Naïve bayes classifier models for cerebral infarction classification

SG Fitri¹, R Selsi², Z Rustam³, J Pandelaki⁴

¹,²,³ Department of Mathematics, Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia
⁴ Department of Radiology, Cipto Mangunkusumo National General Hospital, Dki Jakarta 10430, Indonesia
sari.gita@sci.ui.ac.id, ruhul.selsi@ui.ac.id, rustam@ui.ac.id, jacubp@gmail.com

Abstract. Cerebral infarction is a condition of tissue damage in the brain caused by inadequate oxygen supply caused by obstruction of the flow of regions to the area (ischemia). Brain ischemia is more common than hemorrhagic, but surgery is most often performed in hemorrhagic strokes. This condition is also called stroke infarction. In stroke infarction does not occur bleeding. Changes in brain blood vessel walls can be primary due to congenital or degenerative abnormalities and secondary processes caused by other processes such as inflammation, arteriosclerosis, hypertension, diabetes mellitus and many other processes, so the cause of stroke is very multifactorial. The final sign of cell damage due to ischemia is marked by the nucleus which becomes picnotic and fragmented. Research shows that stroke infarction can occur at the age of 15-55 years. To diagnose the presence or absence of cerebral infarction in the brain is not enough just to use a CT scan, therefore machine learning will also be used to diagnose the presence or absence of cerebral infarction in the brain. For this reason, the authors propose the Naïve Bayes Classifier method as a classification method that has good accuracy, good precision, good memory, and a good F1-score in calculating a patient whose brain has cerebral infarction or not. In this proposed method, Naïve Bayes Classifier is a probabilistic machine learning model used to classify. Naïve Bayes Classifier is a simple probability technique based on the Bayes theorem with the assumption of independence among predictors. In simple terms, Naïve Bayes Classifier assumes that the presence of certain features in a class is not related to the presence of other features. This method can achieve an accuracy value of up to 92.43%, so this method can be an efficient classification tool.

1. Introduction

Stroke is one of the leading causes of disability and death worldwide, and there are large variations in risk factors and outcomes in different populations [1]. Traumatic brain injuries, such as strokes, are usually single events [2]. Basic Health Research, the Indonesian Ministry of Health in 2007 showed that stroke is the leading cause of death in hospitals in Indonesia.

Stroke is one of the non-communicable diseases which is still an important health problem in Indonesia. Along with the increasing morbidity and mortality at the same time, where in Indonesia the increase in cases can negatively impact the economy and productivity of the nation, because stroke treatment requires a long time and requires large costs [4].

Stroke is caused by blocked blood vessel cells in the brain, so the brain lacks oxygen and nutrients because the blood cells that carry these important things are stopped, causing them to die [3]. Stroke is
caused by several factors, that are hypertension, smoking, unhealthy diet, lack of physical activity, high blood pressure, increased blood sugar and increasing blood lipid profile [4].

Hypertension is a major cause of stroke. Increased blood pressure will cause the cerebral vessels to constrict [5]. Stroke is divided into 2 parts, ischemic stroke and hemorrhagic stroke [4]. Hemorrhagic stroke is a stroke caused by leakage or rupture of blood vessels in the brain, causing bleeding in the brain [6].

Whereas ischemic stroke is caused by narrowing or clogging of arteries to the brain, causing reduced blood flow (ischemia) [6]. About 80% of strokes are ischemic strokes. This is because ischemic strokes are more common than hemorrhagic strokes [7].

Risk factors that cause ischemic stroke include age over 50 years, increased cholesterol resulting in narrowing of blood vessels, increased blood sugar, and hypertension that affects brain perfusion pressure [7]. In ischemic stroke, cerebral infarction is a more common condition [8].

Stroke can cause interference with some parts of the brain, while other parts of the brain can work normally, depending on the part of the brain affected by a stroke, how often a stroke occurs and age. Some of the most common strokes are paralysis of one side of the body that is most common, impaired vision that can affect one or both eyes, aphasia, difficulty speaking or understanding speech, impaired perception, fatigue, depression, unstable emotions, impaired memory, and changes personality [9].

Strokes affect everyone differently, some of them take years to recover. Recovery from a stroke involves changes in physical, social and emotional aspects. Stroke treatment usually starts with treatment at the hospital. Stroke rehabilitation requires good coordination between patients, families, doctors, nurses, physical therapists, psychologists, and others. Certain aspects of stroke rehabilitation are well established in clinical practice and are standard of care, for example providing physical therapy for early stroke patients [7].

