DATA MINING APPROACH TO CLASSIFY TUMOR MORPHOLOGY USING DECISION TREE ALGORITHM

Fasya Dzul Fikri Akbar[1], Irfan Darmawan[2], Rahmat Fauzi[3]

[1,2,3]Prodi S1 Sistem Informasi, Fakultas Rekayasa Industri, Universitas Telkom

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

***Abstract*—** Tumors are a general term used to describe the growth of abnormal masses or tissues in the body that include benign tumors, malignant tumors and unidentified tumors. Malignant tumors are known as cancer. The operation of cancer data using a fairly popular tool that is rapidminer. Topics of discussion about classification of patients with tumor disease using Decision Tree algorithm on Rapidminer tools that use supporting variables such as age, sex and place of tumor on the body / topography.

The output of the research is a decision tree with a precision of 85.53% that can be used and implemented by the hospital to facilitate socialize the importance of tumor disease to the community in the hope that the community can prevent as early as possible about the danger of tumor disease. Because most people happen to come to the hospital when it has been affected by a malignant tumor (regardless of cost factor), based on xyz hospital data.

***Keywords*—** Data Mining, Decision Tree, Classification, Tumor Disease.

I. INTRODUCTION

Tumors are an alarming disease as the number one cause of death in Indonesia with 5.7 percent of the total Indonesian population who died (Basic Health Research 2007). Research also states that every 1000 people there are about 4 tumor sufferers. This factor continues to increase in subsequent years so that within 10 years (2005-2015) WHO estimates the number of deaths from tumors averaging 8.4 million annually and by 2015 to 9 million.

Tumors are a general term used to describe the growth of mass (solid / solid) or abnormal tissue in the body that includes benign tumors and malignant tumors. Malignant tumors are known as cancer. This mass arises because of growth imbalances and cell regeneration. Uncontrolled cell growth is caused by DNA damage that causes mutations (decreased genetic changes) in vital genes responsible for controlling cell division. Some mutations may be needed to convert normal cells into cancer cells. These mutations are caused by agents of chemical or physical substances called carcinogens. Mutations can occur spontaneously (obtained) or inherited.

The development of cancer is characterized by tumor cells interacting with surrounding environmental components such as normal cells, immune cells (effector cells), as well as therapeutic agents that can be externally added to the body system. Treatment agents are chemotherapy and immunotherapy. The nature of tumor environmental interactions is complex and depends on many factors, including age, sex and so on. These factors can cause tumors cell changes to be complex.

Public opinion against cancer is a disease most feared by the society because of the difficult healing process, the effects caused and require considerable cost for treatment and also treatment. Many people think that cancer is the same as a tumor, when in fact the tumor that appears does not mean cancer. The appearance of a strange lump in terms of shape and location of the growth needs to be suspected, because it can be ascertained is a tumor. Therefore, it is necessary to socialize the danger of tumor disease so that residents can know the characteristics of tumor diseases and can prevent the occurrence of tumors. Because we often see many people who come to the hospital but with the condition of tumor disease that is very dangerous, because of less information and financial factors.

II. THEORITICAL BASIS

II.1 Tumor Disease

Tumors are a lump caused by cell growth. All tumors of both benign and malignant tumors have two basic components: parenchyma and stroma. Parenchyma is a proliferative tumor cell, which exhibits growth and functional properties varying with the function of the original cell. For example, the production of collagen, mucin, or keratin. Stroma is a supporter of tumor parenchyma, consisting of connective tissue and blood vessels. Presentation of food in tumor cells through the blood vessels by diffusion.

Based on its biological nature, tumors can be distinguished from benign tumors and malignant tumors and tumors located between benign and malignant called "intermediate.that is:

a. Benign Tumor**:** Benign tumors grow slowly and usually have a capsule. It does not grow infiltratively, it does not damage the surrounding tissue and does not cause scattered children in distant places. Benign tumors are generally cured

© Copyright by Directorate of Research and Community Service, Telkom University

perfectly unless they secrete hormones or those that lie in a very important place, such as a spinal assumption that can cause paraplesia or a brain nerve that suppresses brain tissue.

b. Malignant Tumor**:** Malignant tumors generally grow rapidly, infiltratively. And destructive surrounding network. Besides it can spread throughout the body through the flow of limped or bloodstream and often cause death.

c. Intermediate Tumor

Among the 2 groups of benign and malignant tumors are small tumors that have local invasive properties, but their metastatic ability is small. Such tumors are called aggressive tumors of low-grade malignant local tumors. An example is skin basal cell carcinoma.

