Classification of Flood Disaster Predictions using the C5.0 and SVM Algorithms based on Flood Disaster Prone Areas

Saruni Dwiasnati[1], Yudo Devianto[2]

University of Mercubuana, Faculty of Computer Science
South Jakarta, Indonesia

Abstract
Many researchers have been motivated to improve the performance of predictive methods. So that is what prompted researchers to conduct this research in order to find out the object of the Flood Disaster, whether it can be done using the Classification method. The main factor in the occurrence of Flood Disaster is the increasing intensity of rainfall that clogs the river water flow, which further pressures the river water to dike embankments that are no longer strong by carrying materials found in the flow of water from upstream to downstream, such as Wood, Mud, There are even rubbish from home-based industries which are carried away by flood flows which cause many rivers to become clogged. Floods have the meaning of one of the natural disasters that occur due to increased rainfall from normal which can cause casualties and often occur in lowland areas. In this study, a classification will be conducted on how to predict Flood Disasters based on their Prone Areas and this research takes the Bandung area as the object of research. The algorithm used in this study is to compare 2 Classification algorithms namely C5.0 and SVM Algorithms to determine the accuracy value of which algorithm is much higher than other algorithms based on the Prone Areas. C5.0 and SVM algorithms can be used on datasets that have been modeled to produce value accuracy. Data processing in this study uses the Orange application which can be used to create a model of data that has been processed into information that can be given to the public for the early warning of the flood disaster that they will face in the future obtained from past data. Orange is one of the open source software used for processing Data Analytics / Data Mining.

Keywords — *Classification, Bencana Banjir, Algoritma C5.0, SVM, Orange*.

I. INTRODUCTION

According to Krishna S. Pribadi (2008) a disaster is an event or series of events that threatens and disrupts people's lives caused by natural factors or non-natural factors, causing human or animal fatalities, environmental damage, property losses and psychological impacts [1] To avoid or reduce the impact caused by natural disasters, such as landslides, floods or earthquake disasters, disaster management needs to be made. Where disaster management consists of mediation or mitigation, preparedness, emergency response, rehabilitation and reconstruction at the stage after the disaster. Mitigation itself has the meaning of an action taken to reduce the impact / influence caused by a disaster. Mitigation actions consist of structural mitigation and non-structural mitigation. Structural mitigation is an action to reduce or avoid the possibility of physical disasters. Non-structural mitigation is a policy-related action, awareness building, development of existing knowledge and regulations.

According to the Gartner Group, data mining is the process of discovering new relationships that have meanings, patterns and habits by sorting out most of the data stored in storage media by using pattern recognition technologies such as statistical and mathematical techniques.

Data mining is a combination of several scientific disciplines that unite techniques from machine learning, pattern recognition, statistics, databases, and visualization to handle the problem of retrieving information from large databases [2].

Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases. Data mining is a series of processes to explore the added value of a data collection in the form of knowledge that has not been known manually [3]. From the definitions that have been delivered, important things related to data mining according to [3]:

1. Data mining is an automatic process of existing data.
2. Data to be processed is very large data.
3. The purpose of data mining is to get relationships or patterns that might provide useful indications for objects using One of the things that can be done in data mining is classification.

Classification was first applied to plants that classify a particular species, as was done by Carolus von Linne (also known as Carolus Linnaeus) who first classified

species based on physical characteristics. Furthermore he is known as the father of classification [4]. In the classification there are target categorical variables. The methods / models that have been developed by researchers to resolve classification cases include [4]:

a. Decision tree
b. Classifier of bayes / naive bayes
c. Artificial neural network
d. Statistic analysis
e. Genetic algorithm
f. Rough sets
g. K-nearest neighbor classification
h. Rule-based method
i. Memory based reasoning
j. Support vector machine

Predictions have similarities with classification and estimation, the difference between classification and estimation with predictions is in the predicted value generated will be in the future. Example predictions in business and research are:

a. Prediction of rice prices in the next three months.
b. Prediction of the five-year unemployment rate will come.
c. Predict the percentage of traffic accidents next year.
d. Forex value predictions for the next two months.

