STMIK
Amik Riau

JAIA – Journal Of Artificial Intelligence And Applications
Journal Homepage : http://jurnal.sar.ac.id/index.php/JAIA/

Comparison of Support Vector Machine and Random Forest Algorithms for Analyzing Online Loans on Twitter social media

Hamdani [1], Randi N.A [2], M. Khairul Anam [3]
[1,2,3]Teknik Informatika, Jl. Purwodadi Indah, Km 10 Panam, Pekanbaru,

Article Info

Received : January 9, 2024
Revised : March 14, 2024
Accepted : April 1, 2024

Keywords :

Online loans
Sentiment Analysis
Twitter
SVM
Random Forest

Abstract

Online loans represent a form of financial service wherein borrowers can apply for loans through digital platforms without the need to visit physical offices. The application, approval, and disbursement processes are conducted online, leveraging technology to facilitate financial access and transactions. However, some online lending services impose high-interest rates, resulting in a significant financial burden for borrowers. Moreover, there are instances of inappropriate debt collection practices, such as contacting the borrower's friends or family, leading to discussions and comments on social media platforms like Twitter. This research aims to analyze the patterns of comments in Indonesian society regarding online lending. The study utilizes sentiment analysis and compares machine learning algorithms to assess their accuracy. The algorithms employed in this study are Support Vector Machine (SVM) and Random Forest. The results indicate that the SVM algorithm achieves an accuracy of 93.85%, while Random Forest achieves an accuracy of 91.62%.

Hamdani
Email: hamdani@sar.ac.id.

1. Introduction

Online loans represent a financial service facilitated through online platforms, constituting a subset of Financial Technology (Fintech) [1]. This is evident in the utilization of the internet as a transaction medium for banking activities. Among various social media platforms, Twitter stands out as a widely used medium. Presently, numerous studies leverage Twitter as a data source for sentiment analysis.

Sentiment analysis involves seeking public opinions about a particular subject [2]. Considerable risks, such as high-interest rates, potential leakage of personal data during online loan applications, privacy implications in the debt collection process, and frequent occurrences of fraud, need careful consideration. Hence, this research aims to understand public sentiments regarding online loan applications on Twitter. The goal is to assist users in evaluating the risks associated with online loans and minimizing the occurrence of online fraud [3].

In response to the surge of illegal online loans, the Financial Services Authority (Otoritas Jasa Keuangan) has taken swift and firm actions in collaboration with the Indonesian National Police and the Ministry of Communication and Informatics. These efforts include cyber patrols and the blocking/closure of 3,193 illegal online loan applications/websites since 2018. The public is cautioned against online loan offers through SMS/WhatsApp, as they are likely to be associated with illegal online lending.

Twitter users' opinions on online loans are then processed and analyzed to provide valuable information and knowledge for the public. The analysis employs two data mining methods: Support Vector Machine (SVM) and Random Forest, categorizing sentiments into positive, negative, and neutral.

Support Vector Machine (SVM) is recognized for its effective classification capabilities. According to comparative studies on sentiment analysis methods [4], [5], [6], [7], and [8], four of them assert SVM as the superior method. SVM can minimize errors in training sets and is suitable for relatively small data samples. As a statistically grounded method, SVM offers a transparent theoretical foundation, avoiding the black-box nature of some other approaches.

On the other hand, Random Forest is an algorithm used for classifying large datasets. It involves the aggregation of trees through training on sample data [9]. The use of a larger sample data set enhances accuracy.

Motivated by the background above, the author aims to conduct sentiment analysis by comparing Support Vector Machine (SVM) and Random Forest methods regarding the currently popular online loans among the public. The research outcomes will provide a comparison and assess the accuracy of both SVM and Random Forest methods. The data will be divided into 90% for forming the sentiment classification model and 10% for testing purposes.

2. Research Methods

In conducting research, the acquisition of objective data and information is essential, serving as a guideline for the study. The availability of such data is crucial for ensuring the quality of the research outcomes. The research, initiated in December 2021, follows a methodology outlined in a flowchart, as depicted in Figure 1.

