Future Computer Science Journal (FCSJ)
Vol. 1, No. 2, December 2023: 74-80
http://asasijournal.id/index.php/fcsj
http://doi.org/10.xxxxx/fcsj.xxxx.xxx

Flood Disaster Detection Based on Rainfall Using Random Forest Algorithm

Mellisa[1*], Nurul Hidayat[2]
[1]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia
[2]Informatics Department, Faculty of Engineering, Universitas Jenderal Soedirman, Indonesia

Abstract
This research was conducted to detect natural disasters that occur around the world, namely floods based on rainfall, using the Random Forest algorithm. This research can also help the government estimate what to do in the future when a flood disaster occurs. The trick is to collect rainfall data and flood disaster data from Kaggle. The data is then cleaned, processed, and tested for reliability. The use of appropriate datasets for testing and the selection of the right algorithm can ensure the data mining process produces accurate information. Furthermore, the Random Forest algorithm was applied to the data to classify flood disasters based on rainfall. The results showed that the Random Forest algorithm can provide flood disaster classification results with a high accuracy rate of 95.8%. This study also aims to fill the gap of previous research by using the Random Forest algorithm in early detection of flood disasters based on rainfall. The novelty of this research lies in the use of algorithms that are rarely used in flood disaster classification research based on rainfall. With previous research using the Naïve Bayes, CART, and ANN algorithms have a lower level of accuracy than this research using the Random Forest algorithm. Therefore, this research is expected to contribute in developing a more accurate and reliable rainfall-based flood early warning system.

Keywords:
Flood;
Rainfall;
Random Forest Algorithm;
Dataset;
Machine Learning;

Article History:
Received: December 18, 2023
Revised: December 31, 2023
Accepted: December 31, 2023
Published: December 31, 2023

Corresponding Author:
Mellisa
Computer Science Department,
Universitas Negeri Semarang,
Indonesia
Email:
mellisalisa@students.unnes.ac.id

INTRODUCTION

Natural disasters are natural events that occur due to natural processes on the earth's surface. They cause enormous damage, both to property and human life [1]. Natural disasters in the world are numerous [2] and have a huge impact as well. Flood is a natural disaster [3] that occurs in a short time with a high transmission speed [4]. A very dangerous disaster in the world is flooding. The occurrence of large floods is of course due to rainfall and several other triggering factors. Floods also pose a serious threat to society and the environment. Therefore, it requires the right and fast handling to overcome it. By making flood detection based on rainfall. So it can provide predictions when flooding will occur and the government can be helped in preventing flooding. In this flood detection, a machine learning technique is needed. This technique has been widely used in prediction models [5].

This effort will deepen the understanding of urban flooding under nonstationary conditions and provide reliable information to water resource managers [6]. Flood forecasting and warning systems are very important for rural areas and urban centers, but especially in large urban centers [7]. To solve this problem, there have been many studies [8], [9] for prediction models using several data mining techniques, namely Association rule, C4.5, Classification and Random Forest [10]. Although some success has been achieved, most previous studies on flood prediction failed to reflect the buildings and detailed topography of urban basins [11]. This research is an experimental study that tests the Random Forest [12] algorithm for early detection of flood disasters based on rainfall. The results showed that the Random Forest algorithm can provide flood disaster classification results with a high level of accuracy, which is 95.8%. This shows that the Random Forest algorithm is an effective algorithm for use in early detection of flood disasters based on rainfall [13], [14]. This research fills the gap of previous research that only tests the Random Forest algorithm on rainfall data. This research tests the Random Forest algorithm on rainfall data from various sources, so that the data used is morecomplete and accurate. The

significant increase in accuracy of this study compared to previous studies shows that the use of more complete and accurate rainfall data can improve the accuracy of early detection of flood disasters.

The Random Forest algorithm itself is also a popular algorithm due to its high accuracy, resistance to noise and outliers, and ability to handle large data sets with high dimensions [15]. The Random Forest algorithm offers extensive capabilities in making predictions for various purposes, including in handling natural disasters such as floods. Thus, further research using the Random Forest algorithm and the same dataset as previous research can make a valuable contribution to better understand and optimize the potential of this algorithm [16].