Stroke patients who are taken to the hospital using an ambulance can be diagnosed and treated faster than patients who are taken to the hospital not by using an ambulance. Because emergency treatment starts from the journey to the hospital. Emergency workers will provide information that guides care and alert medical staff at the hospital before the patient arrives at the hospital, giving them time to prepare [10].

Before the patient is discharged from the hospital, the patient will be rehabilitated beforehand, to help the patient recover better. In addition, collaboration between the patient, the patient's family and medical staff is also very much needed to find out the cause of stroke in the patient and what steps must be taken to prevent a stroke from returning later [10]. By doing a CT scan of the brain, the presence or absence of infarction in the patient's brain can be observed [11]. But, to diagnose cerebral infarction in a person's brain, it is not enough just to have a CT scan, but it also requires machine learning. Machine learning can be used to classify various diseases, such as sinusitis, breast cancer, and hepatocellular carcinoma (HCC) [12,13,14]. Machine learning has also been used in classifying data on cancer, brain cancer, and high dimensional breast cancer, besides that machine learning can also be used for insolvency prediction in insurance companies [15,16,17,18]. However, in this study we will classify infarction in the brain.

To classify infarction in the brain, we propose the Naïve Bayes Classifier method as a classification method that has good accuracy, good precision, good memory, and a good F1-score in calcifying a patient whose brain has cerebral infarction or not. In this proposed method, Naïve Bayes Classifier is a probabilistic machine learning model used to classify. Naïve Bayes Classifier is a simple probability technique based on the Bayes theorem with the assumption of independence among predictors. In simple terms, Naïve Bayes Classifier assumes that the presence of certain features in a class is not related to the presence of other features [19]. This method can achieve an accuracy value of up to 92.43%, so this method can be an efficient classification tool.
In this study, we examined unbalanced data class samples from hospitals related to cerebral infarction, with 103 majority data and 53 minority data. The main objective of this research is to improve the performance of the classification of machine learning algorithms for prediction of minority and majority classes. The main objective of this study is to improve the performance of the classification of machine learning algorithms for prediction of minority and majority classes, with the Naïve Bayes Classifier using resampling techniques in training data. This method achieves an accuracy of 92.43%.

2. Method

2.1 Naïve Bayes

Bayesian Networks (BNs) introduced by Pearl (1988) are high-level representations of probability distributions over a set of variables $X = \{x_1, x_2, \ldots, x_n\}$ [20].

Naïve Bayes is one of the most efficient and effective inductive learning algorithms for machine learning [21]. Naïve Bayes Classifier is known as a better classification method compared to other classification methods. Because, the main characteristic of Naïve Bayes is a very strong assumption of independence, then the model is simple and easy to make [22].

Specifically, the combined probability distribution for $X$ is given by [20]:

$$P(X) = \prod_{i=1}^{n} P(X_i | P_a(X_i)).$$

(1)

Let denote, $C$ is the observation class $X$, to predict the observation class $X$ using the Bayes rule [20]:

$$P(C | X) = \frac{P(C | P) P(X | C)}{P(X)}$$

(2)

Note:

- $X$: is attributes
- $C$: is class
- $P(C \lor X)$: is the probability of even $C$ given $X$ has occurred
- $P(X \lor C)$: is the probability of even $X$ given $C$ has occurred
- $P(C)$: is the probability of event $C$
- $P(X)$: is the probability of event $X$
Figure 1. A Naïve Bayes

\( X \) is classified as class \( C = \{+\} \) if and only if [23]:

\[
\hat{f}_b(X) = \frac{P(C = + | X)}{P(C = - | X)} \geq 1
\]

(3)

Where \( \hat{f}_b(X) \) is Bayesian classifier.

By substituting \( X = \{x_1, x_2, \ldots, x_n\} \) to equation (2), we get [21]:

\[
P(C | x_1, x_2, \ldots, x_n) = \frac{P(C | x_1, x_2, \ldots, x_n | C)}{P(x_1, x_2, \ldots, x_n)}
\]

(4)

We can decipher \( P(C | x_1, x_2, \ldots, x_n) \) into:

\[
P(C | x_1, x_2, \ldots, x_n) = P(C) \prod_{i=1}^{n} P(x_i | C)
\]

(5)

It can be seen that, equation (5) there are complex factors which make it difficult to calculate the analysis one by one. Thus, the assumption is needed that each factor in equation (5) is free from each other, namely \( P(C | x_1, x_2, \ldots, x_n) \), so that equation (5) can be written as:

\[
P(C | x_1, x_2, \ldots, x_n) = P(C) \prod_{i=1}^{n} P(x_i | C)
\]

(6)

So equation (4) can be written as [20]:

\[
P(C | X) = \frac{P(C) \prod_{i=1}^{n} P(x_i | C)}{P(X)}
\]

