by Irayani *et al.* (2024) demonstrated that a family history of diabetes is a significant risk factor associated with the occurrence of T2DM, increasing the likelihood of developing the disease by up to 14.916 times compared with individuals without a family history of diabetes (36). This finding is further supported by Paramita and Lestari (2019), who reported that individuals with a family history of T2DM tend to have higher blood glucose levels than those without a family history of the disease (37). These findings suggest that hereditary factors play an important role in the development of T2DM and should be considered in early screening and prevention strategies, particularly among individuals with a positive family history of diabetes.

Data Preprocessing

The preprocessing stage was conducted to address potential data quality issues, including missing values, data inconsistencies, and duplicate records. This process was performed using RapidMiner software. The preprocessing results indicated that no missing values were identified across all attributes, as evidenced by a missing value count of zero for each variable. Furthermore, no data inconsistencies or duplicate records were detected. All variables were classified as binominal attributes, except for the Body Mass Index (BMI) variable, which was categorized as a discrete attribute with more than two categories. During this stage, patient medical record numbers were removed in accordance with the Indonesian Ministry of Health Regulation to ensure the confidentiality and privacy of patient information (1).

Classification Using the Naïve Bayes Algorithm

The classification process was performed using two approaches: manual calculation and implementation through RapidMiner software, both utilizing the same dataset. For the manual calculation using a 95%:5% split ratio, the dataset was divided into 598 training records, consisting of 299 T2DM cases and 299 non-T2DM cases, and 32 testing records, comprising 16 T2DM cases and 16 non-T2DM cases.

The prior probability calculations yielded identical values for both classes, with P(T2DM) = 299/598 = 0.50 and P(Non-T2DM) = 299/598 = 0.50. Likelihood values were subsequently calculated for each attribute within each class using the Laplacian smoothing technique to prevent zero-probability estimates. For example, in the T2DM class, the likelihood values were calculated as follows:

P(Female | T2DM) = (1 + 182) / (299 + 2) = 0.608

P(Age ≥45 Years | T2DM) = (1 + 172) / (299 + 2) = 0.574

P(Class I Obesity BMI | T2DM) = (1 + 154) / (299 + 5) = 0.509

Posterior probabilities were then obtained by multiplying the likelihood values of all predictor attributes by the corresponding prior probability for each class. The final classification was determined based on the class with the highest posterior probability. The results obtained from the manual calculations were fully consistent with those generated by RapidMiner. No discrepancies were observed between the predicted class labels produced by the two approaches, indicating complete agreement in the classification outcomes. This consistency confirms the reliability of the implementation process and demonstrates that the Naïve Bayes model was correctly applied within the RapidMiner environment.

Table 6. Prior Probabilities of Class Labels

Class	Prior Probability
Type 2 Diabetes Mellitus	0.50
Non-Type 2 Diabetes Mellitus	0.50

Table 6 shows the prior probability distribution of the target classes. The prior probabilities for the T2DM and non-T2DM classes were both 0.50, reflecting a balanced dataset with an equal representation of positive and negative cases. Such a distribution minimizes class imbalance and contributes to a more reliable evaluation of classification performance.

Confusion Matrix Analysis

To assess the classification performance, a confusion matrix analysis was conducted across multiple training–testing data split ratios. The 95:5 split ratio yielded the most favorable results and was therefore selected as the optimal model configuration.

Table 7. Confusion Matrix Results for the 95:5 Data Split Ratio

Predicted Class	Actual Type 2 Diabetes Mellitus (TP/FN)	Actual Non-Type 2 Diabetes Mellitus (FP/TN)	Precision
Predicted Type 2 Diabetes Mellitus	13 (TP)	0 (FP)	100.00%
Predicted Non-Type 2 Diabetes Mellitus	3 (FN)	16 (TN)	84.21%
Recall	81.25%	100.00%	Accuracy: 90.62%

Based on the confusion matrix presented in Table 7, the model attained an accuracy of 90.62%, a precision of 100.00%, and a recall of 81.25%. The perfect precision score indicates that all observations classified as T2DM were true positive cases, with no occurrence of false positive predictions. Nevertheless, the recall value of 81.25% reveals that three actual T2DM cases were incorrectly classified as non-T2DM, resulting in three false negatives. This finding suggests that while the model demonstrated outstanding reliability in identifying positive predictions, a small proportion of T2DM cases remained undetected. Despite this limitation, the overall classification performance was considered highly satisfactory, as reflected by the high accuracy and precision values. A comprehensive comparison of model performance across different data split ratios is provided in Table 8.