This research aims to test the effectiveness of the Random Forest algorithm for early detection of flood disasters based on rainfall. The Random Forest algorithm was chosen because it has advantages in handling data and is able to overcome overfitting. This is in line with previous research that has proven that the Random Forest algorithm is effective in classifying and predicting various problems, including flood detection based on rainfall. For example, research conducted in 2018 successfully used the Random Forest algorithm to predict floods on the Nil river based on rainfall data and hydrological data. The results showed that the Random Forest algorithm was able to provide flood predictions with high accuracy.

METHOD

A machine learning algorithm that has accurately and efficiently predicted flooding and water resource phenomena has been developed [5]. As an algorithm, Random Forest does not stand alone. Random Forest is a supervised machine-learning algorithm that can be effective [17]. It also has statistical learning for prediction [13]. During prediction, each tree in the forest independently classifies or predicts the target variable and the final result is the mode or average of each tree's output [18]. Therefore, this research started by looking for journals to compare. The learning technique of Random Forest uses the idea of bagging and random feature selection to create a diverse ensemble of decision trees [19], given that this research seeks to get higher accuracy than previous research using the Random Forest algorithm method. Below is a picture of the stages of the research method as in Figure 1.

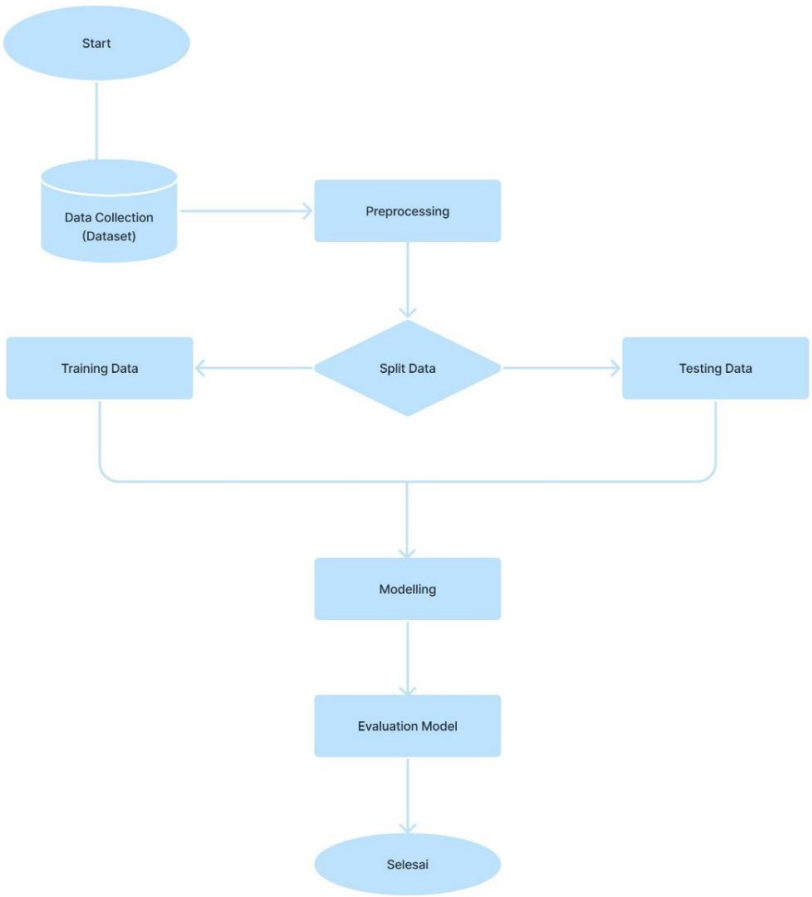

Figure 1. Flowchart of Proposed Method

Dataset

A dataset is a collection of data consisting of some information in a format that is easily accessed or processed by a computer. Datasets can contain various types of data including numbers, text, images, audio, and other types of data. The dataset used is Kerala from the Kaggle website, the following dataset link is shown (https://www.kaggle.com/code/mukulthakur177/flood-prediction-model/input?select=kerala.csv). This dataset has been widely used to test flood detection algorithms. This dataset has a csv format, in the dataset there are several columns, namely year, month, flood information or not, and others.

Preprocessing Data

Data preprocessing is one of the stages in performing the data mining process. Data preprocessing itself has stages that are commonly used in data mining, including exploratory data analysis (EDA), missing value removal, feature selection, and data normalization [20], [21], [22]. At this stage the data will be processed, namely data preprocessing. First the data will be requested to display the top five data in the dataset and four months that have high rainfall peaks. After that the data is entered into the data processing stage. This stage is done to ensure that the data is not empty (null) and it turns out that there is no empty data in it. Next, it will be given a labeling process. In this labeling, numerical features are used to ensure that all features have a similar scale.

