Comparison of Naïve Bayes and Random Forest Methods for Diabetes Prediction

Winda Hasanah
Master of Information System Management
Faculty of Information System
Gunadarma University

Lulu Chaerani Munggaran
Master of Information System Management
Faculty of Information System
Gunadarma University

ABSTRACT
Diabetes is a chronic metabolic disorder in which blood sugar levels exceed normal limits. Riskesdas Ministry of Health in 2018 showed the prevalence of diabetes mellitus in Indonesia increased from 2013. Classification is one of the solutions to decrease the prevalence of diabetes in Indonesia. In this research, Classification is used to predict diabetes by building a classification model. The research steps are data collection, split the dataset into training data and test data, build a classification model using the Naïve Bayes and Random Forest methods, and evaluate the model. The results showed that the Random Forest method has the best performance with accuracy = 100%, error = 0%, precision = 1 and recall = 1. The best ratios in classifying the diabetes dataset are 70:30 and 90:10.

General Terms
Algorithms, Classification

Keywords
Classification, Naïve Bayes Algorithm, Random Forest Algorithm, Diabetes

1. INTRODUCTION
Diabetes mellitus is a chronic metabolic disorder caused by the pancreas not producing enough insulin or the body unable to use insulin effectively. This condition makes the body have blood sugar levels that exceed normal limits. The results of the Riskesdas for the Ministry of Health in 2018 showed the prevalence of diabetes mellitus in Indonesia based on a doctor's diagnosis for ages ≥ 15 years increased by 0.5% compared to Riskesdas in 2013. The prevalence of diabetes mellitus based on the results of blood sugar tests increased by 1.6% compared to 2013.

Based on this percentage, the government has to take early detection to decrease the number of people with diabetes. Classification is one of the solutions that can be used to predict the possibility of someone suffering from diabetes. The purpose of classification is to build a model that can differentiate between positive and negative. This model will be used to predict diabetes based on the symptoms shown. The classification method is Naïve Bayes, Random Forest, Decision Tree, SVM, C4.5, etc. In this research, the dataset was classified using the Naive Bayes and Random Forest Methods. Naive Bayes and the Random Forest method compared to find the best method with ratios 40:60, 50:50, 60:40, 70:30, 80:20, and 90:10.

2. LITERATURE REVIEW
Classification using Naive Bayes and Random Forest methods has been conducted by several researchers. Some of these research are discussed below.

Riski, Annisa [1] compared Decision Tree, Naïve Bayes, k-Nearest Neighbor, Random Forest, and Decision Stump method to predict heart disease. The research divided into training and test data using the 10-fold cross-validation method. Evaluation performance of classification method used the parametric t-test.

Aji and Betha [2] predicted rainfall using the Random Forest method to anticipate flooding. The dataset consists of 2188 data and 16 attributes. Data pre-processing was done by selecting attributes, cleaning data, and transforming data. The implementation of the Random Forest method was carried out using the K-Fold Cross Validation method (K = 10) and used all data.

Sri Widaningsih [3] compared the classification methods C4.5, Naïve Bayes, KNN, and SVM in predicting student graduation time. The research stages used the Discovery Knowledge of Database (KDD). The predictive variable in this research is gender and achievement index. Classification is divided into two classes, appropriated and unappropriated. Achievement index attribute transformed from numeric to categorical data (large, medium, and small). The research evaluation used Confusion Matix and the ROC curve.

Rahmaulidiah [4] compared the Naive Bayes and K-Nearest Neighbor methods for classification the status of value-added tax payments at KPP Samarinda Ulu. Payment status (Y) was divided into two classes, namely adherent and non-adherent. The independent variables (X) in this research are income, government agency, and tax reporting status. The type of the independent variable used is categorical. The research dataset was divided into 80% training data and 20% test data. Evaluation performance of classification model used APER.

