E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1025 - 1032

PERSONALITY CLASSIFICATION OF SOCIAL MEDIA USERS BASED ON TYPE OF WORK AND INTEREST IN INFORMATION

Rizky Yudha Pratama[1)], Sri Suryani Prasetyowati[2)], and Yuliant Sibaroni[3)]
[1,2,3)]School of Computing, Telkom University
Jl. Telekomunikasi No. 1, Bandung, Indonesia
e-mail: yudapratama@student.telkomuniversity.ac.id[1)], srisuryani@telkomuniversity.ac.id[2)],
yuliant@telkomuniversity.ac.id[3)]

ABSTRACT

Social media is a platform that makes it easier for users to interact and get to know each other because in social media there are profiles, statuses, and user uploads. Therefore, many studies utilize social media because there is much information that can be explored on social media, one of which is research on the personality classification of social media users. However, many studies related to personality classification of social media users have failed due to too many model target classes, which result in low accuracy. In this research, the author uses the Myers-Briggs Type Indicator (MBTI) model, which is focused on only two personality classes, namely "Introvert/Extrovert" and "Sensor/Intuitive" with the features type of work and interest in information which are feature representations of the personality class used to reduce the target class with Decision Tree classification method. The best accuracy result is 95.87% after classifying using two personality classes.

***Keywords**: Decision tree, Myers-Briggs Type Indicator (MBTI), personality classification, social media.*

I. INTRODUCTION

SOCIAL media is online media with users who can easily participate, share, and create multi-blog content, social networks, wikis, forums, and virtual worlds. Blogs, social networks, and wikis are the most common forms of social media used by people in the world [1]. Every social media user can interact with other users even though they do not know each other. Therefore, social media users will find many types of personalities from their profiles, status, and uploads.

Personality is the way a person reacts and interacts with others. Personality is a combination of individual behaviors, emotions, motivations, and mindset characteristics [2]. To find out the personality itself, the usual method is to interview or conduct a personality test. However, in recent years, deep learning has made significant progress in the field of natural language processing in text modeling [3], which makes personality classification unnecessary by interviewing or conducting personality tests, simply using existing data on social media.

Classifying the personality of a social media user is a challenge, especially for someone who is not known just by looking at their profile, status, and uploads because a person's personality is very diverse. Models commonly used in personality classification are the Big Five Model and the Myers-Briggs Type Indicator (MBTI). Some studies that use the Big Five Model as used in research [4] with the Naïve Bayes classification method resulting in 63% accuracy, K-Nearest Neighbors (KNN) resulting in 60% accuracy, and Support Vector Machine (SVM) with 61% accuracy, research [5] with the Support Vector Machine (SVM) classification method resulting in 76.23% accuracy, and research [7] with the Naïve Bayes classification method resulting in 97.83% accuracy, Decision Tree resulting in 95.56% accuracy, and Support Vector Machine (SVM) with 95.56% accuracy. Whereas in research [6], using the Myers-Briggs Type Indicator (MBTI) model with the Naïve Bayes classification method resulted in an accuracy of 80%. From the results of several previous studies, there is no certainty about which model is the best in describing personality [7] because the accuracy produced by each classifier method is low due to the target class (personality).

Therefore, the author of this study will conduct personality classification using the MBTI model, which is focused on only two classes, namely Introvert/Extrovert and Sensor/Intuitive to reduce the number of target classes for higher accuracy. The features or attributes used are the type of user's work and the user's interest in information which represents the features or attributes of the Introvert/Extrovert and Sensor/Intuitive classes [8] with Decision Tree classification method. This classification method was chosen because Decision Tree is one of the supervised learning methods and has good accuracy based on previous research [7], so it is relevant to the research topic raised by the author. To measure the method's performance, the author uses confusion matrix to calculate the accuracy and f1-score of the model built.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1025 - 1032

II. RESEARCH METHOD

The personality classification process based on the type of work and interest in information is carried out in several stages from start to finish. In detail, these stages can be seen in Figure 1.

