THE ANALYSIS PERFORMANCE METHOD NAIVE BAYES AND SSVM DETERMINE PATTERN GROUPS OF DISEASE

Rianto Sitanggang¹, Tulus², Zakarias Situmorang³

¹²³Department of Computer Science and Information Technology University of North Sumatra Medan, Indonesia

rianto_sitanggang@students.usu.ac.id

Abstract. Information is a very important element and into the daily needs of the moment, to get precise and accurate information is not easy, this research can help decision makers and make a comparison. Researchers perform data mining techniques to analyze the performance of methods and algorithms naïve Bayes methods Smooth Support Vector Machine (ssvm) in the grouping of the disease. The pattern of disease that is often suffered by people in the group can be in the detection area of the collection of information contained in the medical record. Medical records have information about disease by patients in coded according to standard WHO. Processing of medical record data to find patterns of this group of diseases that often occur in this community take the attribute address, sex, type of disease, and age. Determining the next analysis is grouping of four entities attribute. From the results of research conducted on the dataset fever diabetes mellitus, naïve Bayes method produces an average value of 99% and an accuracy and SSVM method produces an average value of 93% accuracy.

Keywords: Method naïve Bayes, SSVM method, Diabetes Mellitus

1. Background
Data mining is a series of processes for adding additional value of a set of data in the form of knowledge that had been unknown to them manually. Keep in mind that the word mining itself means little effort to get valuable data from a large number of basic data. Because the data mining actually has along root of disciplines such as artificial intelligence (artificial intelligent), machine learning, statistics and databases. Some techniques that are often mentioned in the literature among others, the data mining association rule mining, clustering, classification, neural networks, and others. Piles of data that exist both in the health service and the hospital currently only limited to providing graphs or statistics on the number of patients seeking treatment with disease in misery along with the report of the patient's departure. Reports from this data that is currently used by health authorities to carry out the policies when will provide counseling to the public.

1.2. Formulation of the problem
With reference to the background of the above problems, the issues discussed in this study is to analyze the naïve Bayes methods in getting polakelopok disease classification techniques dataminning smooth support vector machine (ssvm).

1.3. Limitations of the study
The boundary problem in this research are:
   a. Data on the disease General Hospital Sari Mutiara with interfal time data collection in the medical record by 2016.
   b. Disease information being sought is a grouping of diseases by age grouping and the grouping area.

1.4. Research purposes
This study has the objective to apply naïve Bayes methods in searching for the information contained in a stack of medical records

1.5 Benefits Research
The benefits derived from this study are as follows:

a. Assist medical personnel in analyzing the pattern of the disease to the medical record data quickly and accurately.

b. Become a comparison to researchers who focus on Data mining particular field of medicine.

2. LITERATURE REVIEW

2.1. Understanding Data Mining

In a simple data mining is mining or the discovery of new information by looking for patterns or particular rules of a number of very large data (Davies, 2004). Data mining is also referred to as a series of processes for adding additional value in the form of knowledge that had been unknown to them manually from a data set (Pramudiono, 2007). Data mining, often referred to as knowledge discovery in databases (KDD). KDD is an activity that includes the collection, use of data, historical to find regularities, patterns or relationships in large data sets (Santoso, 2007).

2.2 Naïve Bayes

Naïve Bayes is a method that has no rules. Naïve Bayes using a branch of mathematics known as the theory of probability to find the greatest opportunities of the possibilities of classification, by looking at the frequency of each classification on the training data. Naïve Bayes classification is a statistical classification that can be used to predict the probability of membership of a class. Classification bayesian is based on Bayes' theorem, named for the mathematician who is also minister Prebysterian England, Thomas Bayes (1702-1761) [8].