Purnamawati et al [5] predicted the likelihood of early-stage diabetes using a classification method. Data pre-processing was done by using resample technique. Experiments were carried out using the Naïve Bayes method, SVM, and Random Fores with 10-fold cross-validation. Performance evaluation of the classification model used accuracy, f-measure, recall, precision, and ROC.

3. RESEARCH PROCESS
This section explains the research stage used to find the best method for diabetes prediction. The two methods used are Random Forest and Naïve Bayes. The research stage begins with collecting the dataset, splitting the dataset into training data and test data, building a model using Naive Bayes and Random Forest methods, and evaluating a model. Figure 1 illustrates this research methodology.

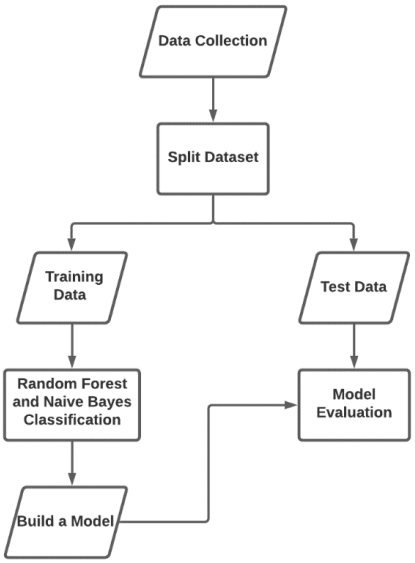

Fig 1: Research Methodology

3.1 Data Collection

This research uses a dataset from the UCI Machine Learning Repository. The data were obtained using questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh, and approved by a doctor. The data consists of 17 attributes with a total of 520 data. Details of the dataset attribute shown in Table 1.

Table 1. Research Dataset Attributes

Attribute	Description
Age	16-90
Gender	Male, Female
Polyuria	Yes, No
Polydipsia	Yes, No
Sudden Weight Loss	Yes, No
Weakness	Yes, No
Polyphagia	Yes, No
Genital Thrush	Yes, No
Visual Blurring	Yes, No
Itching	Yes, No
Irritability	Yes, No
Delayed Healing	Yes, No
Partial Paresis	Yes, No
Muscle Stiffness	Yes, No
Alopecia	Yes, No
Obesity	Yes, No
Class	Positive, Negative

3.2 Split Dataset

Dataset is divided into training data and test data. The ratio for split the dataset are 40:60, 50:50, 60:40, 70:30, 80:20, and 90:10. Details of the number of training data and test data are shown in Table 2.

Table 2. The Ratio of Training Data and Test Data

Ratio	Training Data	Test Data
40:60	208	312
50:50	260	260
60:40	312	208
70:30	364	156
80:20	416	104
90:10	468	52

3.3 Naïve Bayes Method

The classification stages of diabetes prediction using the Naïve Bayes method, that is calculating the prior probability for the possible positive and negative classes ($P(Ci)$), calculating the posterior probability X with terms Ci ($P(X|Ci)$), and calculating $P(X|Ci)P(Ci)$. The following is the flow of classification using the Naïve Bayes method.

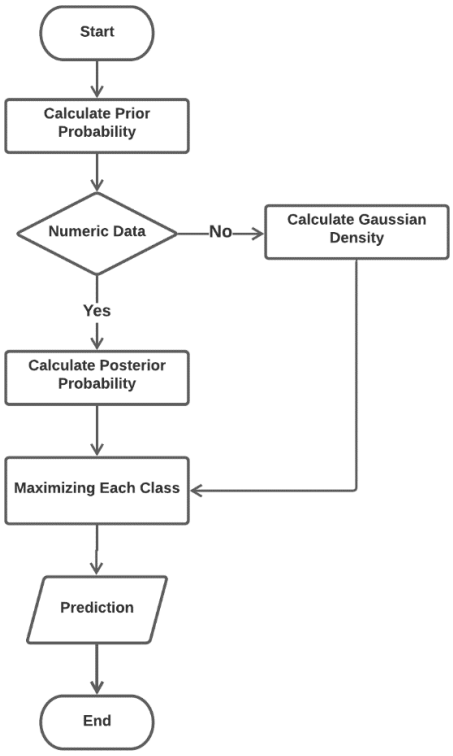

Fig 2: Naïve Bayes Method

3.3.1 Prior Probability

The calculation of the prior probability is divided into the prior probability for the negative class and the prior probability for the positive class. Here is the formula for calculating the prior probability.

