ATLANTIS PRESS

Atlantis Highlights in Engineering (AHE), volume 2

International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018)

Data Mining Approach to Classify Tumour Morphology using Naïve Bayes Algorithm

Shahmirul Hafizullah Imanuddin[1], Irfan Darmawan[2], Rahmat Fauzi[3]

[1,2,3]Information System Study Program, Industrial and System Engineering Faculty, Telkom University

[1]mirurubatsune@gmail.com, [2]irfandarmawan@telkomuniversity.ac.id, [3] rahmatfauzi@telkomuniversity.ac.id

Abstract— Tumours are a very sickly disease and recorded as the second killer disease in the world. This is because until now tumour still not found a drug that can really cure it. In Indonesia itself has many people who have suffered from tumours. Ignorance makes people reluctant to observe the early symptoms of tumour. In addition, the hospital also has problems with what type of tumour is most common in the community and their target of socialization to the community and hospital environment.

In this study, the topic of discussion focused on making patterns of tumour disease patients using Naive Bayes algorithm on Rapidminer tools using supporting variables of sex, age and place of tumour in the body. The output of this study is a posterior probability value of each variable with 66.76 percent accuracy. In addition there is also a density value in the form of a chart on each supporting variable.

Keywords— Naive Bayes Algorithm, Classification, RapidMiner, Tumour.

I. INTRODUCTION

Currently Data mining has been well known in the world of health. Technically Data mining provides potential information with considerable coverage. Classification techniques and predictive data is one example of data mining techniques applied in some cases in the field of medical records.

People think that tumours are the same as cancer, whereas the reality is different. Tumours are a general term that describes the growth of mass (solid / solid) or abnormal tissue in the body that includes benign tumours (benign tumours) and malignant tumours (malignant tumours). Malignant tumours are known as cancer. This mass arises as a result of growth imbalances and cell regeneration. Uncontrolled cells because of DNA damage that presents a mutation (a genetic decline) in a vital gene that carries out cell division. Some mutations may be needed to convert normal cells into cancer cells. These mutations are chemical or physical substances called carcinogens. Mutations can occur spontaneously (obtained) also inherited. Cancer movement with tumour cells with normal environment, immune cells, and also an effective system. Therapeutic agents issued are chemotherapy and immunotherapy (Badan Penelitian dan Pengembangan Kesehatan, 2013).

Cancer is one of the dangerous diseases for humans. Occupying second place as the deadliest disease in the world, making cancer familiar in the minds of society. In 2015 there are 8.8 million deaths from cancer in the world such as lung cancer 1.69 million deaths, liver cancer 788,000 deaths, colorectal cancer 774 000 deaths, stomach cancer 754,000 deaths and breast cancer 571,000 deaths (Plummer et al., 2016).

In Indonesia, tumour has also become a common disease in the community. In addition, the hospital also does not make real efforts to reduce the occurrence of tumours. Due to the difficult healing process, the effects it generates and the considerable cost savings for cancer treatments and treatments become the most feared diseases of the community. The appearance of a lump that is strange in shape and location needs to be suspected, because of the possibility that a tumour has occurred. Often seen many people who come to the hospital already became very dangerous tumour condition because it has turned into cancer. Health institution's ignorance about the most prevalent cancer disease makes it difficult for them to socialize to the public. Types of tumours that vary makes socialization to the community is always not right on target. As a result, socialization is done as far as possible using only general knowledge about the tumour itself. Classification of the tumour type is needed to determine the target of socialization so that the knowledge is useful for the community. In addition to making people more aware of the dangers of tumour, people can also get a picture of what tumour disease that may occur to them if not immediately implement a healthy life

From the things described above, an analysis will be carried out that raised the problem, it is necessary to research, by implementing a data mining classification method that focuses on Naïve Algorithm on patient XYZ hospital disease data. Researchers hope that the classification results can be used to determine the target of socialization to be right on target so as to reduce the number of tumour sufferers in Indonesia.

II. THEORITICAL BASIS

II.1 ICD

ICD (International Classification of Diseases) is the International Standard Diagnostic tool for epidemiology and health management. Up to six characters long, ICD designed to map health condition to corresponding a specific variation of disease and make it became generic categories based on signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease(“WHO | International Classification of

Diseases," 2018).

II.2 Data Mining

Data mining is the process of analyzing the data that the amount is very large so that obtained an information usually in the form of a pattern that will then be used as knowledge to solve problems that are happening. As well as mine which results are uncertain, Information we get from data mining has a different perception depending on what information we are looking for and how to summarize it into important information that can be used as knowledge(Gorunescu, 2011).