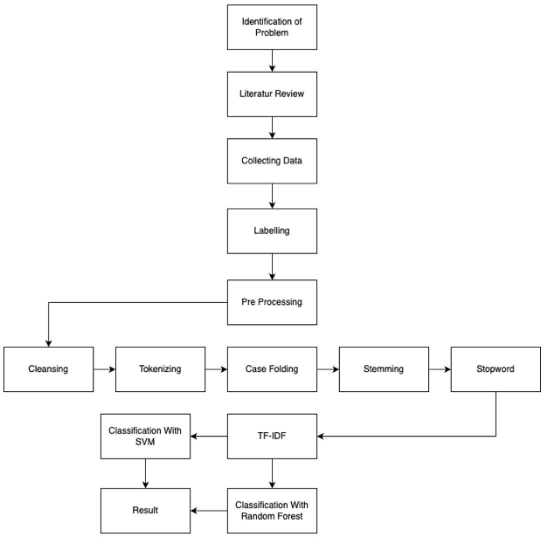


Figure 1. Research Methodology

The following are explanations of the steps in the research methodology:

3.1. Problem Identification

Observing and identifying issues related to sentiment analysis of public opinions on online loans by comparing two data mining methods, namely Support Vector Machine (SVM) and Random Forest.

3.2. Literature Review

Aiming to determine the theories that will be utilized to address the researched issues and to establish a strong foundation of references for the researcher.

3.3. Data Collection and Labelling

During the data collection phase, an analysis of the data intended for the study is conducted, considering data categories and determining the research data requirements. Once the data is obtained, it is then labeled as follows:

(a) Positive: Sentences with a positive meaning or those likely to lead readers to form a positive opinion about a specific statement, such as praises, useful information, good news, support, achievements, breakthroughs, significant events, success, and beauty.

(b) Negative: Sentences with a negative meaning or those likely to lead readers to form a negative opinion about a specific object, such as expressions of disappointment, marked by sarcastic words, rejections, denials, objections, and criticisms.

(c) Neutral: Sentences that do not take sides or support any party.

3.4. Pre-Processing

This stage involves several steps:

a. Case Folding: The process of converting all characters in the text to lowercase. Processed characters include only 'a' to 'z', and other characters such as punctuation marks (.,), commas (,), and numbers are removed [10].
b. Cleansing: The process of removing unnecessary words from comments to reduce noise [11].
c. Tokenizing: The process of breaking down sentences into words or breaking down string sequences into pieces such as words based on each word that makes them up. This restores connecting words [12].
d. Stopwords: Words that are not unique or significant in conveying a message in a sentence. These include conjunctions and adverbs that are not unique, such as "a," "by," "on," and so on. The stopword used is Tala's stopword list [13].
e. Stemming: The process of obtaining the base form of a word by removing prefixes, suffixes, infixes, and combinations of prefixes and suffixes [14].

3.5. TF-IDF

The next stage in this research involves breaking down sentences into individual words and assigning weights to each word using TF-IDF. This stage utilizes Numeric Statistics to express the level of importance of a word for a document within a collection.

3.6. Classification with SVM

In this stage, text data that has undergone previous preprocessing steps will be classified. The classification modeling process begins by dividing the data into training and testing sets based on the text categories. The modeling is implemented using the Python programming language. There are three steps in this stage [15]:

a. Determining the kernel parameter values, with C=1 and degree=20.
b. Building an SVM classification model with the polynomial kernel function and testing the classification results. The polynomial kernel is employed for dimensions of 3 or more, and since the data in this study has more than 3 dimensions, the polynomial kernel is used.
c. Calculating the accuracy of the former model (classification).

3.7. Classification With Random Forest

The random forest method needs to determine the number (m) of predictor variables randomly selected and the number (k) of trees to be formed for optimal results. According to [16], a recommended value for k in bagging methods is k=50, providing satisfactory results for classification problems. Additionally, [17] suggests that k≥100 tends to yield low misclassification rates. When using the random forest method, the choice of m significantly influences the correlation and strength of each tree. The determination of m, the number of randomly selected predictor variables, is based on the number of independent variables (p). The formulas are as follows [18]:

1. For classification, m is determined by using $\left|\sqrt{p}\right|$ with the condition that the minimum node size is 1.
2. For regression, m is determined by using p/3 with the minimum node size being 5. According to [19], there are three methods to determine m by observing error:

$$m = \frac{1}{2}\left|\sqrt{p}\right|$$
$$m = \left|\sqrt{p}\right|$$
$$m = 2 \times \left|\sqrt{p}\right| \qquad (1)$$

Where: p = total variables
Appropriate use of m can result in a random forest with sufficiently small correlations between trees. However, each tree exhibits considerable strength, as indicated by achieving a low error rate. The response of an observation is predicted by aggregating the predictions of k trees. In classification problems, this aggregation is based on the majority vote.