$$P(Ci) = \frac{Si}{s} \quad (1)$$

Where:

Si = the number of training data from category Ci ($C0$ = negative class and $C1$ = positive class)

s = the total number of diabetes dataset

3.3.2 Posterior Probability

The research dataset has two types of data, that is categorical data and numerical data. Attributes that have categorical data types are gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genetic thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity. Example of calculating the posterior probability of a categorical attribute.

Table 3. Posterior Probability of Gender

Gender	Number of events		Probabilitas	
	Negative	Positive	Negatif	Positif
Male	67	52	67/73	52/135
Female	6	83	6/73	83/135
Sum	73	135	1	1

The attribute included in the numeric data is Age. The first step to calculating the age attribute is calculating the *mean*.

$$\mu = \frac{\Sigma age}{\Sigma all\ data} \quad (2)$$

The second step is calculating *standard deviation.*

$$\sigma^2 = \frac{\Sigma\,(Age - \mu)^2}{\Sigma(all\ data - 1)} \quad (3)$$

The last step is calculating *Gaussian Density*.

$$P(X|Ci) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Age-\mu)^2}{2\sigma^2}} \quad (4)$$

The following is the formula for calculating the posterior probability for all attributes.

$$P(C|X_1, \dots, X_n) = P(X_1|C) = \prod_{i=1}^{n} P(Xi|C) \quad (5)$$

Where:

$\prod_{i=1}^{n} P(Xi|C) =$ Multiplication of ratings between attributes

3.3.3 Maximizing Each Class

$$P(Ci|X) = P(X|Ci)\ x\ P(Ci) \quad (6)$$

Where:

$P(X|Ci) =$ Posterior Probability

$P(Ci) =$ Prior Probability.

3.4 Random Forest Method

The classification stages of diabetes prediction using the Random Forest method are determining the "n" tree, bagging from the dataset, calculating the Gini index to choose node, and voting to make a prediction. The following is the flow of classification using the Random Forest method.

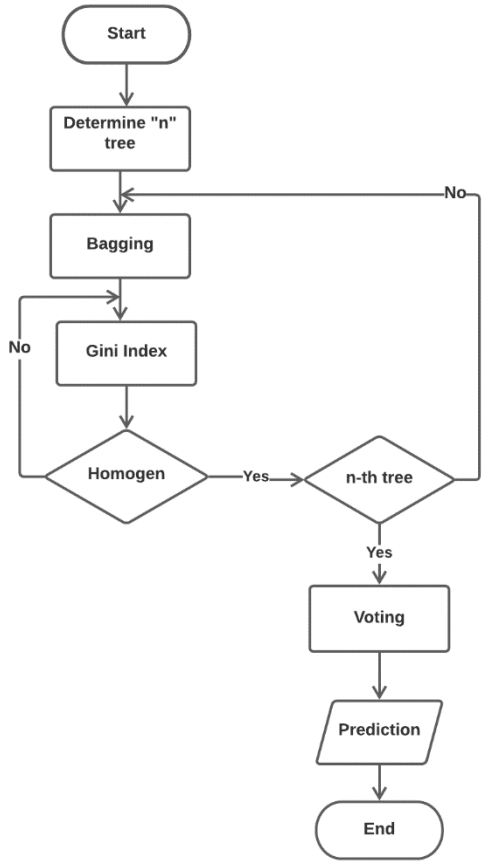

Fig 3: Random Forest Method

3.4.1 Determine n tree

The first stage is to determine the number of trees. The number of trees formed in this study was 100 decision trees.