Bayesian classification has the capability similar to that classification decision tree and neural network [17]. Bayes rule is used to calculate the probability of a class. Naïve algorithm Bayes memberikan away to combine the opportunities foregone on condition likely to be a formula that can be used to calculate the chance of each possibility occurring. The general form of the Bayes theorem as follows:

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

Where:
- $X$ : Data with unknown class
- $H$ : Hypothesis $X$ data is a specific class.
- $P (H | X)$: The probability of the hypothesis $H$ based on the conditions $X$ (a posteriori probability)
- $P (H)$: The probability of the hypothesis $H$ (prior probability)
- $P (X | H)$: Probability $X$ based on the conditions on the hypothesis $H$
- $P (X)$: The probability of $X$

Naïve Bayes is a simplified method of Bayes. Bayes theorem simplifies to:

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

2.3 Definition of SSVM (Smooth Support Vector Machine).

SSVM is a new development of SVM with kernel and non-linear functions for classification analysis using smoothing methods. SVM was first introduced by Boser, Guyon and Vapnik in 1992 as a series of harmonic concepts featured in the statistical learning theory. The basic principle of SVM is a linear classifier and further developed in order to work on a non-linear problem by incorporating the concept of kernel trick on high-dimensional workspace (Vapnik, 1995).

Standard SVM to this problem is given by the following quadratic program:

$$\min_{\omega \in \mathbb{R}^{d+1} \cup H} \left\{ D(\omega) + y \right\}$$

Where:
- $v$ = Sebuah bervektor positif
- $y$ = variable Slack
- $e$ = Kolom vectors atudimensi

In SSVM approach (Lee et al., 2001), modified SVM problem is generated as follows:
Thus, we can replace $y$ in constraints with and convert into SVM SVM similar problem that an optimization problem without constraints as follows:

$$
\min_{(u, \nu, \gamma)} \frac{1}{2} w^T \gamma + \frac{1}{2} \gamma^T \left( w' w + \gamma^2 \right)
$$

This function with smoothing parameter used in replace function hereto plus to get Smooth VectorMachine (SSVM)

$$
\min_{(u, \nu, \gamma) \in \mathbb{R}^{n+1}} \frac{1}{2} \| e - D(Dw - e\gamma) \|_2^2 + \frac{1}{2} (u^u + \gamma^2)
$$

Just as before, it is obtained SSVM for an integral problem:

$$
\min_{(u, \nu, \gamma) \in \mathbb{R}^{n+1}} \frac{1}{2} \| e - D(Dw - e\gamma) \|_2^2 + \frac{1}{2} (u^u + \gamma^2)
$$

### 2.4. Klasifikasi
Classification is the process of finding a set of models / functions that define and distinguish the data into certain classes, with the aim of using the model to determine the class of an unknown object class (Han, et al 2000). There are two processes in classification, namely:

a. The process of learning or training do the construction model using training data. In this study, using the model of Smooth Support Vector Machine. At Smooth Support Vector Machine (SSVM), this is done by labeling each feature value

b. The process of testing conducted tests on the data of testing using a model that has been obtained from the training process. The use of the model to classify new data. Here, a records are fed into the model, and the model will provide answers grade calculation results. Accuracy (%) = $x \times 100\%$ (2.7)

### 3. RESEARCH METHODOLOGY
#### 3.1 Diagram Research
The diagram of the overall study can be seen in Figure 1.

![Figure 1. Diagram Research](image)

#### 3.2 Data Collection
Medical record data to be processed is RSU Sari Mutiara Medan patient medical records. period in 2016. Before the process of data mining, medical records obtained from the data base hospital performed the preprocessing stage. Preprocessing mechanism at this stage of integration and transformation has been collecting data from the tables contained in the formal application in hospitals into the parent table.

#### 3.3. Data cleaning
The initial stage of data mining preprocessing are cleaning the data, this is done to discard data that have incomplete information. It is possible that a lot of data have incomplete information, such as the exclusion of the information age, information about the area, information about sex, disease grouping information, as shown in

3.4 Selection of Data
Data that already have complete information in any subsequent attribute such data be selected. Selection is done to classify the attribute in accordance with the information needed. From attribute in the selection is done at the address attribute, attribute age. While the gender attribute is the target attribute. In this study, the data selection is done on the data that has information area in the city field. Because hospitals are concerned also receive patients from outside the field, the data containing information about the area outside the city of Medan will be discarded.

