International Conference on Health and Biological Science
ICHBS, Page 280-295, December 2024

Machine Learning for Gestational Diabetes Mellitus: Comparing Random Forest, K-Nearest Neighbor and Naive Bayes for Effective Risk Classification

Ulan Juniarti[1*], *Retno* Agus Setiawan[1], and *Rosyid* Ridlo Al-Hakim[1]

[1]Information Systems Study Program, Faculty of Science and Technology, Universitas Harapan Bangsa, Banyumas, Central Java, Indonesia

Abstract. Gestational diabetes mellitus (GDM) is a complication of pregnancy characterized by impaired glucose tolerance arising or initially identified during pregnancy. This condition requires machine learning-based classification methods that can identify risks early, as conventional methods such as screening often cause delays in risk identification and lack accuracy due to variations in maternal health conditions. Machine learning offers a solution by providing faster and more accurate classification of GDM, due to its ability to quickly process large data and analyze data involving many variables. This study explores the use of Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes algorithms for GDM risk classification to determine the most effective model. Using a dataset containing 1012 samples from the Kurdistan region, the researchers performed data pre-processing, including data cleaning, data balancing using SMOTE, and normalization, followed by model training and evaluation based on accuracy, AUC, sensitivity, and specificity metrics. The results showed that Random Forest achieved the highest accuracy of 86.43%, AUC of 93.78%, sensitivity of 89.29%, and specificity of 83.57. Following that, KNN had an accuracy of 83.93%, AUC of 83.93%, sensitivity of 88.57%, and specificity of 79.29%. Lastly, Naive Bayes reached an accuracy of 76.79%. Based on these results, Random Forest is the best performing algorithm for effective GDM risk classification. This study emphasizes the potential of machine learning to enhance the speed and accuracy of early GDM risk prediction, ultimately contributing to better health outcomes for both mothers and their children.

1 Introduction

Gestational diabetes mellitus (GDM) is a prevalent pregnancy complication characterized by impaired glucose tolerance that arises or is initially identified during

---

[*] Corresponding author: ulanjuniarti82@gmail.com

pregnancy (1). The global prevalence of GDM ranges from 5% to 25.5% (2). Southeast Asia, the prevalence of GDM reaches 23.7% and is one of the highest rates (3). While the prevalence of GDM in Indonesia reached 3.6% (4). Maternal conditions such as diabetes or other diseases, including GDM, can lead to serious pregnancy complications, such as preterm birth and even death (5). GDM raises the likelihood of adverse neonatal outcomes, including macrosomia, babies who are large for gestational age (LGA), and low Apgar scores (6). Pregnant women who have gestational diabetes mellitus (GDM) are also more likely to develop type 2 diabetes and cardiovascular disease after birth (7). Genetic factors, such as single nucleotide variations, can increase the risk of gestational diabetes mellitus by affecting insulin signaling pathways, cell transport, and enzyme function (8). In addition, being overweight or obese before pregnancy can also increase the risk of developing GDM (9). Typically, an oral glucose tolerance test (OGTT) conducted between weeks 24 and 28 of pregnancy is used to identify GDM. Because physiological insulin resistance in pregnancy is typically well-developed at this point, this time frame was selected (10). Early detection of GDM is very important, as it can help prevent or reduce the likelihood of poor pregnancy outcomes (11). However, conventional prediction models that rely on maternal characteristics at the beginning of pregnancy often have limited accuracy, as they are unable to account for dynamic changes in health parameters that occur as pregnancy progresses (12). Therefore, Machine learning techniques offer a more precise and reliable solution for early GDM prediction compared to traditional screening methods, especially in diverse populations with significant variability in maternal health conditions (13).

Machine Learning is a method in which computers utilize available data to learn independently and optimize their performance without specific direct instructions (14). Machine learning models have demonstrated their ability to predict and manage chronic diseases such as diabetes, heart disease and cancer. By utilizing in-depth patient data, these models can identify potential disease risks and provide support for faster intervention, allowing the management of complex medical conditions before they develop further (15). Several previous studies have explored the use of machine learning algorithms in GDM risk classification. For example, (16) developed GDMPredictor, a predictive tool that attempts to calculate the risk of gestational diabetes (GDM) based on clinical and biochemical markers, using machine learning techniques. This study used various machine learning algorithms, such as Random Forest, Support Vector Machine (SVM), Gaussian Naive Bayes (NB), K-nearest Neighbors (k-NN), AdaBoost, Gradient Boosting, Bernoulli Naive Bayes (NB), and Decision Tree. Of all the algorithms used, Random Forest showed the highest accuracy of 96%, making it a highly effective tool in predicting GDM risk. This algorithm shows great potential in helping early identification of high-risk patients and providing more precise and personalized treatment recommendations. On the other hand, research conducted by (17) aimed to identify factors influencing the onset of type 2 diabetes (T2DM) in women who had previously experienced gestational diabetes (GDM) by applying machine learning methods. The study utilized several algorithms, including Decision Tree, which achieved an accuracy of 82.7%, and Naive Bayes, which produced a classification accuracy of 81.3%. Another study by (18) aims to develop a prediction model that can recognize gestational diabetes in pregnant women using a combination of machine learning algorithms. The algorithms used include Decision Tree with 95% accuracy, Random Forest has an accuracy of 94%, Support Vector Machine (SVM) has an accuracy of 77%, K-nearest neighbors (KNN) with 93% accuracy, Logistic Regression accuracy of 74%, and Naive Bayes has an accuracy of 73%. This research shows that the combination of K-means clustering with classification algorithms results in a significant increase in accuracy, especially in the Desicion Tree, Random Forest and KNN algorithms.