3.4.2 Bagging Technique

The bagging technique is the selection of a random sample of data to construct a tree with replacements. The data sample used is the training data from the ratio of 40:60, 50:50, 60:40, 70:30, 80:20, and 90:10. Here is an example of a bootstrap sample with Polyuria, Polydipsia, and Polyphagia attributes.

Table 4. Example of The Diabetes Dataset.

No	Polyuria	Polydipsia	Polyphagia	Class
1	No	Yes	No	Positive
2	Yes	Yes	No	Positive
3	No	Yes	Yes	Negative
4	No	No	No	Negative
5	Yes	No	No	Negative

Table 5. The Result of Bagging Technique

No	Polyuria	Polydipsia	Polyphagia	Class
1	No	Yes	No	Positive
1	No	Yes	No	Positive
3	No	Yes	Yes	Negative

4	No	No	No	Negative
5	Yes	No	No	Negative

3.4.3 Gini Index

Gini index calculation is used to determine tree nodes and split the node into two child nodes. The data used to construct the tree is a bootstrap sample. The first step to determine the node of the tree is calculated gini index with the formula below.

$$Gini\ index(D) = 1 - \sum_{i=1}^{m} {P_i}^2 \quad (7)$$

Where:

Pi is probability of data labeled class i in D.If the data D divided into two subsets, the index of the data divided into class m[7].

The second step is calculated gini split with the formula below.

$$Gini_{split} = \sum \frac{n_i}{n} * Gini\ index(D) \quad (8)$$

Where:

$ni =$ the number of data partitions on the set
n = the number of the entire set.

3.4.4 Voting

The tree tested using test data. The prediction of diabetes uses the probability calculation of negative class and positive class from all the decision tree predictions.

3.5 Evaluation of Classification Model

The classification performance measure is described using a confusion matrix.

Table 6. *Confusion Matrix*

Prediction	Actual	
	Positive	**Negative**
Positive	True Positive	False Negative
Negative	False Positive	True Negative

From the confusion matrix above, accuracy, error, precision and recall calculations are performed.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (9)$$

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$Error = \frac{FP + FN}{TP + TN + FN + FP} \quad (12)$$

4. RESULT

This chapter describes an evaluation of the Naïve Bayes and Random Forest classification methods in predicting diabetes. Comparison between Naive Bayes and Random Forest is done to find the method with the best performance.

4.1 Naïve Bayes

The following are the results of the evaluation of the Naïve Bayes method using split data ratios of 40:60, 50:50, 60:40, 70:30, 80:20, and 90:10.

Table 7. Evaluation of Naïve Bayes Method

Ratio	Accuracy	Error	Precision	Recall
40:60	84.94%	15.06%	0.908	0.848
50:50	85%	15%	0.866	0.881
60:40	87.5%	12.5%	0.889	0.903
70:30	90.36%	9.62%	0.901	0.932
80:20	89.42%	10.58%	0.862	0.966
90:10	92.31%	7.69%	0.903	0.966

Ratio 90:10 has the highest accuracy with 92,308% and the lowest error with 7,692%. The highest precision value is a ratio of 40:60 with 0.908, followed by a ratio of 90:10 to 0.903. The highest recall value is ratios 90:10 and 80:20 with 0.966. Based on the comparison above, a ratio of 90:10 has the best performance in classifying diabetes datasets.

4.2 Random Forest

The following are the results of the evaluation of the Random Forest method using split data ratios of 40:60, 50:50, 60:40, 70:30, 80:20, and 90:10.

Table 8. Evaluation of Random Forest Method

Ratio	Accuracy	Error	Precision	Recall
40:60	93.91%	6.09%	0.944	0.958
50:50	95%	5%	0.982	0.942
60:40	97.12%	2.88%	0.96	0.992
70:30	100%	0%	1	1
80:20	97.12%	2.88%	0.969	0.984
90:10	100%	0%	1	1

Ratio 70:30 and 90:10 have the highest accuracy with 100%. The highest precision and recall are ratios 70:30 and 90:10 with precision and recall value equals 1. Based on the comparison above, ratios 70:30 and 90:10 have the best performance in classifying diabetes datasets.