Table 8. Comparison of Classification Performance Across Different Data Split Ratios

Training (%)	Testing (%)	Akurasi (%)	Presisi (%)	Recall (%)
95	5	90,62	100,00	81,25
90	10	90,32	96,30	83,87
80	20	88,10	91,38	84,13
75	25	90,51	93,24	87,34
70	30	90,34	94,19	86,17
66	34	88,32	92,71	83,18
60	40	88,24	92,52	83,19
50	50	87,30	92,09	81,53
40	60	89,15	93,02	84,66
30	70	88,64	93,37	83,18
20	80	88,49	93,30	82,94
10	90	87,28	92,37	81,27

Based on Table 8, the highest accuracy value was achieved with the 95:5 training–testing split ratio, reaching 90.62%. This finding is consistent with the study conducted by Althnian *et al.* (2021), which reported that increasing the proportion of training data enables the model to learn data characteristics more effectively, thereby improving classification accuracy (38). This finding is consistent with the study conducted by Prasetyo *et al.* (2024), which demonstrated that the proportion of training and testing data significantly influences the performance of classification models. A larger proportion of training data enables the model to learn the underlying characteristics and patterns of the dataset more effectively, thereby improving its classification capability. Consequently, the 95:5 training-testing split ratio employed in this study achieved higher accuracy than the other split ratios, as the model was provided with a greater amount of information during the training process (39).

DISCUSSION

The classification results generated by the Naïve Bayes algorithm indicated that most records characterized by risk factors such as overweight or obesity, hypertension, and a history of smoking were classified into the T2DM group. In contrast, records with normal BMI and without hypertension were predominantly classified into the non-T2DM group. These findings are consistent with the study by Salasa *et al.* (2019), which described T2DM as a multifactorial disease resulting from the interaction of genetic factors, lifestyle behaviors, and metabolic conditions (40).

Hypertension was identified as one of the factors most strongly associated with T2DM. According to Susilawati and Rahmawati (2021), many patients with hypertension eventually develop diabetes, and conversely, diabetes patients frequently present with hypertension (27). This relationship is attributed to common underlying mechanisms, including insulin resistance, endothelial dysfunction, and impaired lipid metabolism. Obesity also plays a critical role in the development of insulin resistance. Excessive fat accumulation, particularly in the abdominal region, can impair insulin function in regulating blood glucose levels and is considered a major contributor to the onset of T2DM (23).

Furthermore, smoking may accelerate the progression of atherosclerosis and worsen vascular function, particularly among individuals who already possess other risk factors such as obesity or hypertension (35). Although smoking is not considered a direct cause of T2DM, it may contribute to the development of long-term metabolic complications. Therefore, the classification results provide strong evidence for developing T2DM prevention strategies, with primary emphasis on weight management, blood pressure control, smoking cessation, and the promotion of healthy lifestyle behaviors.

The classification outcomes were also validated by healthcare professionals at Bhayangkara Bondowoso Hospital, who confirmed that the results were consistent with clinical conditions observed in practice. These findings suggest that the Naïve Bayes algorithm has considerable potential as a decision-support tool in clinical practice, particularly for the early screening of Type 2 Diabetes Mellitus (T2DM) risk. The results of this study are consistent with previous research, which demonstrated that the Naïve Bayes algorithm provides reliable classification performance in diabetes prediction and early detection applications (41). The ability of the algorithm to accurately classify patients based on multiple risk factors highlights its potential contribution to supporting healthcare professionals in identifying individuals at high risk of developing T2DM and implementing timely preventive interventions.

Overall, the Naïve Bayes algorithm demonstrated effective performance in classifying T2DM risk factors, achieving consistently high accuracy across different training–testing split ratios, ranging from 87.28% to 90.62%. These findings indicate that the method is both valid and applicable as a medical record–based screening tool for identifying individuals at risk of T2DM and supporting more efficient and evidence-based clinical decision-making.

CONCLUSION

This study successfully developed and evaluated a medical record-based prediction model for Type 2 Diabetes Mellitus (T2DM) risk using the Naïve Bayes algorithm. The proposed model demonstrated excellent predictive performance, achieving an accuracy of 90.62%, precision of 100%, recall of 81.25%, and an AUC of 95.5% using the optimal training–testing data partition. Among the predictor variables, Body Mass Index (BMI), hypertension history, and family history of diabetes were identified as the most influential factors in predicting T2DM risk. These findings demonstrate that the Naïve Bayes algorithm is an effective and reliable approach for predicting T2DM risk using routinely collected inpatient medical records. The proposed model has the potential to support early risk stratification and clinical decision-making, thereby contributing to timely preventive interventions and improved diabetes management in healthcare settings.