Previous research has shown that machine learning techniques like Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN) offer higher accuracy in predicting the likelihood of GDM compared to traditional methods such as the OGTT, which face challenges in accuracy and speed, particularly when detecting GDM at early stages. Based on these existing gaps, the hypothesis of this study is that the Random Forest algorithm will achieve higher accuracy in classifying GDM risk compared to the Naive Bayes and K-Nearest Neighbors (KNN) algorithms and will be superior to traditional methods such as OGTT. This is because Random Forest can handle the complexity of medical data while providing faster and more accurate results. In order to identify the best machine learning algorithm for early DMG illness classification, this study compares three different algorithms, Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN). This research also applies SMOTE to address data imbalance and evaluates model performance with metrics such as accuracy, AUC, sensitivity, and specificity. This comprehensive comparison is critical, as identifying the most effective algorithm can significantly impact early detection and timely intervention for GDM, ultimately leading to better maternal and infant health outcomes by enabling targeted risk management during pregnancy.

2 Method

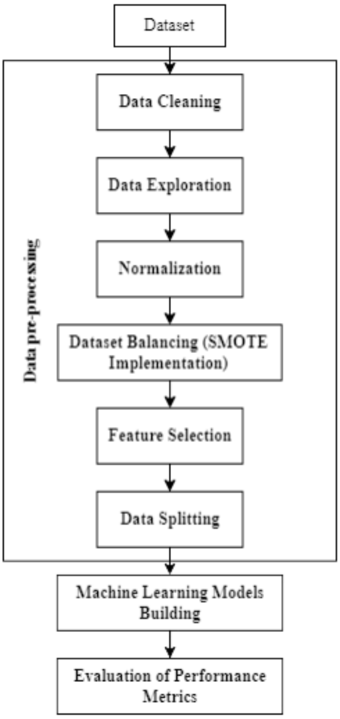


Fig. 1. Research Stages

Figure 1 (19) are the stages of research used in the classification of GDM risk using machine learning. With an explanation of the steps as follows:

2.1 Dataset

Data is a set of information obtained from activities in the real world. After being converted into a format that can be understood by machines, the data is stored in files commonly known as datasets (20). Datasets have become essential in the advancement of machine learning research. They form the basis for the models we develop and deploy, while also acting as the main resource for testing and evaluating model performance (21). The dataset used in this study is a dataset obtained from the Kurdistan region laboratory, which collected information from pregnant women with and without diabetes. The dataset consists of 1012 data and has several important variables that can affect the risk of gestational diabetes, such as age, pregnancy no, weight, height, BMI, heredity and prediction. However, there are some limitations and potential biases to consider such as the data set only comes from 1 region so it may not fully represent populations in other regions that have different genetic backgrounds, environments, or lifestyles. This may limit the ability of the model to work accurately for people outside of this region. Another limitation is the small number of datasets and variables. Future research could use data from different regions with more data and variables.

2.2 Data Preprocessing

Data preprocessing plays a crucial role in ensuring the quality of the analysis, as using unprocessed raw data typically fails to deliver accurate results and may lead to incorrect conclusions (22). This process includes handling missing data, duplicate data, outliners, data exploration, data normalization, data balancing, feature selection, and data splitting. One of the key steps in this study is addressing class imbalance in the dataset, where the number of high-risk GDM cases is lower than non-risk cases. The Synthetic Minority Over-sampling Technique (SMOTE) is used to control this imbalance. By creating synthetic instances for the minority class, SMOTE handles unbalanced data, preventing overfitting and guaranteeing that the model is not skewed toward the majority class (23). In this research, data division is carried out in a ratio of 80:20, where 80% of the dataset is for testing data and 20% of the dataset is for training data. Each step of data preprocessing is done to ensure that the data used in the machine learning model is of high quality and ready to be used in the predictive analysis process. Thus, data preprocessing helps optimize the accuracy and generalization of the model to new data.

2.3 Machine Learning Models Building

Once data preprocessing is complete, the next step is to build and train machine learning models that aim to predict the risk of gestational diabetes mellitus. This process involves selecting algorithms, training models, and evaluating their performance. In this study, various machine learning algorithms are employed, including Random Forest, Naive Bayes, and KNN. Random Forest is a predictive algorithm that works by constructing multiple decision trees using random subsets of the data, with the final prediction determined by the majority vote from all the generated trees (24). Random forest was chosen because this algorithm is not affected by statistical assumptions and preprocessing loads, and is able to process large datasets with many variables (25). Naive Bayes is a straightforward probabilistic approach grounded in Bayes' Theorem, which assumes that all attributes are independent of one another (26). Naive Bayes was chosen because this algorithm is very effective, especially in managing data with high dimensions and unbound attributes, making this algorithm the right choice for large datasets (27). KNN is a non-parametric technique that uses the closest neighbor class to categorize a fresh sample (28). This algorithm was chosen for its effective ability to handle large datasets as well as its robustness against data containing noise (29). By applying these three algorithms, this research aims to compare the performance of each method in predicting the risk of gestational diabetes mellitus.

2.4 Building Evaluation of Performance Metricts

Evaluation has an important role in machine learning (ML). Evaluation is a tool designed to support best practices in measurement, metrics, and comparison of data and models. The goal is to ensure the evaluation is replicable, centralize and document the process, and extend the evaluation to cover more aspects of model performance (30). The research evaluation was conducted using Receiver Operating Characteristic (ROC) curves, Area Under the Curve (AUC), diagnostic accuracy, sensitivity, and specificity. ROC curves are utilized to assess the overall effectiveness of a diagnostic test, compare the performance of multiple tests, and identify the optimal threshold value for determining whether an individual has a particular disease (31). The area beneath the ROC curve, known as AUC, is a measure that evaluates how well the model distinguishes between the positive and negative classes (32). Sensitivity gauges how effectively the model detects actual positive cases, while specificity assesses its capacity to correctly identify true negative cases (33). Combining these evaluation metrics can ensure the reliability, accuracy, and suitability of the model to determine the most effective algorithm in GDM disease classification.

