IJDASEA
International Journal of Data Science
Engineering, and Analytics

# Diagnosis of Diabetes Using Naïve Bayes Classifier Method

**Tasya Ardhian Nisaa [1], Shavira Maya Ningrum [2] and Berlianda Adha Haque [3]**

[1] UPN "Veteran" Jawa Timur; 18081010049@student.upnjatim.ac.id
[2] UPN "Veteran" Jawa Timur; 18081010050@student.upnjatim.ac.id
[3] UPN "Veteran" Jawa Timur, 18081010051@student.upnjatim.ac.id
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials)

**Citation:** Nisaa, T.A.; Ningrum, S.M.; Haque, B.A. Diagnosis of Diabetes Using Naïve Bayes Classifier Method. *C* **2021**, Vol *1*, Page 22-29. https://doi.org/10.3390/xxxxx

Academic Editor: Tasya A. Nisaa

Received: 10 July, 2021
Accepted: 15 July, 2021
Published: 19 July, 2021

**Abstract:** Not a few people suffer from diabetes, diabetes is usually caused by genetic inheritance from parents and grandparents. Not only from heredity but many criteria or characteristics can determine a person has diabetes. This research was conducted by looking for a dataset on Kaggle that contains criteria for someone diagnosed or undiagnosed with diabetes such as age, gender, weakness, polyuria, polydipsia, and others. Furthermore, from these criteria, predictions are calculated using the Naive Bayes classification method where this method is one of the data mining techniques. This prediction calculation uses the Python programming language. From these criteria, each criterion is grouped with similarities and the results of the program that have been made can diagnose someone with diabetes. The prediction calculations that have been carried out have resulted in 90% accuracy, 93% precision, 89% recall, 92% specificity, and 91% F1-Score.

**Keywords:** Diabetes, Naïve Bayes Classifier, Data Mining

## 1. Introduction

Currently many diseases are often encountered by the community, one of which is diabetes, because diabetes does not look at age, old or young does not rule out the possibility of getting diabetes. Diabetes itself is a disease that can be characterized by human blood having increased sugar levels as a result of system disturbances in the body, namely the inappropriate production of the hormone insulin by the pancreas which causes its users to be less effective in the body [1], [2]. Often people assume alone just because the person has offspring from parents who are affected by diabetes and because they often feel tired, then conclude that they have diabetes.

Based on the previous problems, diabetes is very dangerous if they are carried out continuously because they will have very fatal consequences in the future. Because of these problems, the authors conducted research using the Naive Bayes classification to produce a fairly high accuracy value. The Naive Bayes classification algorithm predicts the criteria that have been given by combining the probability of a category/criterion with the probability of a word or grouping a category, class, or group based on previously determined characteristics where between categories, classes, or groups have [3], [4].

In conducting research, the author applies a method where the Naive Bayes classification method is one of the data mining techniques. Due to previous research conducted by [5], the Naive Bayes classification algorithm was applied to predict the level of malignancy of breast cancer and in this study, the accuracy was quite large, namely 97.82%. First of all, the author takes a dataset from Kaggle to predict a diabetes disease or not by looking at the categories/criteria/characteristics of someone who has diabetes in the dataset. Then enter the pre-processing stage where one of the attributes/categories in the dataset, namely age, is categorized into adolescents, adults, and seniors. Furthermore, in the main process stage, the author applies the naive Bayes classification algorithm by training the program that has been made with data training, namely 80% of the data taken from the previous Kaggle to find the probability of each attribute with a certain category. And the last stage is post-processing where the author conducts prediction experiments on data testing, namely 20% of the data taken from Kaggle, or the remaining 80% which was previously used as data training to calculate the values of accuracy, precision, recall, specificity, and F1-Score. as a reference value to assess the program is feasible or not to use.