This study has several limitations. First, the dataset was obtained from a single hospital, which may limit the generalizability of the findings to other healthcare settings or populations. Second, the prediction model was developed using a limited number of clinical variables available in the medical records and did not include lifestyle, laboratory, or genetic factors that may also influence the risk of Type 2 Diabetes Mellitus. Finally, the study evaluated only the Naïve Bayes algorithm without comparing its performance with other machine learning techniques. Future studies are recommended to compare the performance of Naïve Bayes with other machine learning algorithms, such as Decision Tree, Support Vector Machine (SVM), and Random Forest. Furthermore, the findings of this study may serve as a foundation for developing an integrated early detection system for T2DM to support preventive healthcare strategies and improve clinical decision-making.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Bhayangkara Bondowoso Hospital for granting permission to conduct this study and for providing the necessary facilities and support. The authors also extend their appreciation to Politeknik Negeri Jember for its academic support. Finally, the authors would like to thank all individuals and parties who contributed to the successful completion of this research.

REFERENCES

1. Kementerian Kesehatan Republik Indonesia. Peraturan Menteri Kesehatan Nomor 24 Tahun 2022 tentang Rekam Medis. Jakarta: Kementerian Kesehatan Republik Indonesia; 2022.
2. Kementerian Kesehatan Republik Indonesia. Keputusan Menteri Kesehatan (KMK) No. HK.01.07/Menkes/312/2020, standar profesi Perekam Medis dan Informasi Kesehatan (PMIK). Jakarta: Kementerian Kesehatan Republik Indonesia; 2020.
3. American Diabetes Association. Introduction: Standards of Medical Care in Diabetes—2022. Diabetes Care [Internet]. 2022 Jan 1;45(Supplement_1):S1–2. Available from: https://diabetesjournals.org/care/article/45/Supplement_1/S1/138921/Introduction-Standards-of-Medical-Care-in-Diabetes
4. Widiyanti S, Aini DN. Penerapan Pemberian Ekstrak Kayu Manis Terhadap Penurunan Kadar Gula Darah Pada Penderita Diabetes Melitus Di Kelurahan Gemah Semarang. In: The 2nd Widya Husada Nursing Conference. 2022. p. 86–90.
5. Boulton AJM, Armstrong DG, Kirsner RS, Attinger CE, Lavery LA, Lipsky BA, et al. Diagnosis and Management of Diabetic Foot Complications. American Diabetes Association, Inc; 2018.
6. Rahmah F, Ardiansar AM, Bamatraf A. Kaki Diabetik: Pendekatan Diagnosis Sebagai Komplikasi Angiopati Pada Diabetes Melitus. PREPOTIF J Kesehat Masy. 2025;9(2):3245–55.
7. Mutu DA, Yuda HT. Studi Kasus Penerapan Senam Kaki pada Pasien Diabetes Mellitus Desa Giwangretno Kabupaten Kebumen. J Keperawatan Malang. 2019;4(1):42–8.
8. Sumarni N, Rosidin U, Shalahuddin I, Witdiawati. Senam Kaki Diabetes Dalam Upaya Mencegah Ulkus Diabetik pada Lansia di RW 19 Kelurahan Kota Wetan Garut. J Kreat Pengabdi Kpd Masy. 2025;8(4):1941–53.
9. Sitompul Y, Budiman, Soebardi S, Abdullah M. Profil Pasien Kaki Diabetes yang Menjalani Reamputasi di Rumah Sakit Cipto Mangunkusumo Tahun 2008-2012. J Penyakit Dalam Indones. 2015;2(1):9–14.
10. Lasmawati E, Putri DF, Nuryani DD. Analisis faktor yang berhubungan dengan kejadian diabetes melitus tipe 2. Holistik J Kesehat [Internet]. 2023 Jul 20;17(4):334–44. Available from: https://ejurnalmalahayati.ac.id/index.php/holistik/article/view/11190
11. Kementerian Kesehatan Republik Indonesia. Profil Kesehatan Indonesia 2020. Jakarta: Kementerian Kesehatan Republik Indonesia; 2020.
12. Kementerian Kesehatan Republik Indonesia. Riset Kesehatan Dasar 2018. Jakarta; 2013.
13. International Diabetes Federation. IDF Diabetes Atlas 10th Edition 2021. International Diabetes Federation; 2021.
14. Dinas Kesehatan Provinsi Jawa Timur. Profil Kesehatan Provinsi Jawa Timur Tahun 2023. Surabaya: Dinas Kesehatan Provinsi Jawa Timur; 2023.
15. Nuraisyah F. Faktor Risiko Diabetes Mellitus Tipe 2. J Kebidanan dan Keperawatan Aisyiyah [Internet]. 2018 Feb 13;13(2):120–7. Available from: https://ejournal.unisayogya.ac.id/index.php/jkk/article/view/395
16. Purba ML, Wahyu A. Hubungan Jenis Kelamin dan Pola Makan dengan Kejadian Diabetes Melitus Tipe 2 Pada Usia Produktif di Wilayah Kerja Puskesmas Sering Medan. Excell Midwifery J. 2025;7(1):227–32.
17. Suwinawati E, Ardiani H, Ratnawati R. Hubungan Obesitas Dengan Kejadian Diabetes Melitus Tipe 2 Di Posbindu PTM Puskesmas Kendal Kabupaten Ngawi. J Heal Sci Prev [Internet]. 2020 Aug 26;4(2):79–84. Available from: http://jurnalfpk.uinsa.ac.id/index.php/jhsp/article/view/388
18. Nurhayati E, Putri NJY, Slamet S. Analisis Tekanan Darah pada Penderita Diabetes Melitus di Puskesmas Jungkat, Kalimantan Barat. J Lab Khatulistiwa. 2024;8(1):192–8.
19. Jessfry V, Siddik M. Penerapan Data Mining Menggunakan Algortima Apriori Dalam Membangun Sistem Persediaan Barang. JOISIE (Journal Inf Syst Informatics Eng. 2024;8(1):187–99.