Split Data

The split data process is next after the data has been cleaned and processed. Split data is dividing the dataset into two parts, namely training data and testing data. Training data is used to train machine learning models while testing data is used to evaluate model performance. In this study, split data will be compared to 20% for testing data and 80% for training data.

Modelling

In this stage the data will go through the process of building a model using the Random Forest algorithm. So in this stage I use the Random Forest algorithm. I chose this algorithm myself, because there are several studies using Random Forest that show the accuracy obtained is higher and effective for solving the problem.

The Random Forest algorithm is one method that can be used to solve classification problems. To solve the problem, there is a formula Equation (1) used by Random Forest [23]

$$Info\ (D) = -\sum_{i=1}^{m} pi \log^{2(pi)} \dots\dots\dots\dots\dots\dots\dots \quad \dots\dots\dots\dots(1)$$

The above formula calculates the information value using all the data. Where the probability of a tuple in *D* is the default class, also called the entropy of *D*, is the average information required to identify a tuple from *D*. If the value of *A* is discrete, the data *D* will be separated by several values of *A* so that each branch has a clean and similar matter. After the first branch, the number of possible units is measured by Equation (2).

$$InfoA\ (D) = -\sum_{i=1}^{m} \frac{|Dj|}{|D|} x\ InfoA(Dj) \dots\dots\dots\dots\dots\dots\dots\dots (2)$$

After that, calculate the information value with a formula. For each existing attribute, pay attention to the content of the data in detail.

Where:

$\frac{|Dj|}{|D|}$: the contents of partition *j*

$A^{(D)}$: information to classify tuples from *D* in partition *A*

The smaller the result of this equation, the better the partition. The value of an attribute determines whether or not it is essential in building a decision tree. If the attribute's value is continuous, it will be found by split_point by sorting all data from the smallest to the most significant attribute, and then taking the average. The gain value for each attribute will be calculated using the formula above, and the highest gain value will be made a branch in the decision tree, as in Equation (3).

$$Gain(A) = Info(D) - InfoA(D) \dots\dots\dots\dots\dots\dots\dots \quad \dots\dots\dots\dots\dots(3)$$

After all the decision branches have been formed, the calculation is repeated from the first to the last stage. If the branches have reached the maximum allowed branches, leaves will be included with most data values [23].

Evaluation Model

After performing the modeling process with the random forest algorithm, performance testing is carried out to test how well the model has been made. The performance test that will be carried out is the confusion matrix. This confusion matrix provides an overview of how the model predicts the target class in the test data by breaking down the prediction results into four main components, namely True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Through this confusion matrix, we can generate various evaluation metrics such as accuracy, precision. The following presents the confusion matrix as in Equations (4) – (7).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{……..(4)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{……..(5)}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{……..(6)}$$

$$F1\text{-}score = \frac{2TP}{2TP+FP+FN} \quad \text{……..(7)}$$

RESULTS AND DISCUSSION

From the correlation analysis between the numerical variables in the dataset, mainly related to the average temperature over the year, it can be concluded that there is a strong positive correlation between the average temperatures of January and February, as well as between the average temperatures of December and January. On the other hand, significant negative correlations are seen between the average temperatures of June and July, and between the average temperatures of July and August. These findings provide insight into monthly temperature patterns and can be the basis for various applications, such as the prediction of future average temperatures or further understanding of annual weather patterns. The results of the correlation analysis can be seen in Figure 2.