## 2. Related Works

This section will describe research that has similarities in the field of classification using the Naïve Bayes method. Wibawa et all was applied Naïve Bayes classifier for journal classification of each part, more precisely to find the quartile. They used data around 1491 instances, and the results provided an accuracy. Basically, the Naïve Bayes theorem always combining between previous and new knowledge [10]. In the other hand, they assume that methods have high accuracy and simple algorithm.

Naïve Bayes is popular classification method. The algorithm is most straightforward ways to deal with the group. Reddy et all, using Naïve Bayes Classifier to solve clasify the question paper based on difficulty level prediction. The algorithm was implemented to classify based on a decision attribute such as 'Tough', 'Medium', and 'Easy'. They used these procedure to approach the results, and successfully encountered their problems. Furthermore, they provide Bayesian induction, of which the gullible Bayes classifier is an especially straightforward illustration, depends on the Bayes decide that relates restrictive and negligible probabilities [11].

Clustering data usually happened in Data Mining problems. Data mining in recent year is a significant research theory that will have a wide range of application in future [12]. Likewise Shareefunnisa et all., were used Naïve Bayes Algorithm in the Data Mining for Clustering Data. They implemented the algorithm using WEKA as data mining software tools. It can handling of missing data using approach technique of A Unique Category (AUC) for obtaining better performance, then exported the classifier model. The results provide an accuracy with increasing the data.

## 3. Experiment and Analysis

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

### 3.1. Diabetes

Diabetes is a disease caused by disturbances in the metabolic system characterized by the presence of blood glucose levels. The condition is a usually insufficient amount of insulin to transport glucose into cells. This causes blood glucose levels to increase [5]. Diabetes has long-term damaging effects, which can cause damage to the heart, blood vessels, eyes, kidneys, and nerves, this is the result of uncontrolled diabetes and is likely to be cured [2].

### 3.2. Naïve Bayes Classifier Algorithm

Naive Bayes classification algorithm is an algorithm that produces classification using a probability approach, where word probabilities are combined with category probabilities and then generate category possibilities for a given document [3]. The naive Bayes classification algorithm is commonly called Bayesian classification because it is a statistical classification method used to predict the probability of membership of a class. The naive Bayes classification algorithm is derived from Bayes' theorem whose classification is almost the same as the decision tree and neural network [6].

The naive Bayes classification algorithm is carried out using holdout evaluation techniques to build a model. Holdout is an evaluation technique to get the highest level of accuracy by dividing training data and testing data [7]. The Naive Bayes classification algorithm has a fairly high accuracy value when applied to a large database [6].

### 3.3. Data Mining

Data mining identifies related information from various large databases using statistical techniques, artificial intelligence, and machine learning [8]. The essence of the word data mining is a science that aims to find, mining, and extract knowledge from data [9]. According to Witten, data mining is a process to obtain knowledge or patterns from data sets [10].

Data mining according to Larose is divided into several groups that have various tasks, namely:

3.3.1. Description

Analysts or writers usually try to figure out how to describe new patterns and trends in data.

3.3.2. Classification

There is a target variable category. For example, classified in terms of age, namely infants, adolescents, and adults

3.3.3. Estimation

Not much different from classification, the difference is only in the estimated target variable which has a numerical value, not a category

3.3.4. Prediction

Prediction is a result that will appear in the future where the result has not happened or will happen.

3.3.5. Clustering

Clustering is looking for and sorting out one data with another whether it has similar characteristics, which will then be grouped into one.

3.3.6. Association

Association is identifying the relationship between several events that occur at one time [8], [9].

### 3.4. Method Diagram

The research method that the author uses has 3 stages, namely preparation, application, and evaluation.

3.4.1. Pre-processing

1. Literature Review

At this subsection, research references are collected on diabetes, the Naive Bayes classification algorithm, data mining, and other related reading materials in order to support the research.

2. Requirements Analysis

At this subsection, analysis and planning are carried out to determine the implementation process such as the need for a dataset to support this research.

### 3.4.2. Implementation

1. Algorithm Implementation using Python

At this subsection, the implementation process of the Naive Bayes classification algorithm is carried out on the diabetes dataset obtained from Kaggle using the Python programming language.