3. Results and Discussion

The data is extracted from tweets on the social media platform Twitter. This data is identified and categorized into words containing positive, negative, or neutral sentiments. The classification process involves using the Support Vector Machine (SVM) and Random Forest methods. The dataset is collected from the drone emprit and consists of 1000 records.

Out of the 1000 records, 1000 are used as data for building the classification model, manually labeled. This dataset is split into training and testing data in a 90%:10% ratio. To enhance prediction decision-making, the data used to build the model is resampled in the Python programming language, resulting in a total of 1785 records.

The following is a pie chart representing the sentiment proportions regarding online loans from the 1000 tweets. Sentiments include 34.0% positive, 59.5% negative, and 6.5% neutral.

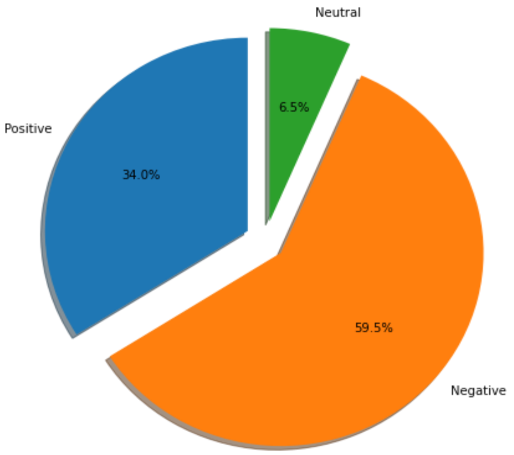


Figure 2. Pie chart of sentiment proportions

In conducting sentiment analysis on online loans, several stages are undertaken. In this document, the researcher provides details for only 10 data points out of the total 1000, as the large dataset necessitates a realistic approach of sampling a small percentage (1%) for inclusion in the report. The following outlines the steps involved in conducting sentiment analysis using the Support Vector Machine and Random Forest methods.

Tabel 1. Data obtained

No	Tweet
1	Ku kan menghilang jauh dari mu Tak terlihat sehelai rambutpun Tapi dimana nanti kau terluka Cari pinjaman online cepat dengan bunga ringan? Di @THEFWCORP bisa beli leot pakai funpaylater, ayo tunggu apa lagi borong itemnya sekarang jugaÃ°ÂŸÂ˜Â‰ https://
2	@ntijamet @anakpagiarto @sekarp110197 @ezash Org Ambon, kebanyakan jadi debt kolektor, penagih pinjol itu org ambon , maaf bukan rasis tapi emg kenyataan begitu
3	Maaf bgt jika keluar topik. aku mohon infonya kok aku jadi dpt tf gini ya?? Aku takut bgt, 2/3minggu sekali dpt, dan itu bisa 50k / 300k. Aku ga ada pinjol samsek. Tolongg aku!! Work! https://t.co/Br4redhQl5
4	@Shibapump98 @DOGEZILLACOIN Lah uang dingin ko panik. Situ pinjol kali JD panik sellÃ°ÂŸÂ¤Â£Ã°ÂŸÂ¤Â£Ã
5	@tandatanya3kali Asal jangan scroll aplikasi pinjol wkwkw
6	RT Kalau negara ini memang serius ingin memberantas pinjol ilegal, pemerintah melalui Kominfo cukup kirimkan list pinjol yg terdaftar di OJK ke Google. Selain yg ada di list tersebut, minta google untuk takedown yg ada di play store. Selanjutnya reject
7	@bravepromote Nazarku gakbakalan pake pinjol lagi pas kepepetÃ°ÂŸÂ¥Â²
8	@Freedomland414l Harus semangat, harus Rajin bekerja , biar Jauh Dari Pinjol Riba .. #KM50MonumenKebiadaban #KM50MonumenKebiadaban https://t.co/TpyTGMvF6G
9	asik banyak data buat daftar pinjol
10	@sakhansaaa @jadijago @gopayindonesia Daftarnya lebih ribet dari pinjol

In Table 1, it is explained that the acquired documents are labeled text documents. At this stage, the data is still intact or not yet cleaned, so the test data to be processed still contains other characters that are attached to the data.