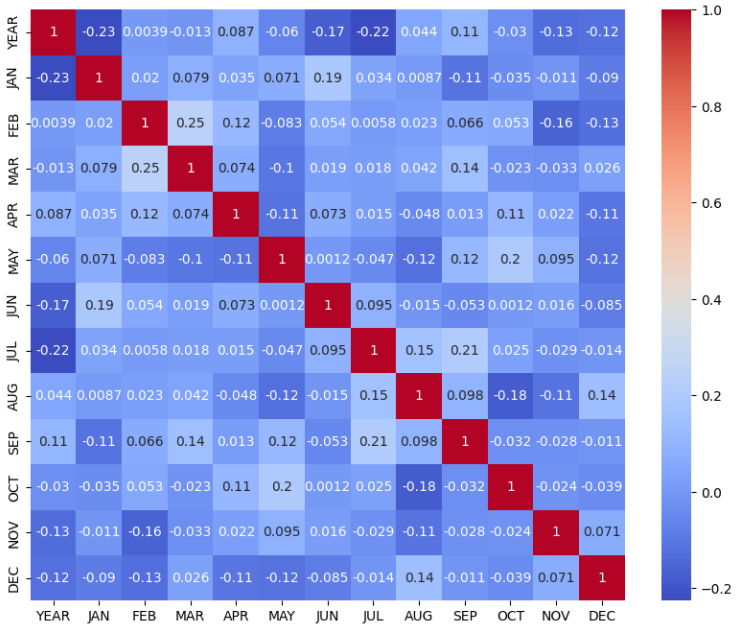

Figure 2. Correlation table

Furthermore, an analysis of the highest and lowest rainfall over the last few years is shown. Based on the graph in Figure 3, it can be concluded that rainfall has become more extreme in recent years. This can be attributed to climate change leading to increased temperatures and flooding. For Figure 4, it can be concluded that the rainfall is quite low and can cause no flooding, but the risk of drought is long.