The block diagram where the block diagram contains the process of implementing the Naive Bayes classification algorithm on the diabetes dataset in detail, the following block diagram in detail can be seen in Figure 1.

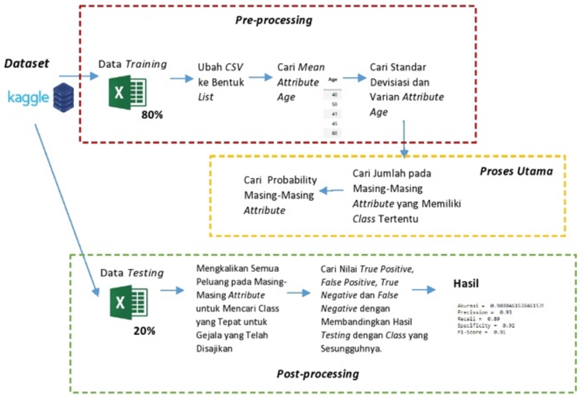

Pre-processing
Dataset
kaggle
Data Training
80%
Ubah CSV ke Bentuk List
Cari Mean Attribute Age
Age
40
50
41
45
60
Cari Standar Devisiasi dan Varian Attribute Age
Proses Utama
Cari Jumlah pada Masing-Masing Attribute yang Memiliki Class Tertentu
Cari Probability Masing-Masing Attribute
Data Testing
20%
Mengkalikan Semua Peluang pada Masing-Masing Attribute untuk Mencari Class yang Tepat untuk Gejala yang Telah Disajikan
Cari Nilai True Positive, False Positive, True Negative dan False Negative dengan Membandingkan Hasil Testing dengan Class yang Sesungguhnya.
Hasil
Akurasi = 0.903846153846153
Precission = 0.93
Recall = 0.89
Specificity = 0.92
F1-Score = 0.91
Post-processing

Figure 1. Block Diagram of Implementation in Naïve Bayes Classifier Algorithm

a. Pre-processing

This subsection is carried out to find the standard deviation of the variance and the mean to overcome continuous data. The steps from this pre-processing stage are first to take 80% of the data from the dataset, which data is training data after that change the CSV file into a list form, then find the mean of the Age attribute. The following is the formula to find the mean of the Age attribute, which can be seen in Figure 2.

```
mean_positive=umur_positive/jumlah_positiv
```

Figure 2. Mean Attribute Age Equation

And the last step is to find the standard deviation and variant of the Age attribute. The following is the formula for finding the mean of the standard deviation and variance of the Age attribute, which can be seen in Figure 3.

```
stdv_positive=math.sqrt(agepos_total/(jumlah_positive
-1))
stdv_negative=math.sqrt(ageneg_total/(jumlah_negativ
e-1))
```

Figure 3. Standard Deviation and Variance of Attribute Age

b. Main-processing

This stage uses the Naive Bayes classifier algorithm to train the program with training data that aims to find the probability of each attribute with a certain category based on Positive and Negative classes. The steps from this main process stage are to continue from the pre-processing stage steps by adding the search for the number of each attribute that has a certain class and looking for the probability of each attribute.

c. Pot-processing
This stage is evaluated by conducting prediction experiments on the testing data that has been separated from the training data, we use a comparison of 80% training data and 20% testing data which aims to calculate the values of accuracy, precision, recall, specificity, and F1-Score. The steps of this post-processing stage are the first to take 20% of the data from the dataset, where the data is testing data, then multiply all the opportunities for each attribute to find the right class for the symptoms that have been presented in the dataset, then look for true positive (tp), false positive (fp), true negative (tn) and false negative (fn) values by comparing the test results with the real class, the last one is the results. The following is the formula for finding the Accuracy, Precision, Recall, Specificity, and F1-Score values, which can be seen in Figure 4

Akurasi = (tp+tn) / (tp+tn+fp+fn)

Precission = tp / (tp+fp)

Recall = tp / (tp+fn)

Specificity = tn / (tn+fp)

F1_score = 2 * (recall*precission) / (recall+precission)

Figure 4. Accuracy Value Equation, Precision, Recall, Specificity, and F1-Score

3.4.3. Evaluation

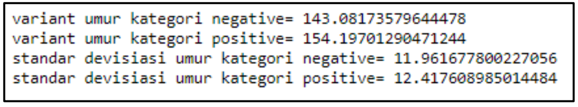
The result will be provided in the next subsection.