3. Result and Discussion

3.1. Gestational Diabetes Mellitus Dataset

The study was conducted using a dataset obtained from the Kurdistan region laboratory, which collected information from pregnant women with and without diabetes. It can be seen in Table 1. Contains examples of Gestational Diabetes Mellitus data used in the study displayed in the first 5 rows and the last 5 rows, where the data amounted to 1012 data and had 7 columns with variables such as age, pregnancy no, weight, height, BMI, heredity and prediction. The data will be an overview of the characteristics of pregnant women who are at risk and not at risk of gestational diabetes mellitus. Each variable has an important role in building an accurate prediction model.

Table 1. Gestational Diabetes Mellitus Data Sampl

No	Age	Pregnancy No	Weight	Height	BMI	Heredity	Prediction
1	17.00	1.0	48.0	165.0	17.6	0	0
2	17.00	1.0	49.0	145.0	23.3	0	0
3	17.00	1.0	50.0	140.0	25.5	0	0
4	17.00	1.0	50.0	145.0	23.8	0	0
5	17.00	1.0	49.0	146.0	23.0	0	0
...	...	...	...	...	...	...	...
1008	35.0	3.0	89.0	159.0	35.2	1	1
1009	41.0	4.0	87.0	165.0	32.0	0	0
1010	34.0	2.0	67.0	160.0	26.2	1	0

1011	33.0	3.0	65.0	167.0	23.3	0	1
1012	28.0	2.0	68.0	156.0	27.9	0	0

3.2. Data Preprocessing

In the data preprocessing stage, several steps are taken to prepare the data before modeling. These steps include data cleaning, data exploration, normalization, dataset balancing, feature selection, and data division.

3.2.1. Data Cleaning

The first step in data preprocessing is to check for missing data, duplicate data, and outliers. If any of these issues are found, data deletion or replacement (using mean or median values) is performed to prevent bias. In this dataset, there were no missing values, but duplicate and outlier data were present. Duplicate data was identified and removed to avoid redundancy, which could have led to overrepresentation of certain samples, affecting the model's performance. Outliers were also detected using box plots and the Interquartile Range (IQR) method. These outliers can distort statistical analysis and negatively impact the model by skewing relationships between variables. While outliers may reflect valid extreme values or rare cases, they can also be the result of data entry errors. To ensure a clean and relevant dataset, both duplicate and outlier data were removed, reducing the dataset size from 1012 to 892 samples. This step aimed to improve the accuracy and performance of the model in predicting the risk of gestational diabetes.

3.2.2. Data Exploration

The next process is data exploration, which is done to gain a deeper understanding of the data's structure and properties before using it to develop machine learning models. Data exploration aims to find patterns, trends, anomalies, and relationships between features in the dataset.

Based on Fig. 2. the distribution of the data shows clear characteristics of the pregnant women in this dataset. Most of the mothers are within the common reproductive age range of 20 to 35 years old, with only a few being under 20 or over 40 years old. The number of pregnancies was dominated by first or second pregnancies, while more than three pregnancies were rare. Maternal weight distribution focused on the 60 to 80 kg range, with a few individuals weighing over 100 kg, which could be an indication of potential obesity-related health risks. In contrast, the height distribution showed that most mothers were between 155 and 165 cm tall, which is a common height range, although there was little variation outside this range. Most mothers in the dataset have a BMI in the normal to overweight category, with the distribution peaking at around 25 to 30. This suggests that most mothers are in normal health, although there are some who have a high BMI above 35, which falls into the obese category. As for the distribution of heredity, most mothers do not have a family history of diabetes, with only a small percentage having such a history. This suggests that heredity may not be the main cause of gestational diabetes risk in this population, and other factors such as BMI and age likely play a greater role.

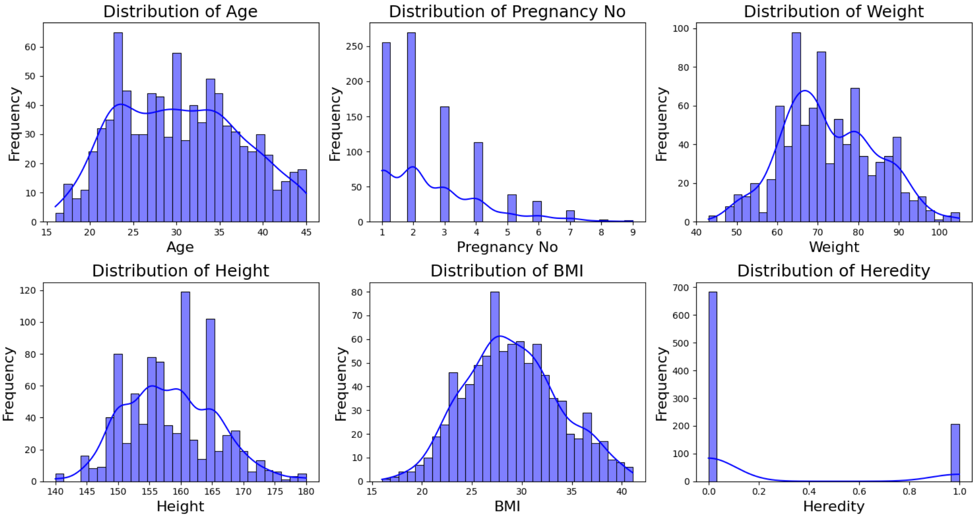
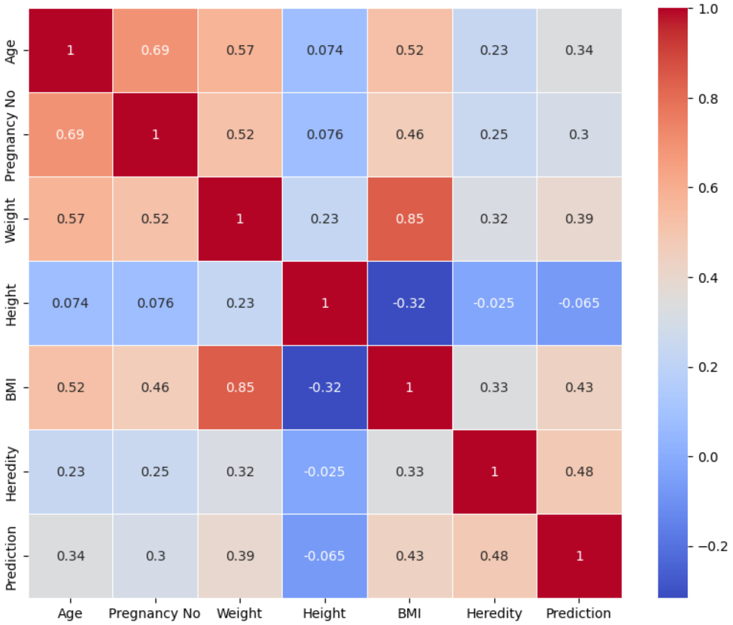