Some methods and techniques used in classification and estimation can also be used for prediction.

C5.0 algorithm is one of the algorithms that are included in the classification section of data mining which is specifically applied to the decision tree technique. C5.0 is an improvement from the previous algorithm formed by Ross Quinlan in 1987, namely ID3, CART and C4.5. C5.0 algorithm produces trees with the number of branches per node varies. C5.0 treats continuous variables similar to those done by CART, but for C5.0 categorical variables treats the value of categorical variables as splitters.

Support Vector Machine (SVM) is a supervised algorithm in the form of classification by dividing data into two classes using vector lines called hyperplane (Octaviani, et al., 2014). Basically the SVM algorithm is an algorithm that uses a hyperplane to be used as a separator between data in a linear manner, so as to overcome data problems

in the form of nonlinear data, the kernel trick technique can be used. The Support Vector Machine (SVM) method is also quite good at solving multiclass classification problems.

According to Erlangga (2007: 10) Flood-Prone Areas have several meanings including:

(1) areas that have high rainfall
(2) rock areas that have low water absorption
(3) the area around the river and becomes a river water flow
(4) dense and slum settlement areas
(5) areas that have experienced floods.

II. DATA MINING

Many uses can be used in processing data mining, which can help get useful information and increase knowledge from various data that can be solved by various algorithms in data mining. The definition of datamining itself is Data mining is a step in carrying out Knowledge Discovery in Databases (KDD). Knowledge discovery as a process consists of data cleaning (data cleaning), data integration (data integration), data selection (data selection), data transformation (data transformation), data mining, pattern evaluation (pattern evaluation) and presentation of knowledge (knowledge presentation) [9].

A. Classification Method

Classification method is a method that is included in the most common supervised process that can be used in data mining. Business issues such as Churn Analysis, and Risk Management usually involve a Classification method to facilitate the resolution of the problem. Classification is the act of giving groups to every situation. Each state contains a group of attributes / indicators, one of which is a class attribute. This method requires a model that can explain the class attribute as a function of the input attribute. The advantage of the classification method is that the dataset used in absolute classification must display the class / target attribute and the knowledge generated by the classification method in the form of clusters (can be Decision Tree, Ruleset, Weight on BackPropagation, etc.)

B. Algorithm C5.0

C5.0 algorithm is one of the algorithms included in the data mining section which is specifically applied to the decision tree technique. C5.0 is an improvement from the previous algorithm formed by Ross Quinlan in 1987, namely ID3 and C4.5. In this algorithm the selection of attributes will be processed using information gain. In selecting attributes for object breakers in several classes, attributes must be chosen that produce the greatest information gain. Attributes with the highest information gain value will be selected as parent for the next node [7].

C. Algorithm SVM

Support Vector Machine (SVM) was first introduced by Vapnik in 1992 as a harmonious series of superior concepts in the field of pattern recognition. As one of the pattern recognition methods, SVM age is relatively young. However, evaluating its capabilities in various applications places it as a state of the art in pattern recognition, and today is one of the fastest-growing themes. SVM is a learning machine method that works on the

principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes on input space.

D. Information Gain

The information gain size is used to select the test attribute on each node in the tree. This size is used to select attributes or nodes in the tree. Attributes with the highest information gain value will be selected as parent for the next node. The formula used for information gain is [8]

$$I\left(S_1, S_2, \ldots, S_m\right) = -\sum_{i=1}^{m} pi * log_2\left(pi\right)$$

E. Orange

Orange is a free datamining tool. The tool built with Python is quite useful for processing data to look for certain patterns in accordance with the concept of datamining. Compared to other Data Mining software, Orange is superior in terms of visualization or what we commonly call visual programming. Orange provides many widgets that we put on the canvas / drawing board and then we connect with other widgets. With this canvas media, it will make it easier for users to play with data and process data analytics intuitively. In addition to widgets, Orange also provides several add-ons / modules for problems in certain domain domains such as Text Mining / Text Analytics, Bioinformatics, Network Data / Social Networks, Model Maps, Prototypes Process, and others.