### 3.5. Results

In this subsection, the results will be discussed in detail about the results of each existing process. The following is the implementation of the Naive Bayes classification algorithm on the diabetes dataset obtained from Kaggle using the Python programming language.

3.5.1. Pre-processing

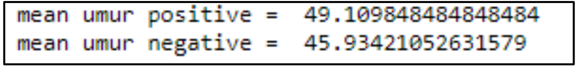
The pre-processing stage is carried out to overcome continuous data where the result of this stage is the discovery of standard deviations and variants of the age attribute in data training where the results can be seen in Figure 5.



```
variant umur kategori negative= 143.08173579644478
variant umur kategori positive= 154.19701290471244
standar devisiasi umur kategori negative= 11.961677800227056
standar devisiasi umur kategori positive= 12.417608985014484
```

Figure 5. Standard Deviation and Variance of Attribute Age

At the pre-processing stage, it also produces the mean/average of the age attribute in the training data where the results can be seen in Figure 6.

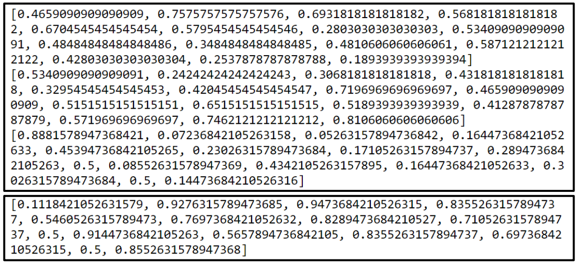


```
mean umur positive =  49.109848484848484
mean umur negative =  45.93421052631579
```

Figure 6. Mean Attribute Age

### 3.5.2. Main-processing

The main process stage is the application of the Naive Bayes classifier algorithm to find the probability of each attribute with a certain category based on the Positive and Negative classes on the training data where the results can be seen in Figure 7.



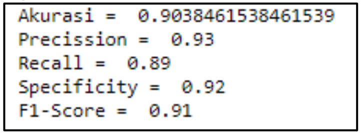
```
[0.4659090909090909, 0.7575757575757576, 0.6931818181818182, 0.568181818181818
2, 0.6704545454545454, 0.5795454545454546, 0.2803030303030303, 0.5340909090909090
91, 0.4848484848484848486, 0.3484848484848485, 0.4810606060606061, 0.587121212121
2122, 0.4280303030303030304, 0.2537878787878788, 0.1893939393939394]
[0.5340909090909091, 0.2424242424242424243, 0.3068181818181818, 0.4318181818181
8, 0.3295454545454545453, 0.4204545454545454547, 0.7196969696969697, 0.465909090909
0909, 0.5151515151515151, 0.6515151515151515, 0.5189393939393939, 0.41287878787
87879, 0.5719696969696970697, 0.7462121212121212, 0.8106060606060606]
[0.8881578947368421, 0.07236842105263158, 0.05263157894736842, 0.16447368421052
633, 0.45394736842105265, 0.23026315789473684, 0.17105263157894737, 0.289473684
2105263, 0.5, 0.08552631578947369, 0.4342105263157895, 0.16447368421052633, 0.3
026315789473684, 0.5, 0.14473684210526316]
```

```
[0.1118421052631579, 0.9276315789473685, 0.9473684210526315, 0.835526315789473
7, 0.5460526315789473, 0.7697368421052632, 0.8289473684210527, 0.71052631578947
37, 0.5, 0.9144736842105263, 0.5657894736842105, 0.8355263157894737, 0.69736842
10526315, 0.5, 0.8552631578947368]
```