4. RESULTS AND DISCUSSION
4.1. Application of Naive Bayes Method
Naive Bayes is based on the simplifying assumption that the attribute values are conditionally independent if given a value of output. In other words, given the value of output, the probability of observing together is the product of the individual probabilities. The advantage of using Naive Bayes is that this method only requires the amount of training data (training data) are small to determine the required parameter estimation in the classification process.

4.2 Read Data Training
To specify the data that will be analyzed by the method of Naive Bayes then the first step is to read training data. The training data used can be seen in Figure 1.

4.1.2 Testing Methods Naive Bayes
From the above probability value will be tested as much data as 60 data and solved by using the tools so that the resulting classification results weka frequent users such as in Figure 2.

![Figure 2](image)

Figure 2. Results of Treatment Naïve Bayes

4.3 Method of Testing Results SSVM
The test results of 10 experiments above can be displayed in the table below. The tables show the percentage of accuracy values for each training and testing. Testing and training value obtained from the calculation as follows:

For training:

\[
\text{Accuracy rate (\%) = } \frac{\text{jumlah pola yang ditemui}}{\text{jumlah pola pelatihan}} \times 100\%
\]

For testing:
Table 1. Percentage of Training and Testing Data In 10 Times Trial

| Percobaan | X1 | X2 | X3 | Akurasi Training (%) | Akurasi Testing (%) |
|-----------|----|----|----|----------------------|---------------------|
| Fold01    | 14 | 18 | 23 | 99.63                | 90.00               |
| Fold02    | 21 | 20 | 10 | 99.63                | 96.67               |
| Fold03    | 30 | 16 | 4  | 99.63                | 90.00               |
| Fold04    | 26 | 6  | 32 | 100.00               | 96.67               |
| Fold05    | 8  | 3  | 17 | 99.63                | 93.33               |
| Fold06    | 1  | 7  | 29 | 100.00               | 93.33               |
| Fold07    | 5  | 22 | 23 | 99.63                | 93.33               |
| Fold08    | 24 | 27 | 19 | 99.63                | 96.67               |
| Fold09    | 12 | 9  | 25 | 99.63                | 93.33               |
| Fold10    | 15 | 11 | 13 | 99.63                | 92.33               |
| Rata-rata |    |    |    | 99.70                | 93.67               |

5. CONCLUSIONS

5.1. Conclusion
The conclusion that can be drawn from this study are as follows:
1. The process of grouping with SSVM method of obtaining the average - average accuracy of the training as much as 99.70% and the average - average test as much as 93.67%.
2. The process of grouping with Naive Bayes algorithm obtain the average - average accuracy as much as 99.40%.
3. SSVM method provides better accuracy than the Naïve Bayes method in the clustering of the disease.

BIBLIOGRAPHY
[1] Andari, S., 2012, Smooth Support Vector Machine and Multivariate Adaptive Regression Splines ToDiagnose Breast Cancer, Thesis, Faculty of Mathematics and Statistics Department Student ITS, Surabaya.
[2] Erdogan, SZ  and M. Timor. 2005. A Data Mining Application In A Student Database. Journal OfAeronautics And Space Technologies. Volume 2 Number 2: 53-57.
[3] Furqan, M., Embong, A. (2009). Smooth Support Vector Machine For FaceRecognition UsingPrincipal Component Analysis, International Journal of Science Engineering and Technology2 (3):1985-3785.
[4] Garcia, EPI and PM Mora. 2011. Prediction Model of Academic Performance for First Year Students. IEEE Computer Society.
[5] Gunn, S. 1998.Support Vector Machines for Classification and Regression, Technical Report, ISIS4(3): 1985-3785.
[6] Habibi, R. 2011. Application of Data Mining Techniques with Smooth Method of Support VectorMachine (SSVM) to predict the chance of students who drop out (polytechnic student case studyfield). Thesis. University of Northern Sumatra.
[7] Han, Jiawei. &Kamber, Micheline. 2000 Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Urbana-Champaign
[8] Hien, NTN and P. Haddawy. 2007. A Decision Support System for Evaluating International Student Applications. IEEE Frontiers in Education Conference.