Symbol of tools orange which is used to process data that has been created on excel, as shown below

Gambar 1. Orange

III. DATA COLLECTION METHODOLOGY

A. Data Set

Data used in the classification of Potential Customer Candidates consists of 100,000 datasets, 332 data used for testing data based on available variables. In determining predictions of Flood Disaster based on Flood Disaster-Prone Areas 4 Variables are the occurrence of floods, the time of occurrence of floods, the longitude and latitude of the occurrence of floods.

IV. Analysis results

The AUC results for the C5.0 algorithm using data 332 from the destination object that is processed using the Orange application, namely 0.919.

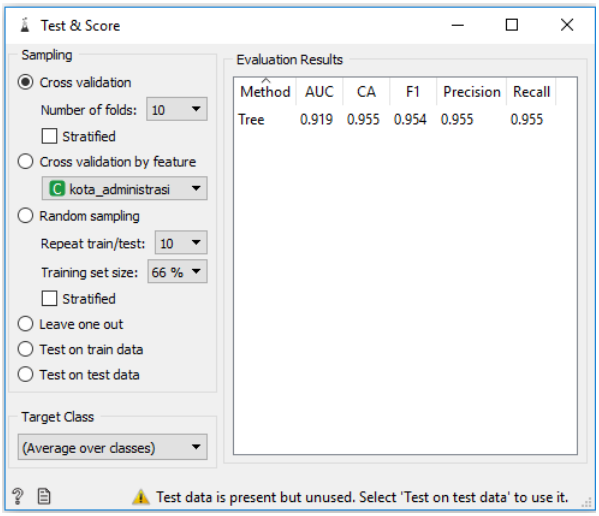

Figure 1. Results of the AUC algorithm C5.0

Figure 2 is a calculation using training data with the C5.0 algorithm which is processed using the Orange application. It is known that training data consists of 332 data records, 70 data classified FLOOD and 262 data predicted NOT FLOOD.

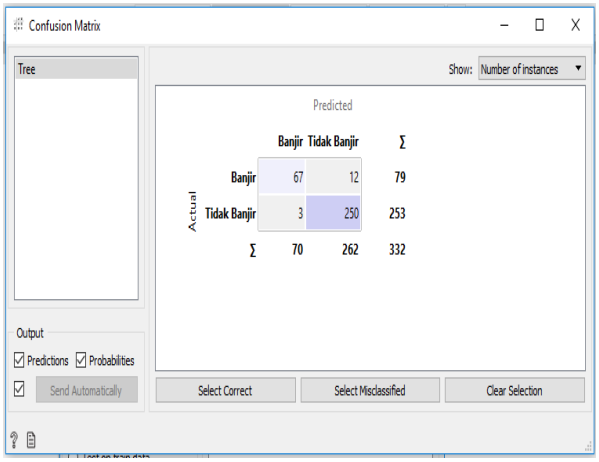

Figure 2. Confusion matrix c5.0 algorithm model

Whereas the results obtained from processing the ROC curve can be seen in Figure 3 for the FLOOD category resulting in the graph below.

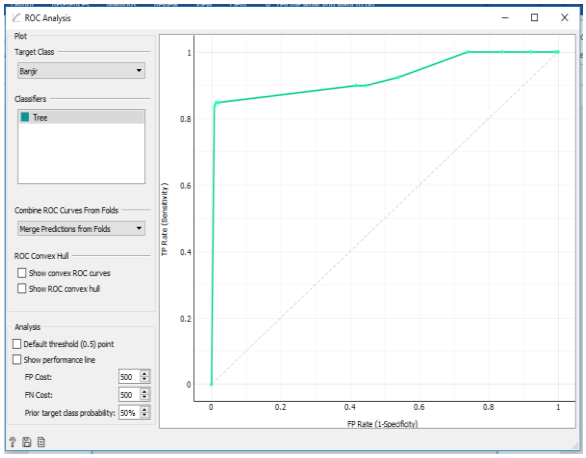

Figure 3 ROC Curve FLOOD category

While the results obtained from processing the ROC curve can be seen in Figure 4 for the NO FLOOD category resulting in the graph below.