(7)

In Naïve Bayes Classifiers models we need to maximize the value of probability of each class, which is called the Maximum A Posteriori Hypothesis (HMAP) [19]:

\[
H_{\text{MAP}} = \arg\max P(C | x_1, x_2, \ldots, x_n) = \arg\max P(C) \prod_{i=1}^{n} P(x_i | C)
\]

(8)
In Naïve Bayes Classifier, we can predict which classes can be used in Naïve Bayes Model. But, if the attribute X in Equation (8) has quantitative types then the probability will be very small such that the value \( P(X \cup C) \) cannot be used to find the value of \( H_{MAP} \). So we need to use other approach such as normal (Gaussian) distribution that is (Rohith, 2018)[19].

\[
P( X_i = x_i | C = c_j ) = \frac{1}{\sqrt{2\pi \sigma_{ij}}} \exp \left( -\frac{(X_i - \mu_{ij})^2}{2 \sigma_{ij}^2} \right) \tag{9}
\]

2.2 Confusion matrix

The confusion matrix is used to measure the performance of a classification algorithm. The matrix can be easily understood, although the terminology associated with mixed matrices is a bit confusing (see table 1).

| Actual | Predicted | True Positive (TP) | False Positive (FP) |
|--------|-----------|--------------------|---------------------|
|        |           | False Negative (FN)| True Negative (TN)  |

Note:
- TP: Correct Positive: Infarction is predicted, and there is an infarction
- FP: False Positive: Infarction is predicted, and there is no infarction
- FN: False Negative: There is no predictable infarction, and infarction exists
- TN: True Negative: There is no predictable infarction, and no infarction

From the confusion matrix, we can find accuracy, accuracy, memory and score f1, with each formula being [24]:

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
Precision = \frac{TP}{TP + FP}
\]

\[
Recall = \frac{TP}{TP + FN}
\]

\[
f_1-score = 2 \cdot \frac{precision \cdot recall}{precision + recall}
\]

3. Experiment

The following data are data from ischemic stroke patients with cerebral infarction in their brain. Data taken from Cipto Mangunkusumo Hospital, Indonesia can be used to build the model of Naïve Bayes Classifier (See Table 2).

| Area | Min | Max | Average | SD | Sum | Length |
|------|-----|-----|---------|----|-----|--------|
| 0.2  | -3  | 38  | 16.88   | 9.3| 5166| 2      |
Table 3. The Feature of Cerebral Infarction Dataset

| No | Feature | Definition of feature |
|----|---------|-----------------------|
| 1  | Area    | Size of area from the infarction point |
| 2  | Min     | Minimum value of infarction |
| 3  | Max     | Maximum value of infarction |
| 4  | Average | Average value of infarction |
| 5  | SD      | Standard error value of infarction |
| 6  | Sum     | Sum value of infarction point |
| 7  | Length  | Length of infarction point |

Where there are 156 data with 7 features proportional to 70% as training data and 30% testing data from the original data, with the actual number 103 main data showing data classes with no infarction and 53 minor data indicating infarction. Table 3 shows an explanation of the infarction data features examined (See Table 3).

4. Result and Discussion

By using the Naïve Bayes algorithm and using python, we can get accuracy results with the training data as follows (See Table 4).

Table 4. Accuracy Value of Naïve Bayes Classifier

| % Data Training | % Accuracy |
|-----------------|-----------|
| 10              | 92.43     |
| 20              | 91.46     |
| 30              | 88.20     |
| 40              | 91.05     |

…
From this table, we can see that the best accuracy value is 92.43% with 10% training data. By using the confusion matrix, we can find the performance of the Naïve Bayes classification system (See Table 5).

Table 5. Confusion Matrix of Naïve Bayes

| Predicted | Positive | Negative |
|-----------|----------|----------|
| Actual    |          |          |
| Positive  | 48.65    | 4.32     |
| Negative  | 3.24     | 43.78    |

From Table 5 it can be seen that from 51.89% of actual positive samples (first column), it is estimated that 48.65% of them are positive with cerebral infarcts in the brain and 3.24% negative with cerebral infarction in the brain. From 48.10% of actual samples negative (second column), it is estimated that 4.32% of them are positive with cerebral infarction in the brain and 43.78% negative for cerebral infarction in the brain.

From these results, we can calculate the value of accuracy, precision, recall, and f1-score with 10% training data (See Table 6).

Table 6. Matrix classification of Naïve Bayes Classifier

|       | Accuracy | Precision | Recall | f₁-score |
|-------|----------|-----------|--------|----------|
|       | 0.9243   | 0.92      | 0.92   | 0.92     |

Table 6 shows the accuracy of Naïve Bayes Classifier in predicting whether or not someone has an infarction in his brain. It can be seen that the accuracy of Naïve Bayes Classifier is 92.43%, the precision of Naïve Bayes Classifier is 92%, recall is 92% and the f1-score is 92%.