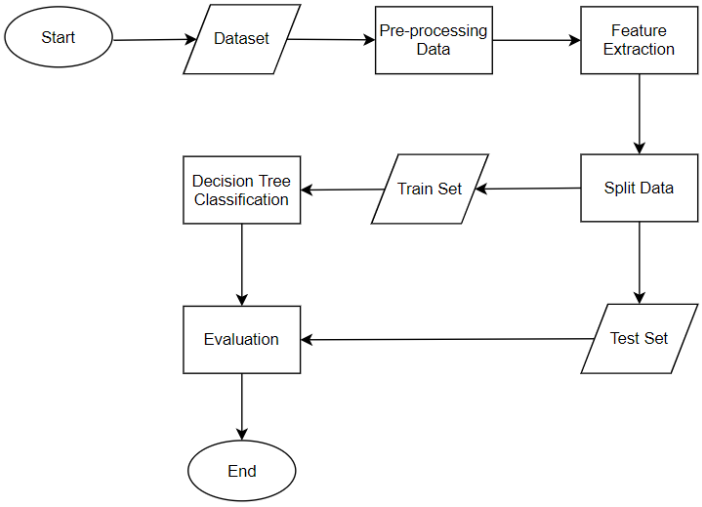

Figure 1. Personality Classification System Design

A. Dataset

The dataset used in this research is obtained from the kaggle.com website, which has been labeled based on the Myers-Briggs Type Indicator (MBTI) type and added from the results of crawling social media data based on the results of responses to surveys given to respondents. The survey provided contains the respondent's name and username for data crawling and 60 online personality test questions for data labeling. Data was crawled with twint tools with the keyword 'job' and retweets set True. Preview of the dataset can be seen in Table 1.

TABLE I
DATASET PREVIEW

Type	Posts
ENTJ	Now I'm interested. But too lazy to go research it, because it's time-consuming :(Welcome to the club, mate!
ENTJ	Still going strong at just over the two years mark. I have made noticeable changes and do not plan on slowing. I have attached my 2 year progress picture, but with my face cropped out, you know to
INFP	Personally, I was thinking this would be more of an SJ type job in a ways. I was having some issues a while back finding a job.
ENFP	He doesn't want to go on the trip without me, so me staying behind wouldn't be an option for him. I think he really does believe that I'm the one being unreasonable.
INTP	Software development. My job is terrible for me because most of it is support-related. I need to be creating something interesting in order to be fulfilled by my work and I can so rarely do that.

B. Pre-processing

Before the data is used to build a classification model, the dataset is first preprocessed. Data preprocessing is used to convert raw data into data that can be processed to facilitate the retrieval of the information contained. The steps in the data preprocessing stage are as follows.

1. Cleaning, a process to remove noise such as usernames, hashtags, special characters, and numbers.
2. Casefolding, a process to replace words in textual data with lowercase.
3. Tokenization, a process to break a sentence into the tokens or words that compose it.
4. Stemming, a process to separate prefix, infix, suffix, and confix in derivative words into basic words. The stemmer used in this study is porterstemmer from Natural Language Toolkit (NLTK).
5. Stopwords, a process to eliminate non-topic words to help reduce irrelevant features. The stopwords used in

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1025 - 1032

this study is the English stopwords dictionary from Natural Language Toolkit (NLTK).

Table 2 is an example of the application of each step in data preprocessing.

TABLE II
STEPS AND RESULTS PRE-PROCESSING

Steps	Posts
Before Pre-processing	Software development. My job is terrible for me because most of it is support-related. I need to be creating something interesting in order to be fulfilled by my work and I can so rarely do that.
Cleaning	Software development My job is terrible for me because most of it is support-related I need to be creating something interesting in order to be fulfilled by my work and I can so rarely do that
Casefolding	software development my job is terrible for me because most of it is support-related i need to be creating something interesting in order to be fulfilled by my work and i can so rarely do that
Tokenization	[software, development, my, job, is, terrible, for, me, because, most, of, it, is, supportrelated, i, need, to, be, creating, something, interesting, in, order, to, be, fulfilled, by, my, work, and, i, can, so, rarely, do, that]
Stemming	[software, development, my, job, is, terrible, for, me, because, most, of, it, is, supportrelate, i, need, to, be, create, something, interest, in, order, to, be, fulfill, by, my, work, and, i, can, so, rarely, do, that]
Stopwords	[software, development, job, terrible, supportrelate, need, create, interest, order, fulfill, work, rarely]