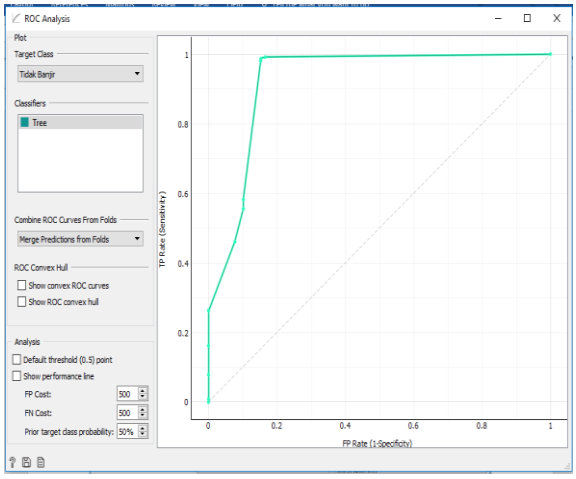

Figure 4 ROC Curve NOT FLOOD category

The AUC results for the SVM algorithm using data 332 from the destination object that is processed using the Orange application, namely 0.921.

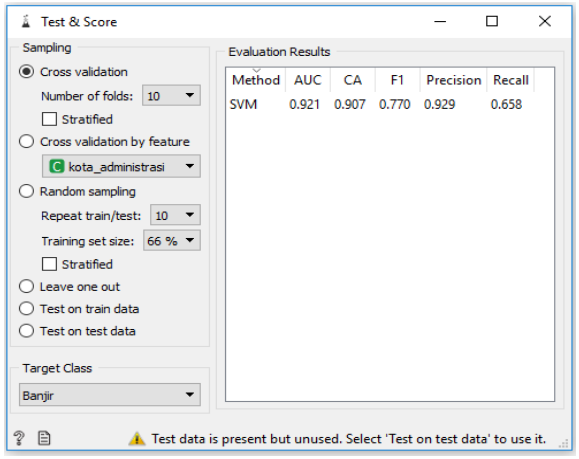

Figure 5 Results of the AUC SVM algorithm

Figure 6 is a calculation using training data with the SVM algorithm which is processed using the Orange application. It is known that training data consists of 332 data records, 56 data classified FLOOD and 276 data predicted NOT FLOOD.

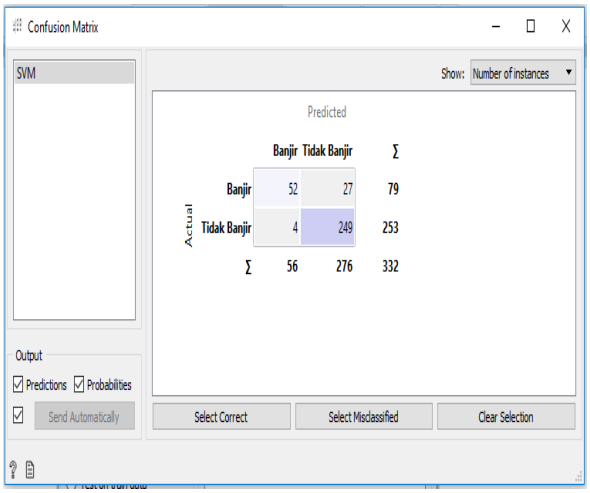

Figure 6. The Confusion Matrix SVM algorithm model

Whereas the results obtained from processing the ROC curve can be seen in Figure 7 for the FLOOD category resulting in the graph below.