Fig. 1. Distribution of Diabetes Mellitus Gestational Data.



	Age	Pregnancy No	Weight	Height	BMI	Heredity	Prediction
Age	1	0.69	0.57	0.074	0.52	0.23	0.34
Pregnancy No	0.69	1	0.52	0.076	0.46	0.25	0.3
Weight	0.57	0.52	1	0.23	0.85	0.32	0.39
Height	0.074	0.076	0.23	1	-0.32	-0.025	-0.065
BMI	0.52	0.46	0.85	-0.32	1	0.33	0.43
Heredity	0.23	0.25	0.32	-0.025	0.33	1	0.48
Prediction	0.34	0.3	0.39	-0.065	0.43	0.48	1

Fig. 2. Heatmap of Diabetes Mellitus Gestational Data.

Fig. 3. is a heatmap of the GDM data showing the relationship between variables. Variables with correlation values close to one indicate a very strong relationship, while variables with correlation values close to zero have a low correlation. Highly correlated variables are marked with dark colors on this map.

Based on the correlation analysis of the heatmap, there are several significant relationships between the variables in the dataset. A high correlation is seen between weight and BMI with a correlation value of 0.85, indicating that these two variables are highly correlated. In addition, there is a fairly strong correlation between age and number of pregnancies (0.69), indicating that the older a pregnant woman is, the more pregnancies she tends to have. For a moderate correlation, BMI and predicted risk had a correlation of 0.43, indicating that the higher the BMI, the greater the risk of gestational diabetes. The dataset's BMI distribution revealed that the majority of pregnant women fell into the normal to overweight range, with a small percentage of pregnant women having a BMI exceeding 35, which is regarded as obese. Since elevated insulin resistance is a primary contributing factor to the development of diabetes during pregnancy, a higher BMI has a considerable impact on the risk of GDM. Heredity, with a correlation of 0.48, also affects the risk of GDM, suggesting that genetic factors have an important role in the increased risk of gestational diabetes mellitus. This correlation reflects that a family history of diabetes tends to increase the likelihood of GDM, as genetic factors can affect how the body regulates glucose levels and insulin response. In addition, body weight had a correlation of 0.39, which suggests that body weight plays a moderate role in the risk of GDM. This correlation could be due to the fact that higher body weight is associated with metabolic changes and increased insulin resistance, both of which contribute to the risk of diabetes during pregnancy. Although the correlation is not as strong as heredity, weight remains an important factor to consider in GDM risk assessment. On the other hand, the height variable has low correlations with most other variables, including a small negative correlation with BMI at -0.32 and a very low correlation with risk prediction at -0.065, indicating that height does not play a significant role in predicting gestational diabetes risk in this dataset.

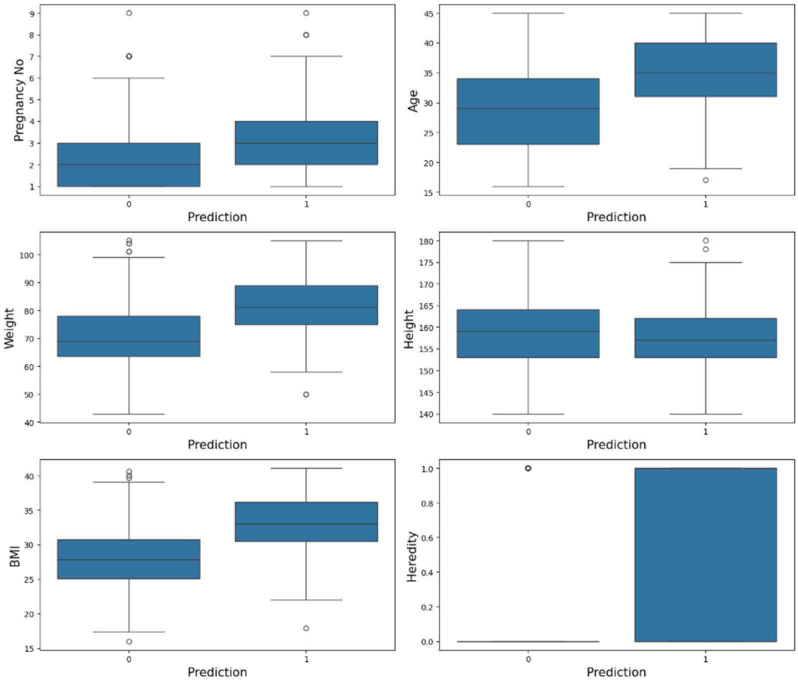


Fig. 3. Box Plot Analysis of Destational Diabetes Data.

Fig. 4. shows the distribution pattern of the important variables. Pregnancy No tends to be on the lower side, indicating that most pregnancy values are in the minimum range. Age shows greater variation in the at-risk group with some outliers in both groups. Weight is evenly distributed in the interquartile range with some outliers, while Height shows a similar distribution in both groups without many significant differences. BMI

in the at-risk group was higher than the non-risk group, with some outliers exceeding the maximum range, suggesting BMI as an important factor in gestational diabetes risk. Meanwhile, Heredity showed most of the non-risk group had no family history of diabetes, while the at-risk group had a wider distribution, indicating heredity as a significant variable in predicting risk.