Figure 7. Each Probability of Attribute with Certain Category based on Positive and Negative Class

### 3.5.3. Post-processing

The post-processing stage is evaluated by conducting prediction experiments on data testing that aims to calculate the values of accuracy, precision, recall, specificity, and F1-Score and the results can be seen in Figure 8.



```
Akurasi =  0.9038461538461539
Precission =  0.93
Recall =  0.89
Specificity =  0.92
F1-Score =  0.91
```

Figure. 8. The results of accuracy value, precision, recall, specificity, and F1-Score

## 4. Conclusions

From the discussion and program above, it can be concluded that the program can work quite well and is feasible to use with an accuracy value of 90%, precision 93%, recall 89%, specificity 92%, and F1-Score 91%. A high accuracy value indicates the program can predict close to accurate and high precision can avoid people who should be positive but predictably negative which can harm that person. For more accurate results, it may be better to use a comparison of 70% training data and 30% testing data.

## References

1. J. G. I. I. A, I. Harjosari, K. M. Amplas, K. Medan, and S. Utara, "UJI AKTIFITAS EKSTRAK DAUN BINAHONG ( Anredera cordifolia ( Ten .) Steenis ) TERHADAP KADAR GULA DARAH MENCIT Rani Ardiani * Haris Munandar Nasution Faisal Amin Tanjung * Fakultas Farmasi Universitas Muslim Nusantara ( UMN ) Al-Washliyah," pp. 484–490, 2020.
2. H. Lingkar, P. Terhadap, K. Gula, M. Tes, T. Glukosa, and O. Pada, "Hubungan lingkar perut terhadap kadar gula darah menggunakan tes toleransi glukosa oral pada remaja akhir," vol. 9, no. 12, pp. 14–20, 2020.
3. D. T. Yudistira, "Penentuan Klasifikasi Status Gizi Orang Dewasa Dengan Algoritma Naïve Bayes Classification ( Studi Kasus Puskesmas Jiken )," pp. 1–10, 2014.
4. H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Cholissodin, "Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 3, p. 180, 2017, doi: 10.25126/jtiik.201743251.
5. Y. V. Via, B. Nugroho, and A. Syafrizal, "Sistem Pendukung Keputusan Klasifikasi Tingkat Keganasan Kanker Payudara Dengan Metode Naive Bayes Classifier," SCAN-Jurnal Teknol. Inf. dan Komun., vol. 10, no. 2, pp. 64–65, 2015.

6. N. Nasution, K. Djahara, and A. Zamsuri, "Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes ( Studi Kasus : Fasilkom Unilak )," J. Fak. Ilmu Komput., vol. 1, no. 1, pp. 1–11, 2015.
7. Y. S. Nugroho, "DATA MINING MENGGUNAKAN ALGORITMA NAÏVE BAYES UNTUK KLASIFIKASI KELULUSAN MAHASISWA UNIVERSITAS DIAN NUSWANTORO," doi: 10.1016/0002-9343(83)90110-9.
8. M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," EECCIS, vol. 7, no. 1, 2013, doi: 10.24076/citec.2017v4i2.106.
9. K. Hastuti, "ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK PREDIKSI MAHASISWA NON AKTIF," Semantik, 2012, doi: 10.2307/j.ctv11hppt6.3.
10. Wibawa, A.P., et all. "Naïve Bayes Classifier for Journal Quartile Classification"., International Journal of Recent Contributions from Engineering, Science, and IT. 2020. https://doi.org/10.3991/ijes.v7i2.10659
11. Reddy, T.R., et all., "Difficulty Level Prediction of a Question Paper Using Naïve Bayes Classifier". Journal of Xi'an University of Architecture and Technology. 2020.
12. Shareefunnisa, S., et all. "A Naïve Bayes Algorithm of Data Mining Method for Clustering of Data". International Journal of Advanced Science and Technology. 2020. Vol 29 No 6, pp. 8021-8028.