4.3 Comparasion Naïve Bayes and Random Forest

Comparison of Naive Bayes and Random Forest methods divided into accuracy comparison, error comparison, precision comparison, and recall comparison. On accuracy, precision, and recall comparison, the method with the highest value has the best performance. Whereas on error comparison, the method with the lowest error value has the best performance. The following is a comparison of the Naïve Bayes and Random Forest methods.

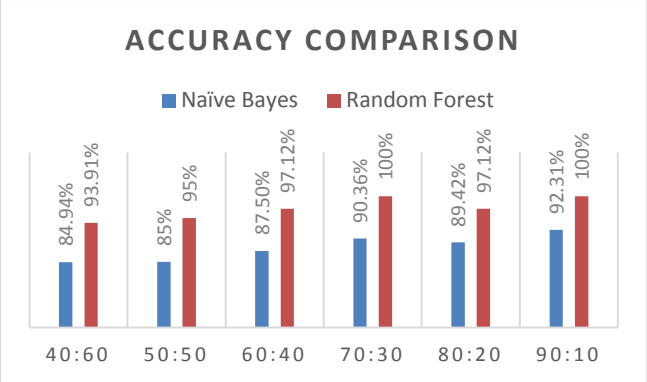

Fig 4: Accuracy Comparison

Based on the accuracy comparison's result, Random Forest has higher accuracy in all ratios than Naïve Bayes. The accuracy average for Random Forest is 97.190% and Naïve Bayes is 88.254%. Ratio 70:30 and ratio 90:10 have the highest accuracy in Random Forest Classification. Whereas on Naïve Bayes Classification, a ratio of 90:10 has the highest accuracy.

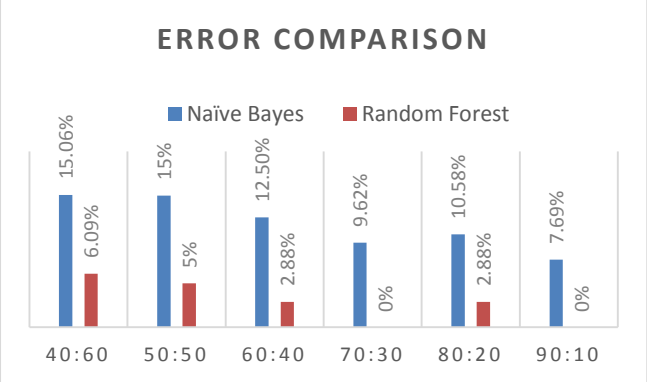

Fig 5: Error Comparison

Based on the error comparison's result, Random Forest has a lower error in all ratios than Naïve Bayes. Error average for Random Forest is 3,374% and Naive Bayes is 14,094%. Split data ratio 90:10 has the best performance for Naive Bayes and Random Forest methods. It has an error percentage of 7,69% and 0%. Split data ratio 40:60 has the highest error to classify the diabetes dataset.

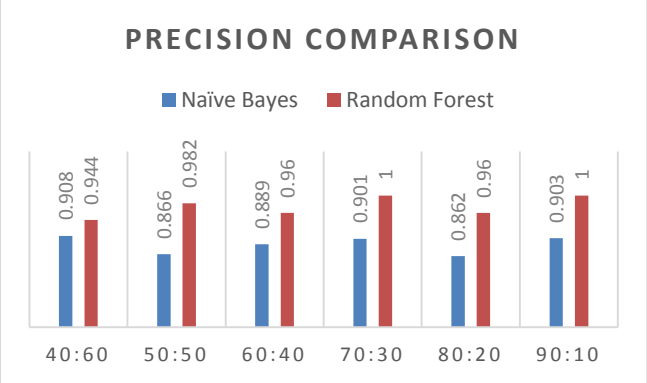

Fig 6: Precision Comparison

Based on the precision comparison's result, Random Forest has a higher score than Naïve Bayes in all ratios. The precision average of Random Forest is 0.976 and Naïve Bayes is 0.888. Ratio 90:10 and 70:30 has the highest score in classification using Random Forest. Whereas on Naïve Bayes, a ratio of 40:60 has the best precision score.