C. Feature Extraction

Feature extraction is a process in text classification to convert unstructured textual formats into structured ones so that they can be processed by machine learning algorithms to be classified into predetermined classes [9]. The features used in this study are the type of work and the user's interest in information which is a representation of the attributes of the "Introvert/Extrovert" and "Sensor/Intuitive" classes in the Myers-Briggs Type Indicator (MBTI) model [8]. The user's interest in this information is determined by what information the user reposts or retweets.

While the weighting method used in this study is TF-IDF because it is efficient, easy, and has accurate results [10]. TF is the frequency of occurrence of words in a sentence. Meanwhile, IDF itself has the following basic formula.

$$IDFdt = log(\frac{D}{Dt}) \quad (1)$$

D is the total data, and Dt is the number of words t that appear in one data D. The weighting formula with TF-IDF is as follows.

$$Wdt = TFdt * IDFdt \quad (2)$$

Description:
Wdt = value or weight of word t in document d
TFdt = frequency of occurrence of word t in document d
IDFdt = Inverse Document Frequency

Tables 3 and 4 are the examples of documents from data collection that have been preprocessed and examples of TF-IDF calculations.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1025 - 1032

TABLE III
SAMPLE DOCUMENTS

Documents	Posts
D1	software development job need create interest order work
D2	strong mark notice change plan slow attach progress crop

TABLE IV
TF-IDF

Words	TF		IDF	TF * IDF	
	D1	D2	Log(D/Dt+1)	D1	D2
software	1/8 = 0.125	0	Log(2/1) = 0.301	0.0376	0
development	1/8 = 0.125	0	Log(2/1) = 0.301	0.0376	0
job	1/8 = 0.125	0	Log(2/1) = 0.301	0.0376	0
need	1/8 = 0.125	0	Log(2/1) = 0.301	0.0376	0
create	1/8 = 0.125	0	Log(2/1) = 0.301	0.0376	0
interest	1/8 = 0.125	0	Log(2/1) = 0.301	0.0376	0
order	1/8 = 0.125	0	Log(2/1) = 0.301	0.0376	0
work	1/8 = 0.125	0	Log(2/1) = 0.301	0.0376	0
strong	0	1/9 = 0.111	Log(2/1) = 0.301	0	0.0334
mark	0	1/9 = 0.111	Log(2/1) = 0.301	0	0.0334
notice	0	1/9 = 0.111	Log(2/1) = 0.301	0	0.0334
change	0	1/9 = 0.111	Log(2/1) = 0.301	0	0.0334
plan	0	1/9 = 0.111	Log(2/1) = 0.301	0	0.0334
slow	0	1/9 = 0.111	Log(2/1) = 0.301	0	0.0334
attach	0	1/9 = 0.111	Log(2/1) = 0.301	0	0.0334
progress	0	1/9 = 0.111	Log(2/1) = 0.301	0	0.0334
crop	0	1/9 = 0.111	Log(2/1) = 0.301	0	0.0334

D. Decision Tree

The classification method used in this research is decision tree. Decision tree is a method that consists of nodes and leaves [11]. While in general, a decision tree is a hierarchical model consisting of discriminant functions applied by partitioning the feature space of a data set into a single, purely recursive subspace of classes [12]. There are several equations used in decision trees, among others.

1. Entropy
   Entropy is a formula for measuring heterogeneity (diversity) of data sets [13]. The entropy formula is as follows [14].

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi \quad (3)$$

2. Information Gain
   Information Gain is a formula for measuring the effectiveness of attributes in classifying data calculated based on entropy [13]. The Information Gain formula is as follows [14].

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(S) \quad (4)$$

3. Gain Ratio
   Gain Ratio is a formula developed from Information Gain, which has shortcomings in the form of bias [13]. The Gain Ratio formula is as follows [14].

$$GainRatio(S, A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (5)$$

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1025 - 1032

4. Split Information
   Split Information is a value that must be calculated before calculating the Gain Ratio. The Split Information formula is as follows [14].

$$SplitInfo(S,A) = -\sum_{i=1}^{n} \frac{Si}{S} \log_2 \frac{Si}{S} \quad (6)$$

E. Evaluation

System performance is based on recall, precision, accuracy, and f1-Score values.