3.1. Process on Google Colab

The following are the steps for data processing on Google Colab:

a. First, open the Google Colab tools on a web browser or through a Google search.
b. Ensure that you have logged in using a Google account.
c. Several feature options will appear, such as example menu, recent, Google Drive, GitHub, upload. Choose the menu according to the needs to access the prepared code.
d. Upload the prepared code, and the code will be uploaded to Google Colab.
e. After uploading the code, import the data to be processed (existing Twitter data).
f. Wait until the data import process is complete.
g. Check and run the code, wait for the code to run. If successful, a checkmark notification will appear, indicating that the code is running.

h. Repeat until all processes are successful.

3.2. Term Weighting (TF-IDF)

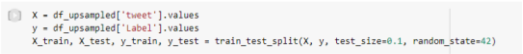
In Tables 4.2 and 4.3, the steps for calculating and obtaining information on TF (Term Frequency), DF (Document Frequency), and IDF (Inverse Document Frequency) are explained. It calculates the document or term based on the frequency of the appearance of that term or document. The probability of the appearance of words/terms in one document (D1 to D10) is calculated. To obtain IDF, the following equation is used:

$$IDF = \log (D/DF) \qquad (2)$$

Then, a document information table is created containing the frequency of words (TF), document frequency (DF), and IDF for each term. Then, the TF-IDF value for each term is calculated. The IDF value used is the IDF value obtained after the system training process.

3.3. SVM and Random Forest Classification

Preprocessing the available data, totaling 1785 divided into 3 sentiment classes: positive, negative, and neutral. Before entering the classification machine, the data is divided into training and testing data, with a testing ratio of 90:10. The process of splitting in Python programming is shown in Figure 3.

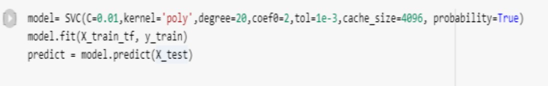


```
X = df_upsampled['tweet'].values
y = df_upsampled['Label'].values
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=42)
```

Gambar 3. Split in Python programming

A. Classification Using SVM

The data to be trained and tested has been divided, and the experiment uses the Support Vector Machine (SVM) method. Classification of the divided feature vector data is done using SVM with a polynomial kernel function that maps non-linear data to obtain a new dataset learning model in each experiment. The primary task in the experiment is the selection of parameters for SVM learning machines and the polykernel function. The important thing is to choose the parameters for SVM classification machines and polykernel functions, namely parameters C and d (degree). If C > 0 or C < 0 is a free parameter in the polynomial. When C = 0, the kernel is called homogeneous. The value of C in this study is 0.01, and the degree is 20. The process of the support vector machine (SVM) is shown in Figure 4.

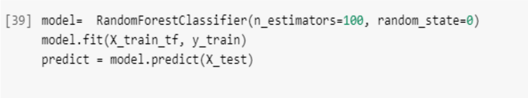


```
model= SVC(C=0.01,kernel='poly',degree=20,coef0=2,tol=1e-3,cache_size=4096, probability=True)
model.fit(X_train_tf, y_train)
predict = model.predict(X_test)
```

Gambar 4. Proses support vector machine (SVM)

B. Classification Using Random Forest

Random Forest is a combination of existing decision tree techniques, which are then merged and combined into a model. There are three main points in the Random Forest method: (1) performing bootstrap sampling to build prediction trees; (2) each decision tree predicts with random predictors; (3) then Random Forest makes predictions by combining the results of each decision tree by majority vote for classification or averaging for regression. The data that has been split will be classified using the Random Forest method. The process of the random forest is shown in Figure 5.



```
[39] model=  RandomForestClassifier(n_estimators=100, random_state=0)
     model.fit(X_train_tf, y_train)
     predict = model.predict(X_test)
```

Gambar 5. Proses Random Forest

The results of the classification learning model with testing experiments obtained a matrix with a size of 3x3 as a representative of actual and predicted classes. After obtaining the results of the machine learning classification model of the support vector machine (SVM) and Random Forest, testing experiments are conducted using the testing data that has been split, where the results of the classification test obtained a matrix with a size of 3x3 as a representation of an actual class and predicted class. The results of these machine learning models are then tested using new data that has not been used before.