3.2.3. Normalization

After the data exploration stage, the next step in data preprocessing is data normalization. In this study, normalization was carried out using StandardScaler, which standardizes the numeric variables to have a mean of 0 and a standard deviation of 1. This process is important because some machine learning algorithms, such as K-Nearest Neighbors (KNN), are sensitive to the scale of the data, and normalization ensures that all features have an equal impact on the model. The normalized features include Age, Pregnancy No, Weight, Height, and BMI. After normalization, each feature in the dataset has the same scale and is ready for the next process.

3.2.4. Dataset Balancing

Continued with dataset balancing to handle data imbalance in the target variable, namely prediction. The proportion of pregnant women in this sample who are at risk for gestational diabetes is unbalanced. Since the model tends to predict the majority class more accurately, this imbalance may have an impact on model performance if left unchecked. This problem was solved by using the SMOTE (Synthetic Minority Over-sampling Technique) approach, which creates a synthetic sample in order to enhance the amount of data in the minority class. By using SMOTE, the dataset's fraction of positive and negative classes is more evenly distributed, improving the model's ability to predict both classes.

Before SMOTE, the number of datasets in class 0 (not at risk of gestational diabetes) was 699 data, while class 1 (at risk of gestational diabetes) was 193 data. This distribution shows that the initial dataset has class imbalance, where the amount of data in class 0 is much more than class 1. After SMOTE, the number of datasets in class 0 is 699 and the number of data in class 1 is 699. This shows that the dataset becomes balanced with the same amount of data in both classes, i.e. 699 data in class 0 and 699 data in class 1. This will help the machine learning model to work better in predicting both classes without bias towards the majority class, which will ultimately improve the overall performance of the model.

3.2.5. Feature Selection

The next stage is feature selection. Through the feature selection process, the main goal is to select the most relevant and significant features in predicting the risk of gestational diabetes in pregnant women. By applying certain statistical methods or algorithms, researchers can identify and eliminate features that are less important or redundant, so that the resulting model becomes more efficient and has more optimal performance. Feature selection also helps in reducing the number of dimensions in the dataset, which in turn can speed up the computational process and reduce the risk of overfitting.

3.2.6. Data Splitting

After going through the feature selection stage, the last step in preprocessing is data splitting. 80% of the data is allocated for training, while the remaining 20% is reserved for testing. This split allows the model to be trained with the training set and its performance assessed using the testing set. With an 80:20 distribution, most of the data is utilized to train the model in recognizing patterns, while the test set is employed to evaluate the model's ability to predict new, unseen data. In sharing the dataset, 1118 samples were used as training data, with a balanced distribution between classes 0 (not at risk) and 1 (at risk), with 559 samples each. Meanwhile, the testing data used 280 samples with a balanced distribution between classes 0 and 1, each with 140 samples.

3.3 Machine Learning Models Building

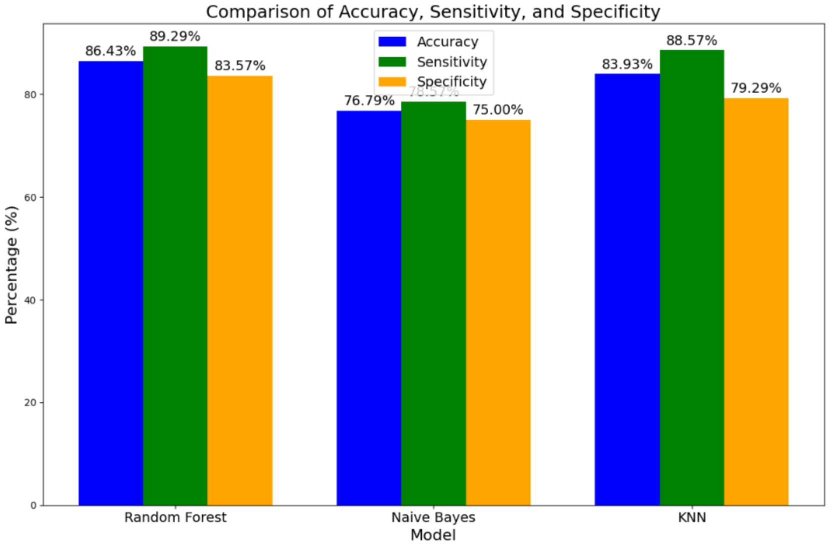


Fig. 4. Comparison of Accuracy, Sensitivity, and Specificity

Fig. 5. presents a comparison of the performance of three machine learning algorithms, namely Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes based on accuracy, sensitivity, and specificity. According to the diagram, Random Forest outperformed the other models in all metrics, achieving 86.43% accuracy, 89.29% sensitivity, and 83.57% specificity. KNN also showed strong results, with 83.93% accuracy, 88.57% sensitivity, and 79.29% specificity. In contrast, Naive Bayes performed the least well, with an accuracy of 76.79%, sensitivity of 78.57%, and specificity of 75.00%. These results indicate that Random Forest is the most consistent and reliable model, demonstrating a well-balanced performance across all evaluation metrics. Its ability to manage intricate and non-linear relationships between features by averaging multiple decision trees enhances its classification accuracy. This strength makes Random Forest the best option for classifying GDM disease.