1. Recall
   Recall is the amount of data with a positive category that is correctly classified by the system compared to all existing positive data [16]. The recall formula is as follows.

$$Recall = \frac{TP}{FN+TP} \quad (7)$$

2. Precision
   Precision indicates the amount of data with a positive category that is correctly classified by the system compared to all positive prediction data [16]. The precision formula is as follows.

$$Precision = \frac{TP}{FP+TP} \quad (8)$$

3. Accuracy
   Accuracy indicates the number of correctly classified data compared to the total data [16]. The accuracy formula is as follows.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

4. F1-Score
   F1-Score shows the harmonic average of precission and recall [16]. The F1-Score formula is as follows.

$$F1-Score = \frac{2\ x\ precision\ x\ recall}{precision+recall} \quad (10)$$

Confusion Matrix is a formula to analyze a classifier and how well it recognizes tuples from various classes. There are four terms for measuring performance using a confusion matrix, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [8]. Where TP is the number of correctly classified positive data, TN is the number of correctly classified negative data, FP is the number of positive but incorrectly classified data, and FN is the number of incorrectly classified negative data [15]. The confusion matrix table can be seen in Table 5.

TABLE V
CONFUSION MATRIX

Category		**Predicted Class**	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

III. RESULT

In this study, there are several steps taken, starting from the first step of preparing the dataset to the last step, namely performance evaluation. The dataset used is obtained from the kaggle.com website, which has been labeled and added from the results of crawling data independently based on the responses to surveys given to respondents. Furthermore, the dataset is preprocessed to make it easier to get information on the dataset. After preprocessing, feature extraction is carried out so that the dataset becomes structured with the tf-idf method to give a value or

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1025 - 1032

weight to each word. Furthermore, the data is split into training and test sets with a ratio of 90:10. Then, the data is classified using the decision tree model. Here are some scenarios used in this research.

1. Scenario 1: Imbalance handling
2. Scenario 2: Split data
3. Scenario 3: Modelling for 4 personality types (IN, EN, IS, ES)

A. Scenario 1 Imbalance Handling

The dataset used has uneven data distribution. This scenario is to determine the difference in accuracy and f1-score before and after imbalance handling. Three datasets are used in this scenario, the baseline dataset, oversampled dataset, and undersampled data with the division of training data and test data with a ratio of 90:10. A comparison of the data distribution of the three datesets is shown in Figure 2.

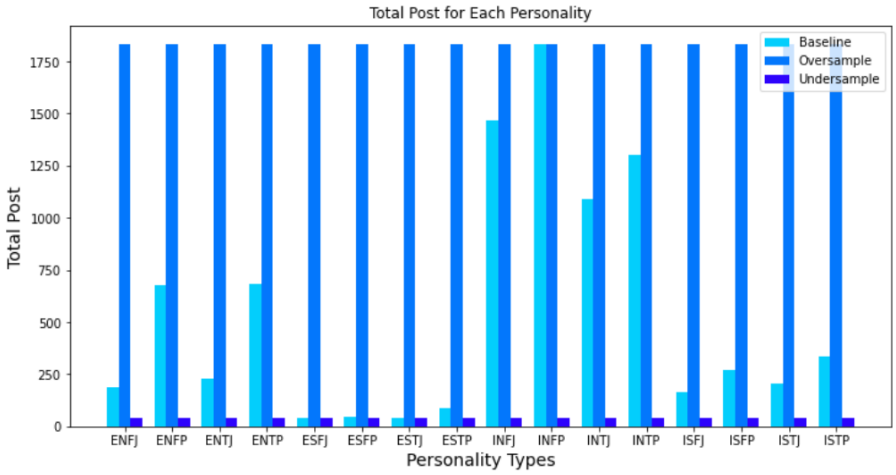

Figure 2. Comparison of Dataset Distribution

The results of testing scenario 1 can be seen in Table 6.