Tabel 2. Confusion matrix

Actual Data	Predicted NegativeData		
	Negative	Neutral	Positive
Negative	TNg	NgN	FN
Neutral	NNg	TN	NP
Positive	FP	PN	TP

In Table 2, the confusion matrix shows the results of predictions using SVM and Random Forest classification machines, measuring the performance of each class by calculating precision, recall, and F1-score.

Precision is used to calculate the accuracy of the predicted class according to the actual class for accuracy results. Then, Recall is used to measure the sensitivity of the measurement to the dataset or the system's predictive ability according to the level of truth to recall relevant documents for each class word.

C. Comparison of Accuracy Result

Data that has gone through the preprocessing stage is then randomly divided into training and testing data. The total number of data is 1000, with 900 data used for training in SVM and Random Forest classification modeling and 100 data as testing. The data is then resampled, so unbalanced data is balanced first so that the total data becomes 1785. The data division of 1785 is done with a 90% training data and 10% testing data ratio.

The next step is to build a classification model using the Support Vector Machine method and Random Forest on the training data. The kernel function on SVM used is a polynomial. The SVM model obtained is then tested on testing data to see the accuracy of the model classification. The accuracy of the classification can be determined by evaluating the model. Table 4.14 shows the model evaluation.

Tabel 4. 1 ***Evaluasi model***

Method	Evaluation		
	Training	*Testing*	Difference
SVM	0.995018	0.938547	0.056471
Random Forest	0.995018	0.916201	0.078817

Based on Table 4.5, it can be observed that the accuracy with SVM is 0.995018, indicating that the SVM model with the polynomial kernel function correctly predicts at a rate of 99.7971%. The polynomial model does not experience overtfitting or underfitting because the training accuracy value is not significantly different from the testing accuracy value. Similarly, the accuracy using the Random Forest method has a difference of 0.078817%. Here is a comparison of the output from models using the SVM and Random Forest methods.

Tabel 4. 2 Comparison Result

No	Method	Accuracy	Precision	Recall	F1-score
1	SVM	93.85	94.00	93.00	93.33
2	Random Forest	91.62	91.00	91.33	91.00

4. Conclusion

Based on the conducted research, the conclusions are as follows:

1. The poly kernel with C=0.01 and degree=20 on 1000 tweets yields a model formation accuracy of 93.85% for the Support Vector Machine method and 91.62% for Random Forest, with 90% training data and 10% testing data.
2. The positive, neutral, and negative class percentages for the Support Vector Machine method are 0.88%, 1.00%, and 0.92%, respectively.

3. The positive, neutral, and negative class percentages for the Random Forest method are 0.88%, 0.99%, and 0.86%, respectively.
4. The total data count is 1000, with 900 data used for SVM classification modeling training and 100 data for testing. The data is then resampled to balance the imbalanced data, resulting in a total of 1785 data.
5. The data division of 1785 is done with a ratio of 90% training data and 10% testing data. The SVM accuracy is 0.995018, indicating that the SVM model with a poly kernel function correctly predicts 99.7971%.
6. The poly model does not experience overfitting or underfitting because the training accuracy is not far from the testing accuracy. Similarly, the accuracy using the Random Forest method has a difference of 0.078817%.

Subsequently, recommendations for this research are:

1. Using a larger dataset would undoubtedly enhance the classification model.
2. The addition of techniques for high accuracy is essential, necessitating the utilization of various additional techniques or methods to improve accuracy.
3. Future research is expected to use more than two methods for comparison or just one method for classification.

5. Reference

[1] J. Z. Y. Arvante, “Dampak Permasalahan Pinjaman Online dan Perlindungan Hukum Bagi Konsumen Pinjaman Online,” *Ikatan Penulis Mahasiswa Hukum Indonesia Law Journal*, vol. 2, no. 1, pp. 73–87, Feb. 2022, doi: 10.15294/ipmhi.v2i1.53736.

[2] M. K. Anam, M. I. Mahendra, W. Agustin, Rahmaddeni, and Nurjayadi, “Framework for Analyzing Netizen Opinions on BPJS Using Sentiment Analysis and Social Network Analysis (SNA),” *Intensif*, vol. 6, no. 1, pp. 2549–6824, 2022, doi: 10.29407/intensif.v6i1.15870.

[3] T. P. Lestari, “Analisis Text Mining pada Sosial Media Twitter Menggunakan Metode Support Vector Machine (SVM) dan Social Network Analysis (SNA),” *Jurnal Informatika Ekonomi Bisnis*, pp. 65–71, Aug. 2022, doi: 10.37034/infeb.v4i3.146.