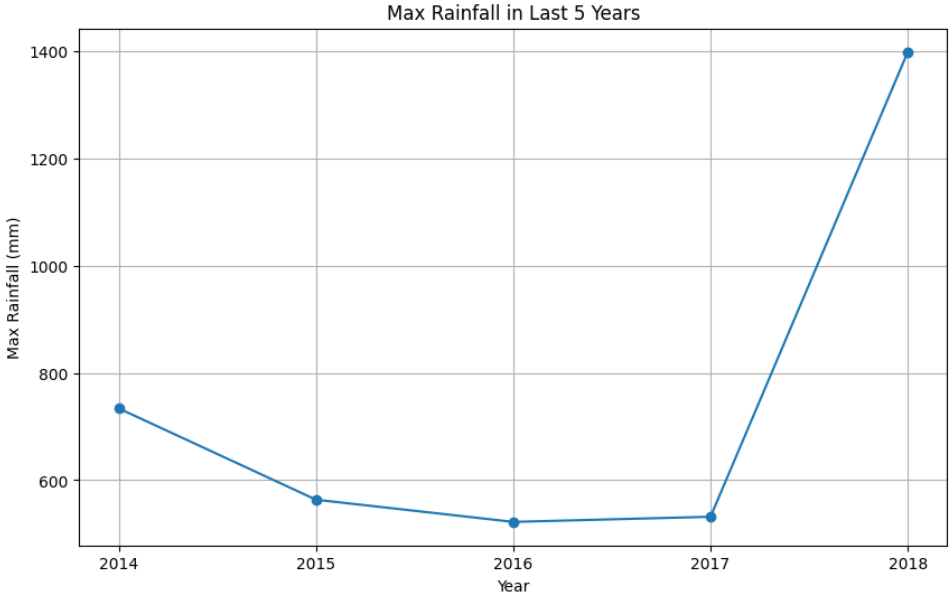

Figure 3. Highest Rainfall

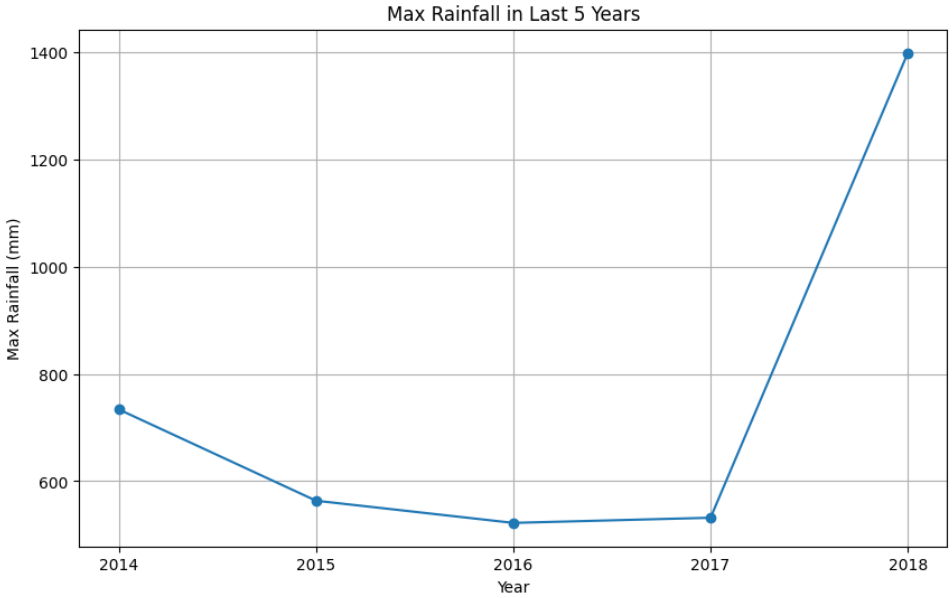

Figure 4. Highest Rainfall

The results of flood prediction from rainfall have an accuracy of 95.8%. Here, the accuracy results that I got in Figure 1 and Figure 2 show the evaluation of the results with precision, recall, and F1-score.

```
print('Random Forest:', accuracy_rf)
```

```
Random Forest: 0.9583333333333334
```

Figure 5. Random Forest Accurate

After conducting the analysis, I present a comparison table between my results and previous studies that used different models, as shown in Table 1. This table is designed to provide a more complete picture of the

comparison of model performance in the two studies, clarifying the differences and similarities in the results obtained from the different approaches.

Table 1. Comparison of the accuracy of different models

Author	Methods used	Accuracy
Slamet Triyanto et al. [16]	Naïve Bayes	79.16%
Msy Aulia Hasanah et al [24]	CART	89.4%
Xingyu Yan at al [25]	ANN	87.6%
Proposed Model	**Random Forest**	**95.8%**

Based on this analysis, it can be concluded that the Random Forest algorithm is the most appropriate method to use in flood detection. This algorithm has high accuracy, is able to handle complex and unstructured data, and produces a model that is relatively simple and easy to understand. In this study, the proposed method is Random Forest, which achieved the highest accuracy of 95.8%. The comparison shows that the Random Forest method performs better in detecting flood disasters based on rainfall data than the Naïve Bayes, CART, and ANN methods used in previous studies.

CONCLUSION

This research uses the Random Forest algorithm to detect flooding based on rainfall data. The research stages include data collection and preprocessing, model training, model evaluation, and metric calculation. The results showed that the accuracy of the model reached 95.8%, with Precision 99.9%, Recall 92.3%, and F1-score 96%. These findings indicate that the Random Forest algorithm has the potential to be developed as an effective flood early warning system.

REFERENCES

[1] M. Dhanushree, S. Chitrakala, dan C. M. Bhatt, "Robust human detection system in flood related images with data augmentation," *Multimed. Tools Appl.*, hal. 10661–10679, 2023.

[2] M. Akter, D. Cumming, dan S. Ji, "Natural disasters and market manipulation," *J. Bank. Financ.*, vol. 153, 2023, doi: 10.1016/j.jbankfin.2023.106883.

[3] T. Sharma, A. Pal, A. Kaushik, A. Yadav, dan A. Chitragupta, "A Survey on Flood Prediction analysis based on ML Algorithm using Data Science Methodology," *IEEE Delhi Sect. Conf.*, 2022, doi: 10.1109/DELCON54057.2022.9753396.

[4] H. D. Nguyen, "GIS-based hybrid machine learning for flood susceptibility prediction in the Nhat Le – Kien Giang watershed , Vietnam," *Earth Sci. Informatics*, hal. 2369–2386, 2022, doi: https://doi.org/10.1007/s12145-022-00825-4.

[5] A. Mosavi, P. Ozturk, dan K. Chau, "Flood Prediction Using Machine Learning Models : Literature Review," *water Rev.*, hal. 1–40, 2018, doi: 10.3390/w10111536.

[6] L. Yang, J. Li, A. Kang, S. Li, dan P. Feng, "The Effect of Nonstationarity in Rainfall on Urban Flooding Based on Coupling SWMM and MIKE21," *Water Resour. Manag.*, 2020.

[7] L. A. V Brito, R. I. Meneguette, R. E. De Grande, dan C. M. Ranieri, "FLORAS : urban flash-flood prediction using a multivariate model," *Appl. Intell.*, no. December 2022, hal. 16107–16125, 2023, doi: https://doi.org/10.1007/s10489-022-04319-0.

[8] A. Alamsyah dan T. Fadila, "Increased accuracy of prediction hepatitis disease using the application of principal component analysis on a support vector machine Increased accuracy of prediction hepatitis disease using the application of principal component analysis on a support vect," *J. Phys. Conf. Ser.*, 2021, doi: 10.1088/1742-6596/1968/1/012016.