20. Martantoh E, Yanih N. Implementasi Metode Naïve Bayes Untuk Klasifikasi Karakteristik Kepribadiaan Siswa Di Sekolah MTS Darussa'adah Menggunakan Php Mysql. J Teknol Sist Inf [Internet]. 2022 Sep 14;3(2):166–75. Available from: https://jurnal.mdp.ac.id/index.php/jtsi/article/view/2896
21. Khasanah LU, Nasution YN, Amijaya FDT. Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier. BASIS J Ilm Mat. 2022;1(1).
22. Qois Al'Ariq, Permadi GS, Mashuri C, Andriani A. Implementasi Naive Bayes dalam Memprediksi Penyakit Diabetes Mellitus. Inov J Ilm Inov Teknol Inf [Internet]. 2024 Sep 21;9(1):144–53. Available from: https://ejournal.unhasy.ac.id/index.php/inovate/article/view/7267
23. Rohmatulloh VR, Riskiyah R, Pardjianto B, Kinasih LS. Hubungan Usia dan Jenis Kelamin Terhadap Angka Kejadian Diabetes Melitus Tipe 2 Berdasarkan 4 Kriteria Diagnosis Di Poliklinik Penyakit Dalam RSUD Karsa Husada Kota Batu. PREPOTIF J Kesehat Masy. 2024;8(1):2528–43.
24. Arania R, Triwahyuni T, Esfandiari F, Nugraha FR. Hubungan Antara Usia, Jenis Kelamin, dan Tingkat Pendidikan dengan Kejadian Diabetes Mellitus di Klinik Mardi Waluyo Lampung Tengah. J Med Malahayati. 2021;5(3):146–53.
25. Nora P, Gultom N, Harahap F, Edi S, Sipahutar H. Hubungan Antara Jenis Kelamin dan Usia pada Penyakit Diabetes Melitus di Puskemas Kota Medan Tahun 2024-2025. J BIOSHELL J Pendidik Biol Biol Pendidik IPA. 2025;14(1):142–50.
26. Imelda S. Faktor-Faktor Yang Mempengaruhi Terjadinya diabetes Melitus di Puskesmas Harapan Raya Tahun 2018. Sci J. 2019;8(1):28–39.
27. Susilawati, Rahmawati R. Hubungan Usia, Jenis Kelamin dan Hipertensi dengan Kejadian Diabetes Mellitus Tipe 2 di Puskesmas Tugu Kecamatan Cimanggis Kota Depok. ARKESMAS. 2021;6(1):15–22.
28. Haryono DA, Arifin S, Shinta HE, Widodo T, Yuliani NNS. Hubungan obesitas dan aktivitas fisik dengan kejadian diabetes melitus tipe II pada usia > 40 tahun di wilayah kerja Puskesmas Bukit Hindu. Barigas J Ris Mhs [Internet]. 2023 Aug 29;1(2):53–60. Available from: https://e-journal.upr.ac.id/index.php/medica/article/view/7987
29. Sabrini AM, Febrianty F, Shafira NNA. Karakteristik Pasien DM Tipe 2 Dengan Hipertensi di Poliklinik Penyakit Dalam RSUD Raden Mattaher Jambi Tahun 2016-2019. J Med Stud [Internet]. 2023 Jan 6;2(2):72–80. Available from: https://online-journal.unja.ac.id/joms/article/view/23254
30. Zulkarnaini A, Mahatma G, Puspita D, Vani AT, Abdullah D. Aktivitas Fisik, Pola Makan, dan Konsumsi Makanan Glikemik Tinggi Meningkatkan Risiko Kejadian Diabetes Mellitus Tipe 2. J Kesehat Metro Sai Wawai. 2022;15(2):154–62.