TABLE VI
SCENARIO 1 IMBALANCE HANDLING RESULTS

Dataset	Accuracy	F1-Score
Baseline	48.38%	30.94%
Oversample	91.81%	91.12%
Undersample	36.50%	35.32%

Based on the scenario results above, the oversampled dataset produces the highest accuracy and f1-score compared to the baseline and undersampled datasets. By using the oversampled dataset, the highest accuracy and f1-score are obtained, which are 91.81% and 91.12%.

B. Scenario 2 Split Data

In scenario 2, the dataset used is the oversampled dataset because it produces the highest accuracy and f1-score in the previous scenario test. Experiments in this scenario use a division of training and test sets with a ratio of 90:10 and are carried out five times with different random state values of 11, 13, 15, 17, and 19. The results of scenario 2 testing can be seen in Table 7.

TABLE VII
SCENARIO 2 SPLIT DATA RESULTS

Random State	Accuracy	F1-Score
11	91.64%	91.23%
13	91.47%	90.96%
15	92.19%	91.82%
17	91.88%	91.35%
19	91.91%	91.67%

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1025 - 1032

Based on the scenario results above, the highest accuracy and f1-score were obtained in the third experiment with a random state value of 15. From the five experiments conducted, the average accuracy and f1-score were 91.81% and 91.40%.

C. Scenario Modelling For 4 Personality Types

In scenario 3, the dataset used is an oversampled dataset with the division of training and test set with a ratio of 90:10 because it has the highest accuracy and f1-score in testing the previous two scenarios. In the dataset, one new column was added, a column containing personality types with only two classes that are introvert/extrovert and sensor/intuitive, resulting in 4 personality types IN, EN, IS, and ES. Table 8 is a preview of the dataset after adding one new column for 4 personality types

TABLE VIII
DATASET PREVIEW WITH 4 PERSONALITY TYPES

Type	Posts	Type4
ENTJ	Now I'm interested. But too lazy to go research it, because it's time-consuming :(Welcome to the club, mate!	EN
ENTJ	Still going strong at just over the two years mark. I have made noticeable changes and do not plan on slowing. I have attached my 2 year progress picture, but with my face cropped out, you know to	EN
INFP	Personally, I was thinking this would be more of an SJ type job in a ways. I was having some issues a while back finding a job.	IN
ENFP	He doesn't want to go on the trip without me, so me staying behind wouldn't be an option for him. I think he really does believe that I'm the one being unreasonable.	EN
INTP	Software development. My job is terrible for me because most of it is support-related. I need to be creating something interesting in order to be fulfilled by my work and I can so rarely do that.	IN

The results of testing scenario 3 can be seen in Table 9.

TABLE IX
SCENARIO 3 MODELLING FOR 4 PERSONALITY TYPES RESULTS

Model	Accuracy	F1-Score
16 Types	90.89%	90.37%
4 Types	95.87%	95.66%

Based on the scenario results above, modeling with 4 personality types yields higher accuracy and f1-score than using 16 personality types. Modeling with 4 personality types resulted in accuracy and f1-score of 95.87% and 95.66%.

After testing the three scenarios, where each scenario uses training and test sets with a ratio of 90:10, it can be concluded that each scenario test affects performance and accuracy results. In scenario 1, the results of testing with the baseline dataset and the dataset that has been done imbalance handling with oversample and undersample, resulting in the dataset that has been done imbalance handling with oversampled has the highest accuracy. In scenario 2, experiments were conducted five times with different random state values of 11, 13, 15, 17, and 19, resulting in the highest accuracy in the third experiment with a random state value of 15. In scenario 3, modeling with 4 personality types resulted in higher accuracy than using 16 personality types.

Based on the three scenarios that have been carried out, scenario 1 imbalance handling produces the highest accuracy of 91.81%. In scenario 2, split data produces an average accuracy of 91.81%. In scenario 3, modeling with 4 personality types produces the highest accuracy of 95.87%.