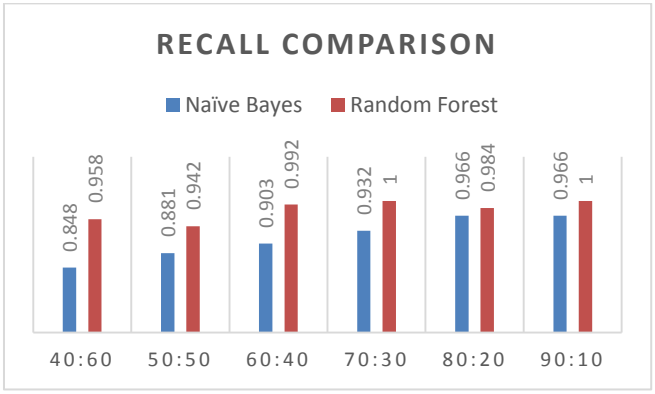

Fig 7: Recall Comparison

Based on the recall comparison's result, Random Forest has a higher score than Naive Bayes in all ratios. The recall average of Random Forest is 0.979 and Naïve Bayes is 0.916. Ratio 70:30 and ratio 90:10 have the highest recall score in Random Forest Classification. Whereas on Naïve Bayes Classification, ratio 80:20 and 90:10 has the highest recall score.

5. CONCLUSION

The diabetes dataset was successfully classified using Naïve Bayes and Random Forest methods. The ratio 70:30 and 90:10 has the best performance in the classification of diabetes dataset with Random Forest Method. In the Naïve Bayes, ratio 90:10 has the best performance in classifying the diabetes dataset. Based on the comparison of accuracy, error, precision, and recall, Random Forest has better performance than Naïve Bayes on each data ratio. The Random Forest classification has the highest accuracy value of 100%. Future studies can implement the data ratio 70:30 or 90:10 on different datasets for classification, and can also be implemented into an expert system for disease diagnosis.

6. REFERENCES

[1] Annisa, R. (2019). Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung. JTIK (Jurnal Teknik Informatika Kaputama), 3(1), 22-28.

[2] Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. Indonesian Journal of Artificial Intelligence and Data Mining, 1(1), 27-31.

[3] Widaningsih, S. (2019). Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4, 5, Naïve Bayes, Knn Dan Svm. Jurnal Tekno Insentif, 13(1), 16-25.

[4] Rahmaulidyah, F. N. (2020). Perbandingan Metode Klasifikasi Naive Bayes dan K-Nearest Neighbor pada Data Status Pembayaran Pajak Pertambahan Nilai di Kantor Pelayanan Pajak Pratama Samarinda Ulu (Doctoral dissertation, universitas mulawarman).

[5] Purnamawati, A., Nugroho, W., Putri, D., & Hidayat, W. F. (2020). Deteksi Penyakit Daun pada Tanaman Padi Menggunakan Algoritma Decision Tree, Random Forest, Naïve Bayes, SVM dan KNN. InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan, 5(1), 212-215.

[6] Effendi, M. T., Hidayat, N., & Dewi, R. K. (2019). Sistem Diagnosis Penyakit Tumbuhan Mangga Menggunakan Metode Naive Bayes. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X.

[7] Amiarrahman, M. R., & Handhika, T. (2018). Analisis dan implementasi algoritma klasifikasi Random Forest dalam pengenalan Bahasa Isyarat Indonesia (BISINDO). In Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi) (Vol. 2, No. 1, pp. 083-088).