II.2 International Statistical Classification of Diseases (ICD)

According to Hatta (2013) ICD (International Classification of Diseases) is a comprehensive and internationally recognized classification system. ICD is a standard diagnostic tool for the epidemiology of health management and clinical goals. Includes an analysis of the health conditions of a population group in general. ICDs are used to monitor incidents or incidents and diseases in general as well as other health problems, by providing an overview of the general health conditions of the population in the countries of the world.

The function of ICD used to :

- Indexing disease records and actions in health care facilities.
- Input for medical diagnosis reporting system.
- Facilitate the process of storing and retrieving data related to the diagnosis of patient characteristics and service providers.
- Analysis of health service financing.

II.3 Data Mining

Data mining is defined as the process of finding a pattern from a set of data (Witten, 2011). The found pattern must have knowledge that will be used to create profit, especially in the field of economy and business. Understanding other data mining is extracting knowledge from a set of data that the number is very large(Han, 2006).

II.4 Decision tree

The decision tree is a flow chart that is shaped like a tree structure in which each internal node declares testing of an attribute, each branch representing the output of the test and the leaf node representing classes or class distributions. The topmost node is called the root node or root node.

a. Entropy is the value of information that states the size of the uncertainty (impurity) of the attributes of a collection of data objects in units of bits.

$$\text{Entropy}(S) = \sum_{i=1}^{n} - pi * Log_2\ pi$$

Type:

S: case set

n: partition number S

Pi = proportion of Si to S

b. Information Gain

Information Gain is a measure of the effectiveness of an attribute in classifying data. Used to specify the order of attributes where the attribute has the largest value of Information Gain selected.

$$\text{Gain}\ (S,A)\ =\ \text{Entropy}\ (S)\ -\ \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

Type:
S: The set of cases
A: Attribute
n: number of partition attributes a
| Si | : the number of cases on the i-partition
| S | : the number of cases in S

III. RESEARCH METHODOLOGY

III.1 Conceptual Model

The Conceptual Model is defining the model as the main body of information about the system collected to study the system.

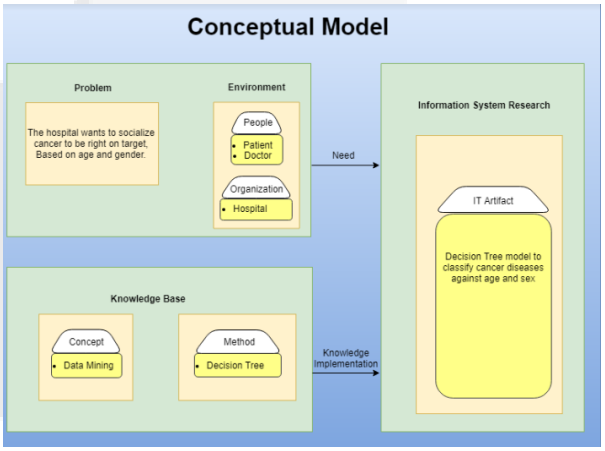

Fig 1 Conceptual Model

III.2 Systematics Research

Systematic research is a research method to synthesize the results of research. In using this

© Copyright by Directorate of Research and Community Service, Telkom University

systematic research, we can map out how the process from beginning to end, so we can do it systematically.

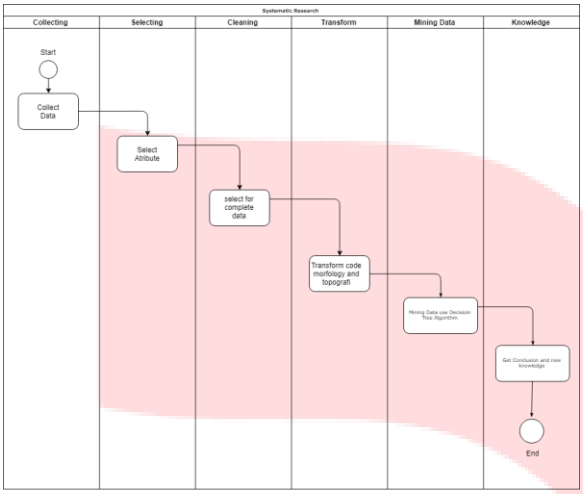

Fig 2 Systematic Research

1) Collecting Data

• This data is obtained in one hospital in Indonesia, this data has not been done data mining before. This data collection process is collective. this raw data will then be in the selection of attributes and this process will be described in the sub next chapter.