[9] Walid dan Alamsyah, "Recurrent Neural Network For Forecasting Time Series With Long Memory Pattern," *J. Phys. Conf. Ser.*, 2017, doi: 10.1088/1742-6596/755/1/011001.

[10] A. Novandya dan I. Oktria, "Penerapan Algoritma Klasifikasi Data Mining C4 . 5 Pada Dataset Cuaca Wilayah Bekasi," *Penerapan Algoritm. Klasifikasi Data Min. C4. 5 Pada Dataset Cuaca Wil. Bekasi*, vol. 6, hal. 98–106, 2017.

[11] H. J. Keum, K. Yeun, dan H. Il Kim, "Real-Time Flood Disaster Prediction System by Applying Machine Learning Technique," *KSCE J. Civ. Eng.*, vol. 24, hal. 2835–2848, 2020, doi: 10.1007/s12205-020-1677-7.

[12] J. Jumanto, M. A. Muslim, Y. Dasril, dan T. Mustaqim, "Accuracy of Malaysia Public Response to Economic Factors During the Covid-19 Pandemic Using Vader and Random," *J. Inf. Syst. Explor. Res.*, vol. 1, no. 1, hal. 49–70, 2023.

[13] M. Schonlau dan R. Y. Zou, "The random forest algorithm for statistical learning," *random For. algorithm Stat. Learn.*, no. 1, hal. 3–29, 2020, doi: 10.1177/1536867X20909688.
[14] Irfan, R. Mardiati, dan M. R. Effendi, "Early Warning System of Flood Disaster Using JSN-SR04 and Rainfall Sensor Based on Internet of Things," *Int. Conf. Wirel. Telemat.*, 2022, doi: 10.1109/ICWT55831.2022.9935139.
[15] C. Iwendi, A. K. Bashir, A. Peshkar, dan R. Sujatha, "COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm," *Front Public Heal.*, vol. 8, no. July, hal. 1–9, 2020, doi: 10.3389/fpubh.2020.00357.
[16] S. Triyanto, A. Sunyoto, dan M. R. Arief, "Analisis Klasifikasi Bencana Banjir Berdasarkan Curah Hujan Menggunakan Algoritma Naïve Bayes," *JOISIE J. Inf. Syst. Informatics Eng.*, vol. 5, no. 2, hal. 109–117, 2021.
[17] S. Ahmed dan A. El, "Random forest and naïve Bayes approaches as tools for flash flood hazard susceptibility prediction , South Ras El-Zait , Gulf of Suez Coast , Egypt," *Arab. J. Geosci.*, hal. 1–12, 2022, doi: 10.1007/s12517-022-09531-3.
[18] K. VijiyaKumar, B. Lavanya, I. Nirmala, dan S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," *IEEE Int. Conf. Syst. Comput. Autom. Netw.*, 2019, doi: 10.1109/ICSCAN.2019.8878802.
[19] O. Okun, "Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations," *IGI Glob.*, 2011, doi: 10.4018/978-1-60960-557-5.
[20] S. Alexandropoulos, S. Kotsiantis, dan M. N. Vrahatis, "Data preprocessing in predictive data mining," *Knowl. Eng. Rev.*, no. April 2020, 2019, doi: 10.1017/S026988891800036X.
[21] E. D. Wahyuni, A. A. Arifiyanti, dan M. Kustyani, "Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining," *Explor. Data Anal. dalam Konteks Klasifikasi Data Min.*, vol. 2019, no. November, hal. 263–269, 2019.
[22] I. Oktanisa dan A. A. Supianto, "Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank Direct Marketing," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, 2018, doi: https://doi.org/10.25126/jtiik.201855958.
[23] N. L. Hanun, A. U. Zailani, P. Studi, T. Informatika, dan U. Pamulang, "Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera," *J. Technol. Inf.*, vol. 6, no. 1, hal. 7–14, 2020, doi: https://doi.org/10.37365/jti.v6i1.61.
[24] M. A. Hasanah, S. Soim, dan A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *J. Appl. Informatics Comput.*, vol. 5, no. 2, 2021.
[25] X. Yan, K. Xu, W. Feng, dan J. Chen, "A Rapid Prediction Model of Urban Flood Inundation in a High-Risk Area Coupling Machine Learning and Numerical Simulation Approaches," *Int. J. Disaster Risk Sci.*, vol. 12, no. 6, hal. 903–918, 2021, doi: 10.1007/s13753-021-00384-0.