[4] E. P. Nuansa, “Analsis Sentimen Pengguna Twitter Terhadap Pemilihan Gubernur Dki Jakarta Dengan Metode Naïve Bayesian Classification Dan Support Vector Machine,” 2017.

[5] G. A. Buntoro, “Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter,” *Integer Journal Maret*, vol. 1, no. 1, pp. 32–41, 2017.

[6] S. Kurniawan, W. Gata, D. A. Puspitawati, Nurmalasari, M. Tabrani, and K. Novel, “Perbandingan Metode Klasifikasi Analisis Sentimen Tokoh Politik Pada,” *Resti*, vol. 1, no. 10, pp. 2–8, 2019.

[7] N. Saputra, B. T. Adji, and E. A. Permanasari, “Analisis Sentimen Data Presiden Jokowi dengan Preprocessing Normalisasi dan Stemming Menggunakan Metode Naive Bayes dan SVM,” *Jurnal Dinamika Informatika*, vol. 5, no. November, p. 12, 2015.

[8] G. A. Buntoro, “ANALISIS SENTIMEN HATESPEECH PADA TWITTER DENGAN METODE NAÏVE BAYES CLASSIFIER DAN SUPPORT VECTOR MACHINE,” *J Chem Inf Model*, vol. 53, no. 9, pp. 1689–1699, 2016, doi: 10.1017/CBO9781107415324.004.

[9] M. A. Abubakar, M. Muliadi, A. Farmadi, R. Herteno, and R. Ramadhani, “Random Forest Dengan Random Search Terhadap Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung,” *Jurnal Informatika*, vol. 10, no. 1, pp. 13–18, Mar. 2023, doi: 10.31294/inf.v10i1.14531.

[10] M. K. Anam, Rahmaddeni, M. B. Firdaus, H. Asnal, and Hamdani, “Sentiment Analysis to analyze Vaccine Enthusiasm in Indonesia on Twitter Social Media,” *JAIA – Journal Of Artificial Intelligence And Applications*, vol. 1, no. 2, pp. 23–27, 2021.

[11] Junadhi, Agustin, M. Rifqi, and M. K. Anam, “Sentiment Analysis of Online Lectures using K-Nearest Neighbors based on Feature Selection,” *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 3, pp. 216–225, Dec. 2022, doi: 10.23887/janapati.v11i3.51531.

[12] M. K. Anam *et al.*, “Sentiment Analysis for Online Learning using The Lexicon-Based Method and The Support Vector Machine Algorithm,” *ILKOM Jurnal Ilmiah*, vol. 15, no. 2, pp. 290–302, 2023, doi: 10.33096/ilkom.v15i2.1590.290-302.

[13] F. Z. Tala, “Pembelajaran stemming pada bahasa indonesia,” 2003.

[14] R. S. Putra, W. Agustin, M. K. Anam, L. Lusiana, and S. Yaakub, “The Application of Naïve Bayes Classifier Based Feature Selection on Analysis of Online Learning Sentiment in Online Media,” *Jurnal Transformatika*, vol. 20, no. 1, p. 44, Jul. 2022, doi: 10.26623/transformatika.v20i1.5144.

[15] P. S. Ayu, “Analisakompetitifsosialmediamenggunakanmetodeklasifikasi naive bayesdan supportvector machine,” 2018.

[16] N. K. Dewi, U. Dyah Syafitri, and S. Y. Mulyadi, “Penerapan Metode Random Forest Dalam Driver Analysis,” *Forum Statistik dan Komputasi*, vol. 16, no. 1, pp. 35–43, 2011.

[17] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, “Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, pp. 393–399, Apr. 2021, doi: 10.29207/resti.v5i2.3008.

[18] A. G. Rakhmat and W. Mutohar, “Prakiraan Hujan menggunakan Metode Random Forest dan Cross Validation,” *MIND (Multimedia Artificial Intelligent Networking Database) Journal*, vol. 8, no. 2, pp. 173–187, 2023, doi: 10.26760/mindjournal.v8i2.173-187.

[19] E. Christy and K. Suryowati, “Analisis Klasifikasi Status Bekerja Penduduk Daerah Istimewa Yogyakarta Menggunakan Metode Random Forest,” *Jurnal Statistika Industri dan Komputasi*, vol. 6, no. 1, pp. 69–76, 2021.