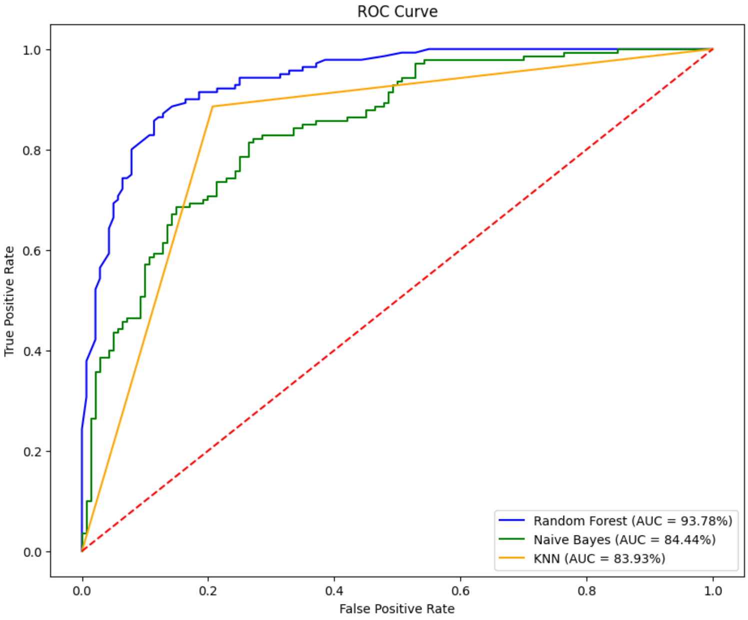


Fig. 5. Receiver Operating Characteristic Curve

Fig. 6. shows the Receiver Operating Characteristic (ROC) Curve of the three machine learning algorithms used in the study, namely Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN). The ROC curve plots the relationship between the True Positive Rate (TPR) on the vertical axis and the False Positive Rate (FPR) on the horizontal axis, allowing for an evaluation of how well the model distinguishes between positive and negative classes. A model's performance is considered better when its ROC curve is closer to the top-left corner, indicating higher predictive accuracy.

AUC (Area Under the Curve) provides a numerical score for this ability. AUC values range from 0 to 1, values closer to 1 mean the model performs better at identifying positive and negative cases correctly, while values closer to 0.5 indicate the model performs similarly to random guessing. In simpler terms, a higher AUC value means the model makes more accurate predictions and fewer errors. Based on the ROC curves, Random Forest achieves the highest performance, with an AUC of 93.78%, which is very close to 1. This indicates that Random Forest is highly accurate at separating positive and negative cases. Naive Bayes follows with an AUC of 84.44%, showing good performance, though not as accurate as Random Forest. Lastly, KNN has an AUC of 83.93%, slightly lower than Naive Bayes, indicating it performs reasonably well but is less accurate than both Random Forest and Naive Bayes.

Table 1. Model Performance Comparison

Algoritma	Accuracy (%)	AUC (%)	Sensitivity(%)	Specificity(%)
RF	86.43	93.78	89.29	83.57
KNN	83.93	83.93	88.57	79.29
NB	76.79	84.44	78.57	75.00

Abbreviations: RF, Random Forest ; KNN, K-Nearest Neighbor; NB, Naive Bayes; AUC, Area Under the Curve.

Tabel 2 shows the performance of Random Forest, KNN, and Naive Bayes is evaluated based on key metrics, including accuracy, AUC, sensitivity, and specificity. Random Forest has an accuracy of 86.43%, AUC 93.78%, sensitivity 89.29%, and specificity 83.57%, indicating that this model is very good at predicting gestational diabetes risk with high accuracy and good ability to separate positive and negative classes. KNN also performed well with 83.93% accuracy, 83.93% AUC, 88.57% sensitivity, and 79.29% specificity. Although slightly lower than Random Forest, KNN remains a fairly reliable algorithm in detecting risk, especially with a sensitivity that is almost close to Random Forest. On the other hand, Naive Bayes has a lower performance with an accuracy of 76.79%, AUC of 84.44%, sensitivity of 78.57%, and specificity of 75.00%, indicating that this algorithm is less optimal in predicting gestational diabetes risk than RF and KNN. Although its AUC is quite good, its lower sensitivity and specificity indicate that Naive Bayes more often makes mistakes in predicting the risk and non-risk classes. Overall, Random Forest emerged as the best performing model, followed by KNN, and Naive Bayes showing the lowest performance among the three algorithms. This is because Random Forest has resilience against overfitting and its ability to capture complex patterns in the data due to the ensemble of decision trees. This robustness makes Random Forest very effective for classification problems that involve complex relationships between features, such as in gestational diabetes prediction.

3.4. Evaluation

Based on the results of modeling three models used to predict the risk of gestational diabetes in pregnant women, the evaluation shows that Random Forest (RF) has the best performance with the highest accuracy, which is 86.43%, AUC 93.78%, sensitivity 89.29%, and specificity 83.57%. These results are consistent with previous research by [16], who used a machine learning approach in the development of GDMPredictor, a gestational diabetes risk prediction tool based on clinical and biochemical parameters. In that study, Random Forest also showed the highest accuracy among other algorithms, making it one of the most effective algorithms in predicting GDM risk.

The KNN algorithm had an accuracy of 83.93%, AUC of 83.93%, sensitivity of 88.57%, and specificity of 79.29%. These results show that KNN is one of the effective algorithms in classification tasks, especially in data related to diabetes problems. This is in line with research [18], where KNN also showed strong performance with an accuracy of 93%. Although there are differences in accuracy results between the table and previous research, these results validate that KNN is a reliable and widely used algorithm in prediction modeling, including in the case of gestational diabetes prediction.

The Naive Bayes (NB) algorithm recorded an accuracy of 76.79%, AUC of 84.44%, sensitivity of 78.57%, and specificity of 75%. These results show that NB is one of the effective algorithms in classification which is supported by research [17] where the accuracy of NB was 81.3%, higher than the results in the table. This difference may be due to the different characteristics of the data used in the two studies, or due to different data processing methods. Although NB is usually not as good as other algorithms such as Random Forest or KNN in some classification cases, the results from this table validate that NB is still effective and suitable for classification tasks, especially in medical prediction problems such as diabetes prediction.