2) Selecting Data

• The raw data will then be done in the process of selecting data, namely: the process of selecting attribute data to determine which attributes that have correlation to our final project, and also the selection of this attribute is used to determine which label / reference attribute.

3) Cleaning data

• The raw data will then be done in the process of selecting data, namely: the process of selecting attribute data to determine which attributes that have correlation to our final project, and the selection of this attribute is used to determine which label / reference attribute.

4) Transform data

• Transform data is how we change the data to make it easier in the process of data mining classification, and in graphical results later easier to understand. The data to be transformed is morphology and typography data codes.

5) Mining Data

• As explained above. This process is the core process towards the outcome we will be able to achieve. In this data mining process, we use the data mining decision tree algorithm and the tool in use is rapidminer.

6) Knowledge

• This is the final process in the process to get the result. The process of this knowledge after we know the results of the data mining above then if we become a new knowledge in order to support in a decision.

IV. RESULT

IV.1. Data Set

Data set result is data gathered from read csv parameter. After we do the selection of attribute and prose clinsing then the empty data will be lost and attributes that are not in use will be deleted as well. The data already in the process of select attribute and cleansing then the data will be in data mining. Figure 14 is data set after processed

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

Fig 3 Data set

IV.2. Decision Tree Model

The root node of this decision tree is topografi kode , the branch nodes are age, gender. And leave class is morphology. Figure 15 shows the decision tree result.

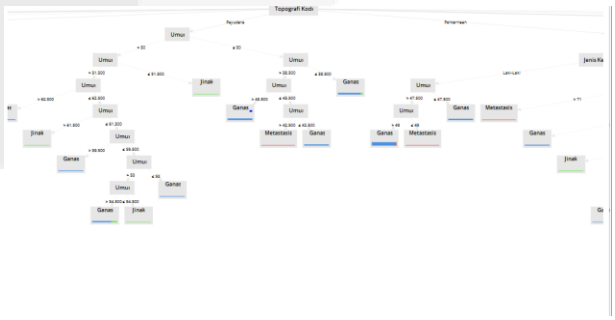

Fig 4 Decision Tree model

© Copyright by Directorate of Research and Community Service, Telkom University

From the decision tree model that has been run, we get the result of 85.53% accuracy. from the above data we can get info that:

- both male and female can be affected by tumor disease

- Under 30 years of age usually only affected by benign tumor disease

- Most people affected by malignant tumor disease at the age of about 50 years

- And others.

V. CONCLUSION

V.1. Conclusion

Regarding the decision tree model that has an accuracy of 85.53%, from data that has been cleaned that there are 698 patient data. Can be concluded that: all humans can be affected by cancer both young and old and even old age. But from this data we conclude age under 30 is usually indicated benign tumor disease. And for benign tumors usually appear most at the age of about 50 years. And I state that the Decision tree algorithm is well suited for the case of "Data Mining for Tumor Morphology Classification with Decision Tree Algorithm.

V.2. Suggestion

This decision tree model is possible to be more useful in classifying tumor disease against age and gender, so that data more accurately we need more attributes and have a high correlation to support the accuracy of data mining such tumor disease.

REFERENCES

[1] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine*

[2] Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques 2nd edition.* San Francisco: Morgan Kaufmann Publishers

[3] Birta, L. G., Edward, K., Box, P. O., & Stn, A. (2015). Conceptual Modelling:Definition, Purpose and Benefits, 2812–2826.

[4] Goyal, Anshul dan Mehta, Rajni. 2012. *Performance Comparison of Naïve Bayes and J48 Classification Algorithms*. Kurukshetra: Kurukshetra University.

© Copyright by Directorate of Research and Community Service, Telkom University