31. Audina M, Maigoda TC, Wahyu T. Status Gizi, Aktivitas Fisik dan Asupan Serat Berhubungan dengan Kadar Gula Darah Puasa Penderita DM Tipe 2. J Ilmu Dan Teknol Kesehat. 2018;6(1):59–71.
32. Murtiningsih MK, Pandelaki K, Sedli BP. Gaya Hidup sebagai Faktor Risiko Diabetes Melitus Tipe 2. e-CliniC [Internet]. 2021 Mar 16;9(2):328–33. Available from: https://ejournal.unsrat.ac.id/index.php/eclinic/article/view/32852
33. DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, et al. Type 2 diabetes mellitus. Nat Rev Dis Prim [Internet]. 2015 Jul 23;1(1):15019. Available from: https://www.nature.com/articles/nrdp201519
34. Galicia-Garcia U, Benito-Vicente A, Jebari S, Larrea-Sebal A, Siddiqi H, Uribe KB, et al. Pathophysiology of Type 2 Diabetes Mellitus. Int J Mol Sci [Internet]. 2020 Aug 30;21(17):1–34. Available from: https://www.mdpi.com/1422-0067/21/17/6275
35. Sastrawan IKW, Darmini AAAY, Suantika PIR, Megayanti SD. Hubungan Perilaku Merokok dengan Kejadian Komplikasi Kronis pada Pria dengan Diabetes Melitus Tipe II. J Persat Perawat Nas Indones [Internet]. 2023 Mar 29;7(3):113–9. Available from: http://www.jurnal-ppni.org/ojs/index.php/jppni/article/view/386
36. Irayani SP. Hubungan Riwayat Keluarga, Aktivitas Fisik, dan Pola Makan terhadap Kejadian Diabetes Melitus. J Public Heal Educ [Internet]. 2024 Jul 29;3(4):145–52. Available from: https://journals.prosciences.net/index.php/JPHE/article/view/227
37. Paramita DP, Lestari A. W. Pengaruh Riwayat Keluarga Terhadap Kadar Glukosa Darah Pada Dewasa Muda Keturunan Pertama dari Penderita Diabertes Mellitus Tipe 2 di Denpasar Selatan. E-Jurnal Med. 2019;8(3):1–6.
38. Althnian A, AlSaeed D, Al-Baity H, Samha A, Dris A Bin, Alzakari N, et al. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. Appl Sci [Internet]. 2021 Jan 15;11(2):1–18. Available from: https://www.mdpi.com/2076-3417/11/2/796
39. Prasetyo YA, Utami E, Yaqin A. Pengaruh Komposisi Split Data Terhadap Performa Akurasi Analisis Sentimen Algoritma Naïve Bayes dan SVM. J Electr Eng Comput [Internet]. 2024 Oct 4;6(2):382–90. Available from: https://ejournal.unuja.ac.id/index.php/jeecom/article/view/9188

40. Salasa RA, Rahman H, Andiani A. Faktor Risiko Diabetes Mellitus Tipe 2 Pada Populasi Asia: A systematic Review. J BIOSAINSTEK [Internet]. 2019 Nov 24;1(01):95–107. Available from: http://www.jurnal.ummu.ac.id/index.php/BIOSAINSTEK/article/view/306
41. Septiani WD, Marlina M. Comparison of Decision Tree, Naive Bayes, and Neural Network Algorithm for Early Detection of Diabetes. J Pilar Nusa Mandiri [Internet]. 2021 Mar 5;17(1):73–8. Available from: https://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/2213