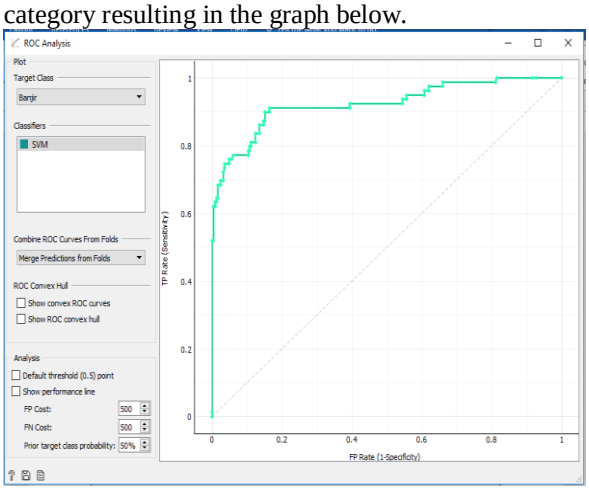

Figure 7 ROC Curve FLOOD category

While the results obtained from processing the ROC curve can be seen in Figure 8 for the FLOOD category resulting in the graph below.

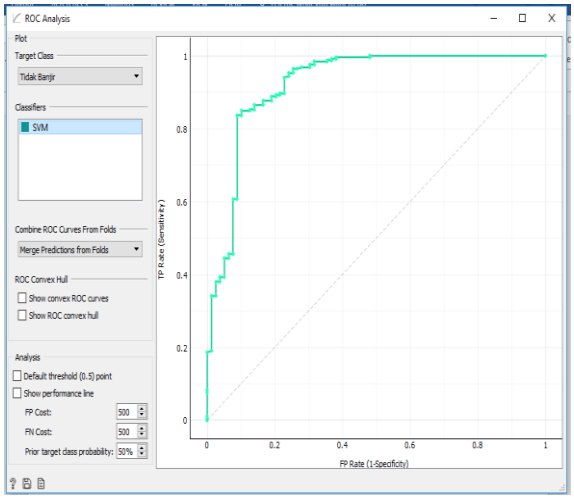

Figure 8 ROC Curve NOT FLOOD category

V. CONCLUSIONS

Based on the discussion of the results of the research discussed in the previous chapter, then in the Flood Disaster Prediction Classification study using the C5.0 and SVM Algorithms based on Flood Disaster Prone Areas, the conclusions from this study can be drawn:

1. Based on the research that has been done using 2 algorithms, it can be found that the results of the SVM algorithm have a higher AUC value of 0.921 compared with C5.0 algorithm having an AUC value of 0.919.
2. This research uses one of the tools that can be used to process a data to be able to produce a value from data mining, namely Orange.

ACKNOWLEDGMENT

Thank you to Mercu Buana University and the Faculty of Computer Science who have played a role in conducting research activities and specifically to PUSLIT Mercu Buana University, which has financially supported the completion of research that can be a journal that can be used as useful information for the field.

REFERENCES

[1] Krisna S. Pribadi (2008)
[2] Larose, Daniel T (2005) Discovering Knowledge in Data Mining An Introduction to Data Mining, Wiley Interscience
[3] Bramer, Max (2007)Principles of Data Mining, Springer Science
[4] Mardi, Yuli (2014) Analisa Data Rekam Medis untuk Menentukan Penyakit Terbanyak Berdasarkan International Classification Of Disease (ICD) Menggunakan Decision Tree C4.5 (Studi Kasus : RSU. CBMC Padang). UPI YPTK Padang
[5] Octaviani, P. A., Wilandari, Y. & Ispriyanti, D., 2014.Penerapan Metode Klasifikasi Support Vector Machine (SVM) Pada Data Akreditasi Sekolah Dasar (SD) Di Kabupaten Magelang. Jurnal Gaussian, 3(4), pp. 811-820.
[6] Han, Jiawei and Khamber, Micheline.2001. Data Mining : Concepts and Techniques. Morgan Kaufmann Publishers. San Francisco.
[7] Ernawati, Lin.2008. Prediksi Status Keaktifan Studi Mahasiswa Dengan Algoritma C5.0 dan K-Nearest Neighbor. Sekolah Pasca Sarjana Institute Pertanian Bogor: Bogor
[8] Katardzic M. 2003. Data Mining Concepts Models, Methods and Algorithms. New Jersey, USA: A John Wiley & Sons.
[9] Firdaus, Diky. 2017. "Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer". Jurnal FORMAT, Vol 6 Nomor 2, ISSN : 2089-5615.