The steps in performing data mining that showed in Figure 1 are as follows:

1. Data cleaning
   Used to remove noise and inconsistent data
2. Data integration
   Combine the different data but with a same meaning
3. Data selection
   Retrieved relevant data from the database
4. Data transformation
   Transform data so the operation can perform well
5. Data mining
   The process to extract data patterns
6. Pattern evaluation
   identify the meaning of patterns representing knowledge
7. Knowledge presentation
   knowledge representation so it can be useful to solve the

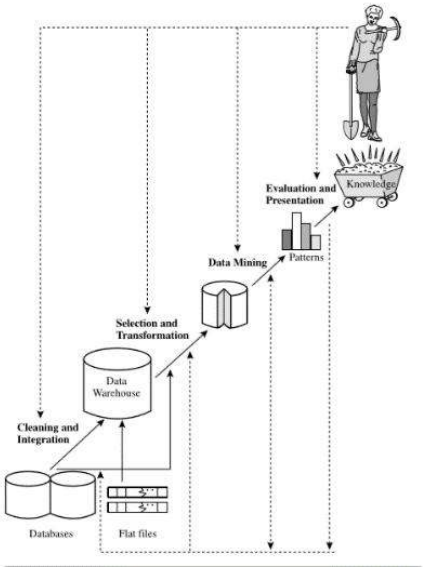

Figure 1 Data Mining Process (Data Mining Concept and Techniques)

II.3 Naïve Bayes Algorithm

The Naive Bayes algorithm is a classification method using the probability and statistical methods proposed by the British scientist Thomas Bayes (NFORMATIKALOGI, 2017). The Naive Bayes algorithm predicts future opportunities based on past experiences so known as Bayes Theorem. The main feature of this Na Nave Bayes Classifier is a very strong assumption of the independence of each condition / occurrence. The formula shows in (2.1)

$$P(C|X) = \frac{P(C|x)P(c)}{P(x)}, \quad (2.1)$$

The above formula illustrates that the probability of entering a sample of certain characteristics in class C (Posterior) is the probability of the emergence of class C (before the entry of the sample, often called prior), multiplied by the probability of occurrence of sample characteristics in class C (also called likelihood) probability of occurrence of sample characteristics globally (also called evidence). Therefore, the above formula can also be written as follows (2.2) formula:

$$posterior = \frac{prior\ x\ likelihood}{evidence}, (2.2)$$

For classification with continuous data used Gauss Density (2.7) formula below:

$$P = (X_i = x_i \mid Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i - \mu_{ij})^2}{2\pi\sigma_{ij}^2}}, (2.7)$$

For mean using (2.8) formula below:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, (2.8)$$

And for standard deviation using (2.9) formula below:

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5} \quad (2.9)$$

II.4 Rapidminer

RapidMiner is the Java-based open source tool to perform data mining technique. RapidMiner can work in collaboration with Rapid Analytics.

RapidMiner GUI-based tool that has many ways of data viewing, make RapidMiner the most advantageous and easier to use. Because RapidMiner has friendly graphical user interface make RapidMiner a commonly used data mining tool. RapidMiner graphic user interface allows users to drag and drop the data analytic blocks to the working space in order to create the data analysis workflow which is the combination of the data analytic operators. RapidMiner can work efficiently by using the provided operators in form of blocks (Kitcharoen, Kamolsantisuk, Angsomboon, & Achalakul, 2013).

III. RESEARCH METHODOLOGY

III.1 Conceptual Model

Figure 2 presents the problem that occur in this research which is there are hospital that want to predict cancer growth rates in patient by age and sex. Using classification method and Naive Bayes Algorithm, researcher solved this problem based on data patient that has indicated suffering tumour.

The purpose of Naive Bayes is as an prediction for the next patient tumour type growth for each age, sex, and the area where the tumour growth. This conceptual method is supported by knowledge of data.

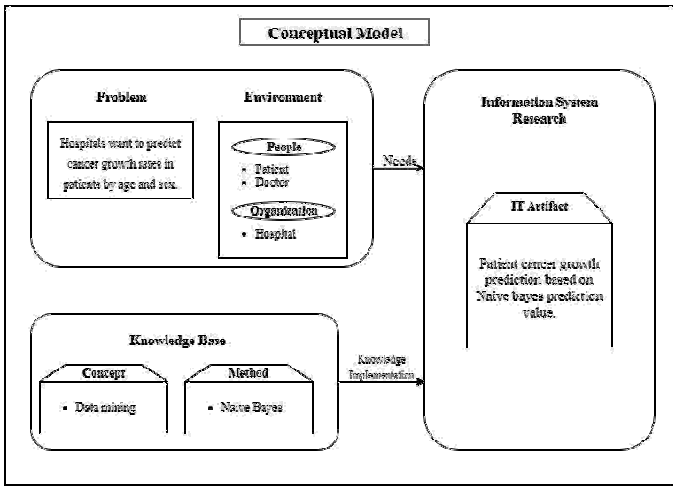

Figure 2 Conceptual Model

III.2 Systematics Research

The research in Figure 3 will be explain of each process as follows:

1. Collecting Stage
   Collecting the data is the first and the most important stage that should be done. Before continue to the next step, the data that already collected need to be understood to know what information can be had from that data.
2. Stage
   Selecting data means to choose the most influential variable from data so the result can be more accurate. Selecting data minimizes the scope so the research will become focused and right on target
3. Cleaning Stage
   Data cleaning is the stage next after selecting data. In this stage Rapid Miner will do the process that make the data pure without duplication and null data. Duplication and null data can make mining process goes error, so remove it is important to make the mining process perform smoothly.
4. Transform Stage
   Transform stage change the form of data to fit with the used algoritm. This stage must perform because there are some algoritm that not perform well when the value of data not match in the requirement.
5. Data Mining Stage
   Data mining stage is purposed to gain information on patient of tumour disease data. The stage starts from choosing the data mining task which is classification, the data from data cleansing stage will be classify based on the atribute of data patient known as Age, Gender, Topology, Morphology. The prediction will be calculated using Naive Bayes algorithm. The data that has been cleansed will be processed to get output the prediction value of tumour patient based on sex and age.
6. Evaluation Stage
   When the data mining process finished and get the prediction probability of the tumour patient, the data will be evaluated and processed by read the result of naive bayes algoritm.

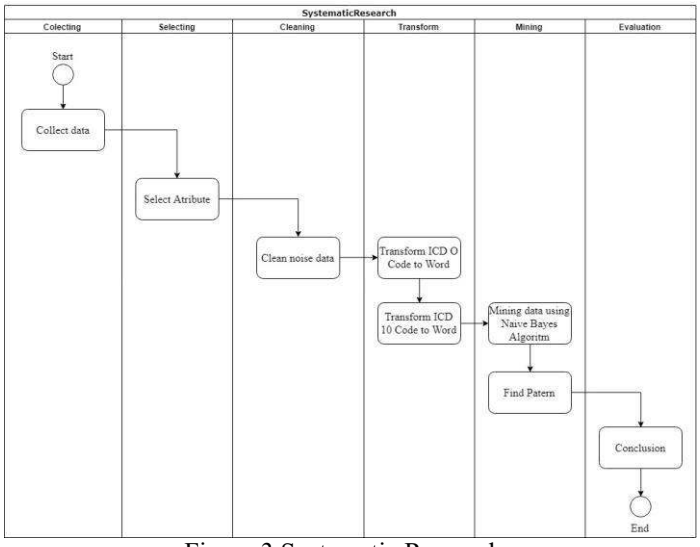

Figure 3 Systematic Research

III.3 Data Gathering

Data set is gathered from read csv operator before processed by Naive Bayes Algorithm. The data (Table 1) consists of 4 columns testing which one column in green is a label or variable target of the data in morphology column and the other is support or predictor variable know as gender, age and topology.

Table 1 Tumour Patient Dataset

Row No.	Morfologi Ko...	Topografi K...	Umur	Jenis Kelamin
1	Ganas	Pencernaan	26	Laki-Laki
2	Ganas	Kemaluan W...	57	Perempuan
3	Ganas	Saluran Kemih	75	Perempuan
4	Ganas	Pencernaan	47	Laki-Laki
5	Ganas	Pencernaan	19	Perempuan
6	Ganas	Kemaluan Pria	42	Laki-Laki
7	Ganas	Kemaluan W...	20	Perempuan
8	Ganas	Kemaluan W...	52	Perempuan
9	Ganas	Pencernaan	33	Laki-Laki
10	Ganas	Kemaluan W...	30	Perempuan
11	Ganas	Pernafasan	58	Laki-Laki
12	Ganas	Kulit	53	Perempuan
13	Ganas	Mulut	30	Perempuan
14	Ganas	Kulit	42	Laki-Laki
15	Ganas	Kemaluan W...	49	Perempuan
16	Ganas	Kemaluan W...	58	Perempuan

IV. RESULT

IV.1. Naive Bayes Distribution Table

Using Naive Bayes operator in RapidMiner to do the Naive Bayes Algorithm, the result is shown in the Table 2 below.