5. Conclusion

Based on the experiments, it can be seen that using 10% training data produces an accuracy value of 92.43%. The method used in this paper achieves the classification results of the presence or absence of cerebral infarction in a person's brain with good accuracy. The weakness of this method is that the assumption of independence between attributes reduces accuracy, because there are several pieces of data where the attributes are bound to each other.

Acknowledgement

This research supported financially by the Ministry of Research and Higher Education Republic of Indonesia with PDUPT 2019 research grant scheme, ID number 1621/UN2.R3.1/HKP.05.00/2019

References

[1] Rachel D, Karthik M, Sudhakar V N, Kumar D, Serene I C, Ruijun C, Brian W, Isuru R 2014 Most Important Outcomes Research Papers on Stroke and Transient Ischemic Attack (Aha Journal vol.7) No.1
[2] Wei D, Bernard Y, Beth P J, Lynette M, Stephen D, Joanne F, Owen B W Ischemic Stroke: the Ocular Motor System As a Sensitive Marker for Motor and Cognitive Recorvey. [Internet]. [Accesses on 2019]. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3582066/

[3] Rida D, Nurdiina, Tina H N 2017 Analysis of Predictive Factors of Mortality in Hemorrhagic Stroke Patients at Soebadi Hospital Jember (NurseLine Journal vol.2) No.2

[4] Stroke: A Stroke Survivor Essay. [Internet]. [Accesses on 2019]. Available at https://www.123helpme.com/stroke­a­stroke­survivor­preview.asp?id=664060

[5] Zuryati T Q and Adiyo W Stroke 2016 Hemoragik e.c Hipertensi grade II Vol.5.No.2

[6] Stroke-Symptoms and Causes. [Internet]. [Accesses on 2019]. Available at https://www.mayoclinic.org/diseases­conditions/stroke/symptoms­causes/syc­20350113

[7] Dian H and Dwi D 2018 Gambaran Drug Related Problems (DRP’S) pada Penatalaksanaan Pasien Stroke Hemoragik dan Stroke Non Hemoragik di RSUD DR M Yunus Bengkulu (Jurnal Farmasi dan Ilmu Kefarmasian Indonesia vol.5) No.1

[8] Zuherman R, Dea A U, Jacub P and Widyo A N 2019 Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalance Cerebral Infarction Dataset Indonesia vol.9.No.2

[9] Aldy S R Stroke: Sekilas Tentang Definisi, Penyebab, Efek dan Faktor Risiko (Medan: Departemen FK­USU/RSUP H)

[10] Stroke Treatment. Centers for Disease Control and Prevention. [Internet]. [Accesses on 2019]. Available at https://www.cdc.gov/stroke/treatments.htm

[11] Amanda Rizki Bagasta et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052016

[12] Rezki Aulia Putri et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052060

[13] Zuherman Rustam and Sri Hartini 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052067

[14] Rafiqtul Khairi et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052038

[15] Rachman A A, Rustam Z 2016 Cancer Classification using Fuzzy C-Means with Feature Selection 12th International Conference on Mathematics, Statistics, and Their Applications, ICMSA 2016, AAC Dayan Dawood, Syiah Kuala University Banda Aceh; Indonesia

[16] Panca V, Rustam Z 2017 Application of Machine Learning on Brain Cancer Multiclass Classification AIP Conference Proceedings Vol.1862

[17] Lestari A W, Rustam Z 2017 Normed Kernel Function-based Fuzzy Probabilistic C-Means (NKFPFCM) Algorithm for High Dimensional Cancer Database Classification with Feature Selection is Based on Laplacian Score AIP Conference Proceedings Vol.1862

[18] Rustam Z, Yaurita F 2018 Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means Journal og Physics: Conference Series Vol.1028, Issue 1

[19] Nafizatus Salmi and Zuherman Rustam 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052068

[20] Sona T and Musa M 2013 Learning the Naive Bayes Classifier with Optimization Models Int. J. Appl. Math. Comput. Sci. Vol.23.N0.4

[21] H Zhang The Optimally of Naive Bayes Canada

[22] Jiawei H, Micheline K, Jian P 2012 Data Mining Concepts and Techniques (3rd ed). USA: Elsevier Inc.

[23] Pouria K and Sunita D 2017 Short Survey on Naive Bayes Algorithm Scientific Journal of Impact Factor (SJIF): 4.72 Vol.4.No.11.

[24] Durrabida Zahras et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052089