IV. CONCLUSION

After conducting personality classification research of social media users based on the type of work and interest in information, where the type of work is a feature or attribute representing the introvert/extrovert class, and interest in information is a feature or attribute representing the sensor/intuitive class, it can be concluded that imbalance

handling with oversampled produces the highest accuracy of 91.81% because the dataset used in this study has an uneven distribution, while imbalance handling with undersampled produces the lowest accuracy because the target class is too much and undersample reduces the training data with the majority label. Splitting the dataset into training and test sets with several experiments resulted in a not-so-significant difference in accuracy ranging only around 0-1%, with an average accuracy of 91.81%. Classifying with 4 personality types produces a higher accuracy of 95.87% compared to using 16 personality types because the target class in 16 personality types is too many, which causes lower accuracy.

Suggestions for future research, use the dataset with more balanced labels than the dataset used in this study so that the data distribution is more evenly distributed. Use other features that represent each personality type class because this research only focuses on features that represent introvert/extrovert and sensor/intuitive classes to compare the effectiveness of classification with 4 personality types and 16 personality types.

REFERENCES

[1] N. Istiani and A. Islamy, "Fikih Media Sosial Di Indonesia," *Asy Syar'Iyyah J. Ilmu Syari'Ah Dan Perbank. Islam*, vol. 5, no. 2, pp. 202–225, 2020, doi: 10.32923/asy.v5i2.1586.

[2] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2313–2339, 2020, doi: 10.1007/s10462-019-09770-z.

[3] X. Wang, Y. Sui, K. Zheng, Y. Shi, and S. Cao, "Personality classification of social users based on feature fusion," *Sensors*, vol. 21, no. 20, 2021, doi: 10.3390/s21206758.

[4] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," *Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015*, no. November, pp. 170–174, 2016, doi: 10.1109/ICODSE.2015.7436992.

[5] V. Ong *et al.*, "Personality prediction based on Twitter information in Bahasa Indonesia," *Proc. 2017 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2017*, vol. 11, pp. 367–372, 2017, doi: 10.15439/2017F359.

[6] L. C. Lukito, A. Erwin, J. Purnama, and W. Danoekoesoemo, "Social media user personality classification using computational linguistic," *Proc. 2016 8th Int. Conf. Inf. Technol. Electr. Eng. Empower. Technol. Better Futur. ICITEE 2016*, no. September, 2017, doi: 10.1109/ICITEED.2016.7863313.

[7] A. Souri, S. Hosseinpour, and A. M. Rahmani, "Personality classification based on profiles of social networks' users and the five-factor model of personality," *Human-centric Comput. Inf. Sci.*, vol. 8, no. 1, 2018, doi: 10.1186/s13673-018-0147-4.

[8] M. H. Amirhosseini and H. Kazemian, "Machine learning approach to personality type prediction based on the Myers–Briggs type indicator®," *Multimodal Technol. Interact.*, vol. 4, no. 1, 2020, doi: 10.3390/mti4010009.

[9] irwan budiman, M. R. Faisal, and D. T. Nugrahadi, "Studi Ekstraksi Fitur Berbasis Vektor Word2Vec pada Pembentukan Fitur Berdimensi Rendah," *J. Komputasi*, vol. 8, no. 1, pp. 62–69, 2020, doi: 10.23960/komputasi.v8i1.2517.

[10] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for IDF," *J. Doc.*, vol. 60, no. 5, pp. 503–520, 2004, doi: 10.1108/00220410410560582.

[11] S. Ruggieri, "Efficient C4.5," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 2, pp. 438–444, 2002, doi: 10.1109/69.991727.

[12] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *J. Chemom.*, vol. 18, no. 6, pp. 275–285, 2004, doi: 10.1002/cem.873.

[13] Suyanto, *Data mining : untuk klasifikasi dan klasterisasi data / Suyanto*. Bandung: Penerbit Informatika, 2017.

[14] D. Noviana, Y. Susanti, and I. Susanto, "Analisis Rekomendasi Penerima Beasiswa Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Algoritma C4.5," *Semin. Nas. Penelit. Pendidik. Mat. 2019 UMT*, pp. 79–87, 2019.

[15] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.

[16] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," *Data Min. Concepts Tech.*, 2012, doi: 10.1016/C2009-0-61819-5.