Based on the three algorithms that have been compared, namely Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bayes (NB), all three show that each algorithm has its own effectiveness in predicting the risk of gestational diabetes. These models, especially Random Forest, are highly relevant in a clinical context as they can help obstetricians predict the risk of gestational diabetes early on, allowing for timely interventions and treatments that are better suited to the patient's needs. However, these findings are based on a data set limited to one region, which may limit their applicability to populations with different characteristics or healthcare systems. To improve generalizability, future studies should use more diverse data and explore more advanced machine learning techniques, such as deep learning, which can recognize more complex patterns. In addition, the addition of other factors such as genetic data or lifestyle information may improve the accuracy of the model. Ethical considerations, such as the protection of patient data privacy and the importance of human oversight of AI-generated predictions, are also crucial to ensure safe and responsible use of these models in clinical practice.

Conclusion

This study compared three algorithms, namely Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes in gestational diabetes mellitus (GDM) risk classification and found that Random Forest outperformed the other algorithms with the highest accuracy of 86.43%, making it the most reliable model for GDM prediction. KNN also performed well with 83.93% accuracy, while Naive Bayes, with 76.79% accuracy, remained usable but less than optimal. These findings highlight the potential of machine learning, specifically Random Forest, in improving early detection of GDM, enabling timely interventions that can improve maternal and fetal health outcomes. Healthcare providers can use these models as decision support tools to assess risk early in pregnancy, leading to better management through lifestyle modifications or increased monitoring. Although Random Forest offers the best accuracy, KNN and NB still provide valuable alternatives depending on the context or data set. Integrating machine learning into clinical practice can significantly improve the accuracy and efficiency of GDM risk assessment and drive better outcomes for patients. Going forward, healthcare systems can prioritize the adoption of these Machine Learning technologies to ensure timely diagnosis and intervention thereby improving health for mothers and children.

References

1. Moon JH, Jang HC. Gestational Diabetes Mellitus: Diagnostic Approaches and Maternal-Offspring Complications. Diabetes & Metabolism Journal [Internet]. 2022 Jan 31;46(1):3–14. Available from: http://e-dmj.org/journal/view.php?doi=10.4093/dmj.2021.0335
2. Choudhury AA, Devi Rajeswari V. Gestational diabetes mellitus - A metabolic and reproductive disorder. Biomedicine & Pharmacotherapy [Internet]. 2021 Nov;143:112183. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0753332221009677
3. Wang H, Li N, Chivese T, Werfalli M, Sun H, Yuen L, et al. IDF Diabetes Atlas: Estimation of Global and Regional Gestational Diabetes Mellitus Prevalence for 2021 by International Association of Diabetes in Pregnancy Study Group's Criteria. Diabetes Research and Clinical Practice [Internet]. 2022 Jan;183:109050.

Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168822721004095

4. Sumarah S, Anies A, Rahfiludin MZ, Sulistiyani S. Factors related to HbA1c in the first trimester of pregnancy. International Journal of Public Health Science (IJPHS) [Internet]. 2023 Dec 1;12(4):1491. Available from: https://ijphs.iaescore.com/index.php/IJPHS/article/view/23189
5. Purwono P, Wirasto A, Setiawan RA, Triwibowo DN, Zuhrufillah I, Sumantri RBB, et al. Analysis of the relationship between fetal health prediction features with machine learning feyn qlattice regression model. In 2023. p. 020183. Available from: http://aip.scitation.org/doi/abs/10.1063/5.0120255
6. Muche AA, Olayemi OO, Gete YK. Gestational diabetes mellitus increased the risk of adverse neonatal outcomes: A prospective cohort study in Northwest Ethiopia. Midwifery [Internet]. 2020 Aug;87:102713. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0266613820300863
7. Diaz-Santana M V., O'Brien KM, Park Y-MM, Sandler DP, Weinberg CR. Persistence of Risk for Type 2 Diabetes After Gestational Diabetes Mellitus. Diabetes Care [Internet]. 2022 Apr 1;45(4):864–70. Available from: https://diabetesjournals.org/care/article/45/4/864/141055/Persistence-of-Risk-for-Type-2-Diabetes-After
8. Ortega-Contreras B, Armella A, Appel J, Mennickent D, Araya J, González M, et al. Pathophysiological Role of Genetic Factors Associated With Gestational Diabetes Mellitus. Frontiers in Physiology [Internet]. 2022 Apr 4;13. Available from: https://www.frontiersin.org/articles/10.3389/fphys.2022.769924/full
9. Song X, Shu J, Zhang S, Chen L, Diao J, Li J, et al. Pre-Pregnancy Body Mass Index and Risk of Macrosomia and Large for Gestational Age Births with Gestational Diabetes Mellitus as a Mediator: A Prospective Cohort Study in Central China. Nutrients [Internet]. 2022 Mar 3;14(5):1072. Available from: https://www.mdpi.com/2072-6643/14/5/1072
10. McIntyre HD, Kapur A, Divakar H, Hod M. Gestational Diabetes Mellitus—Innovative Approach to Prediction, Diagnosis, Management, and Prevention of Future NCD—Mother and Offspring. Frontiers in Endocrinology [Internet]. 2020 Dec 3;11. Available from: https://www.frontiersin.org/articles/10.3389/fendo.2020.614533/full
11. Artzi NS, Shilo S, Hadar E, Rossman H, Barbash-Hazan S, Ben-Haroush A, et al. Prediction of gestational diabetes based on nationwide electronic health records. Nature Medicine [Internet]. 2020 Jan 13;26(1):71–6. Available from: https://www.nature.com/articles/s41591-019-0724-8
12. Snyder BM, Baer RJ, Oltman SP, Robinson JG, Breheny PJ, Saftlas AF, et al. Early pregnancy prediction of gestational diabetes mellitus risk using prenatal screening biomarkers in nulliparous women. Diabetes Research and Clinical Practice [Internet]. 2020 May; 163:108139. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168822720303892
13. Zhang Z, Yang L, Han W, Wu Y, Zhang L, Gao C, et al. Machine Learning Prediction Models for Gestational Diabetes Mellitus: Meta-analysis. Journal of Medical Internet Research [Internet]. 2022 Mar 16;24(3):e26634. Available from: https://www.jmir.org/2022/3/e26634
14. A. A. Permana, R. Darmawan and RRA-H. Artificial Intelligence Marketing [Internet]. Available from: https://www.researchgate.net/publication/373043823
15. Goecks J, Jalili V, Heiser LM, Gray JW. How Machine Learning Will Transform