The table show that in Jakarta the highest patient tumour is female with tumour benign posterior value = 0.456, and tumour malignant posterior value = 0.214. Its conclude because there is no male that can have womb tumour and the number of females in data set is 411. Second place is lung tumour with tumour benign posterior value = 0.285, and tumour malignant posterior value = 0.175. For the last, the data show the place that not clear where the tumour is with metastasis attribute as the highest posterior value that is 0.459. in the Chapter 2 metastasis mention as a tumour that already spread to another organ. It happened because the tumour is already becoming cancer and the dangerous cancer can spread the disease to other organ in body. For the age the patient evenly when they reach 50 years old.

Table 2 Naive Bayes Posterior Probability Result

Attribute	Parameter	Ganas	Jinak	Metastasis
Topografi Kode	value=Pencernaan	0.115	0.031	0.059
Topografi Kode	value=Kemaluan Wanita	0.214	0.456	0.071
Topografi Kode	value=Saluran Kemih	0.022	0.013	0.000
Topografi Kode	value=Kemaluan Pria	0.011	0.006	0.000
Topografi Kode	value=Pernafasan	0.285	0.175	0.235
Topografi Kode	value=Kulit	0.009	0.019	0.024
Topografi Kode	value=Mulut	0.139	0.031	0.059
Topografi Kode	value=Payudara	0.104	0.044	0.012
Topografi Kode	value=Tidak Jelas	0.033	0.044	0.459
Topografi Kode	value=Mesothelioma	0.009	0.031	0.047
Topografi Kode	value=Endokrin	0.033	0.131	0.000
Topografi Kode	value=Tulang	0.022	0.019	0.035
Topografi Kode	value=Sistem Saraf Pusat	0.004	0.000	0.000
Topografi Kode	value=unknown	0.000	0.000	0.000
Umur	mean	50.898	48.494	51.071
Umur	standard deviation	14.876	14.139	15.853
Jenis Kelamin	value=Laki-Laki	0.470	0.213	0.471
Jenis Kelamin	value=Perempuan	0.530	0.787	0.529
Jenis Kelamin	value=unknown	0.000	0.000	0.000

V. CONCLUSION

V.1. Conclusion

Based on the results of analysis and processing of existing data it can be concluded as follows:

1. From the research the socialization target for male is in respiratory organ. The scatter chart shows the dot for male gather around respiratory organ. For female the target will be in womb and breast because in scatter chart the dot gathers more in womb and not a few are also located in breast area. For age the patient that already have tumour even for malignant, benign or metastasis type is around age of 57 years old same for male and female. So, the socialization target will be the male respiratory organ and for female is womb and breast with the age ten year before 57 years old.
2. The accuracy of tumour morphology using Naïve Bayes algorithm is 66.76%. The test result comes with 466 from 700 patient match with the paternt. This value show up because there are a limit in variable that make the algorithm calculation can
3. The prediction not work smoothly because there is a limit in variable that used in this reasearch. The small amount of variable make the prediction can not be trusted. There is impossible for people to predict the tumour growth rate just based on gender and age without any other variable because is not make

Rapidminer also can be said a good tool to do datamining process. The reason is because it simple, have a lot of tool that can be drag and drop on the process panel. Not just that rapid minner has a various type of algorithm as it default algorithm.

V.2. Suggestion

The naive bayes algorithm actuallly good for the prediction type of mining. So, the reasearcher sugest to Hospital to use other type of data maybe like history of another desease from patient and their familly, their behaviour in environment, what their eat and drink very often and many more. This can make the accuracy more higher that before. Not just that, the result can bo not just a clasification but also a prediction of the tumour itself.

REFERENCES

[1] Badan Penelitian dan Pengembangan Kesehatan. (2013). Riset Kesehatan Dasar (RISKESDAS) 2013. *Laporan Nasional 2013*, 1–384. https://doi.org/1 Desember 2013

[2] Gorunescu, F. (2011). *Data Mining: Concepts and Techniques*. *Elsevier* (Vol. 12). https://doi.org/10.1007/978-3-642-19721-5

[3] Kitcharoen, N., Kamolsantisuk, S., Angsomboon, R., & Achalakul, T. (2013). RapidMiner Framework for Manufacturing DataAnalysis on the Cloud. *2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 1–6. https://doi.org/10.1109/JCSSE.2013.6567336

[4] NFORMATIKALOGI. (2017). Algoritma Naive Bayes | INFORMATIKALOGI. Retrieved July 12, 2018, from https://informatikalogi.com/algoritma-naive-bayes/

[5] Plummer, M., de Martel, C., Vignat, J., Ferlay, J., Bray, F., & Franceschi, S. (2016). Global burden of cancers attributable to infections in 2012: a synthetic analysis. *The Lancet Global Health*, *4*(9), e609–e616. https://doi.org/10.1016/S2214-109X(16)30143-7

[6] WHO | International Classification of Diseases. (2018). *WHO*. Retrieved from http://www.who.int/classifications/icd/en/