Biomedicine. Cell [Internet]. 2020 Apr;181(1):92–101. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867420302841

16. Xing J, Dong K, Liu X, Ma J, Yuan E, Zhang L, et al. Enhancing gestational diabetes mellitus risk assessment and treatment through GDMPredictor: a machine learning approach. Journal of Endocrinological Investigation [Internet]. 2024 Mar 9;47(9):2351–60. Available from: https://link.springer.com/10.1007/s40618-024-02328-z
17. Ilari L, Piersanti A, Göbl C, Burattini L, Kautzky-Willer A, Tura A, et al. Unraveling the Factors Determining Development of Type 2 Diabetes in Women With a History of Gestational Diabetes Mellitus Through Machine-Learning Techniques. Frontiers in Physiology [Internet]. 2022 Feb 17;13. Available from: https://www.frontiersin.org/articles/10.3389/fphys.2022.789219/full
18. Rasool Jader, Sadegh Aminifar. An Intelligent Gestational Diabetes Mellitus Recognition System Using Machine Learning Algorithms. Tikrit Journal of Pure Science [Internet]. 2023 Feb 20;28(1):82–8. Available from: https://www.tjpsj.org/index.php/tjps/article/view/1269
19. Widodo AO, Setiawan B, Indraswari R. Machine Learning-Based Intrusion Detection on Multi-Class Imbalanced Dataset Using SMOTE. Procedia Computer Science [Internet]. 2024;234:578–83. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1877050924004009
20. Reddy Sankepally S, Kosaraju N, Mallikharjuna Rao K. Data Imputation Techniques: An Empirical Study using Chronic Kidney Disease and Life Expectancy Datasets. In: 2022 International Conference on Innovative Trends in Information Technology (ICITIIT) [Internet]. IEEE; 2022. p. 1–7. Available from: https://ieeexplore.ieee.org/document/9744211/
21. Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns [Internet]. 2021 Nov;2(11):100336. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2666389921001847
22. Madrakhimov S, Makharov K, Lolaev M. Data preprocessing on input. In 2021. p. 030003. Available from: https://pubs.aip.org/aip/acp/article/1028483
23. Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya INH, Ismail M. SMOTE for Handling Imbalanced Data Problem : A Review. In: 2021 Sixth International Conference on Informatics and Computing (ICIC) [Internet]. IEEE; 2021. p. 1–8. Available from: https://ieeexplore.ieee.org/document/9632912/
24. Schonlau M, Zou RY. The random forest algorithm for statistical learning. The Stata Journal: Promoting communications on statistics and Stata [Internet]. 2020 Mar 24;20(1):3–29. Available from: https://journals.sagepub.com/doi/10.1177/1536867X20909688
25. Zhu T. Analysis on the Applicability of the Random Forest. Journal of Physics: Conference Series [Internet]. 2020 Aug 1;1607(1):012123. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/1607/1/012123
26. Zhang H, Jiang L, Yu L. Class-specific attribute value weighting for Naive Bayes. Information Sciences [Internet]. 2020 Jan;508:260–74. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0020025519308217
27. Pajila PJB, Sheena BG, Gayathri A, Aswini J, Nalini M, R SS. A Comprehensive Survey on Naive Bayes Algorithm: Advantages, Limitations and Applications. In: 2023 4th International Conference on Smart Electronics and Communication (ICOSEC) [Internet]. IEEE; 2023. p. 1228–34. Available from: https://ieeexplore.ieee.org/document/10276274/

28. Saadatfar H, Khosravi S, Joloudari JH, Mosavi A, Shamshirband S. A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning. Mathematics [Internet]. 2020 Feb 20;8(2):286. Available from: https://www.mdpi.com/2227-7390/8/2/286
29. Salim AP, Laksitowening KA, Asror I. Time Series Prediction on College Graduation Using KNN Algorithm. In: 2020 8th International Conference on Information and Communication Technology (ICoICT) [Internet]. IEEE; 2020. p. 1–4. Available from: https://ieeexplore.ieee.org/document/9166238/
30. L. Von Werra, L. Tunstall, A. Thakur, A. S. Luccioni, T. Thrush, A. Piktus, F. Marty, N. Rajani, V. Mustar, H. Ngo, O. Sanseviero, M. Šaško, A. Villanova, Q. Lhoest, J. Chaumond, M. Mitchell, A. M. Rush, T. L. Von Werra, L. Tunstall, A. Thakur ASL. Processing: System Demonstrations. Proceedings of the The 2022 Conference on Empirical Methods in Natural Language. 2022;128–136.
31. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean Journal of Anesthesiology [Internet]. 2022 Feb 1;75(1):25–36. Available from: http://ekja.org/journal/view.php?doi=10.4097/kja.21209
32. Namdar K, Haider MA, Khalvati F. A Modified AUC for Training Convolutional Neural Networks: Taking Confidence Into Account. Frontiers in Artificial Intelligence [Internet]. 2021 Nov 30;4. Available from: https://www.frontiersin.org/articles/10.3389/frai.2021.582928/full
33. J. Li, L. Wang, X. Zhang, L. Liu, J. Li, M. F. Chan, J. Sui and RY. Biology*Physics. International Journal of Radiation Oncology. 2019;105:893.