Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 6, No.3, July 2024, pp. 02403016-01 ~ 02403016-08
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v6i3.676

Classification of Movie Recommendation on Netflix Using Random Forest Algorithm

Alifia Salwa Salsabila*, Christy Atika Sari, Eko Hari Rachmawanto

Faculty of Computer Science, Dian Nuswantoro University, Jl. Imam Bonjol No.207,
Pendrikan Kidul Semarang, 50131, Central Java, Indonesia

*111202113245@mhs.dinus.ac.id

Abstract. Netflix is one of the most popular streaming platforms in this world. So many movies and shows with various genres and production countries are available on this platform. Netflix has their own recommendation systems for the subscribers according to their data and algorithm. This research aims to compare two methods of data classifications using Decision Tree and Random Forest algorithm and make a recommendation system based on Netflix dataset. This paper use feature importance to selecting relevant feature and how n_estimators affect the classification. In this research, Random Forest with 50 trees estimator with 96.84% accuracy before feature selection and 96.92% accuracy after feature selection has the best accuracy compared to the Decision Tree classification. Besides, Decision Tree has only 95.64% accuracy before feature selection and increases to 96.07% accuracy after feature selection. Trees estimator also affect the accuracy of Random Forest classification. After comparing the results, Random Forest with 50 trees estimators using feature selection provides best accuracy and it will be used to predict some similar movies and shows recommendation.

Keywords: Decision Tree, Feature Selection, Random Forest, Netflix, Recommendation

(Received 2024-06-06, Accepted 2024-07-09, Available Online by 2024-07-25)

1. Introduction

Netflix is one of the most popular subscription platforms to watch movies or series around the world. This platform allows their subscribers to watch various movies and shows in any genre and language all around the world. By purchasing a subscription, Netflix provides online and offline streaming from any device. Subscribers can watch any available series or movie on this platform anytime, anywhere [1]. Netflix always updates their movies or series periodically. Subscribers can enjoy watching the latest or popular movies on the internet through this platform. And Netflix not only offers popular series or movies, it also provides programs that were released decades ago. By adapting data and algorithms, Netflix constantly updates their system. Such as provides a recommendation system based on user viewing data. And with the massive data, machine learning algorithms are needed [2].

Classification is one of the Data Mining categorization or grouping techniques based on its class [3]. The technique will be useful for recommendation system implementation, so user will get some recommendations based on the groups generated after the classification process [4]. Several methods of

Data Mining can be applied to make recommendation systems based on classification algorithms. For example, Decision Tree is used to classify Ratings and make a Movie Recommendation [1], also Random Forest Classification is used with Genetic Algorithms to compare how far Genetic Algorithms impact the Music Genre Classifier [4]. To classify the movie dataset, [5] made a recommendation system based on its rating using kNN classification method. However, this recommendation system still can be developed using another algorithm to get better quality and increase the accuracy.

This research will use Decision Tree and Random Forest algorithms to develop a better recommendation system. Decision Tree Algorithm is known as an effective classification method [6]. Along with Random Forest Classification, results could be better compared to the Decision Tree method, since it can combine the most votes of various decision trees to get the best accurate results [7]. By comparing the accurate results of both algorithms, the best one will be chosen to make a recommendation system according to its genres, movie age certification, production countries, and movie release year.

2. Methods

Random Forest as a classification method and Netflix dataset from Kaggle repository are used in this research. Selection features and target will be used in this dataset. It will be divided into data train and data test with an 8:2 ratio for the classification process. Data train will be used to process the classification algorithm and will compare the accuracy of each algorithm using data test. It will be used to compare the accuracy score between Decision Tree and Random Forest Classification. The best classification will be used to develop a movie recommendation system.

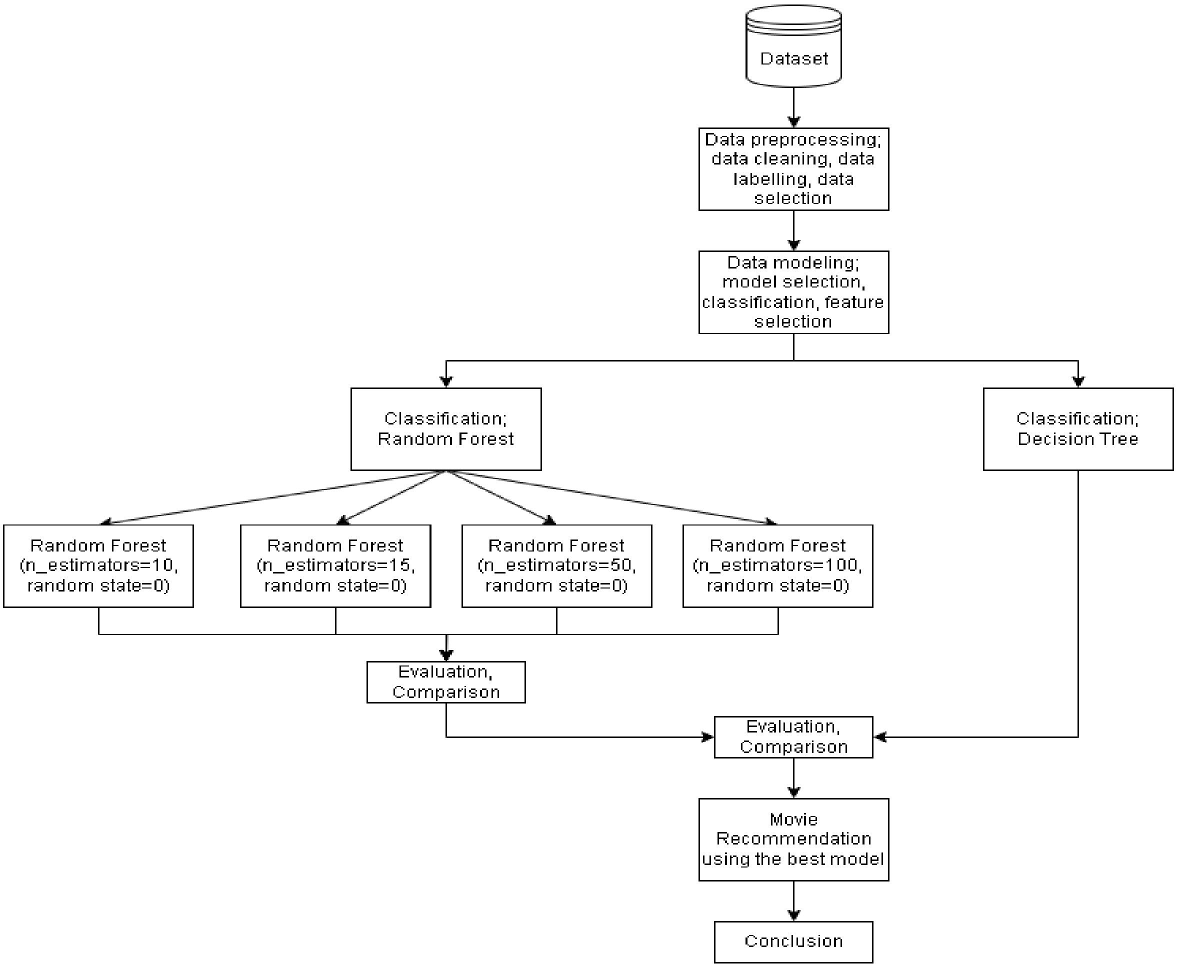

Dataset
Data preprocessing; data cleaning, data labelling, data selection
Data modeling; model selection, classification, feature selection
Classification; Random Forest
Classification; Decision Tree
Random Forest (n_estimators=10, random state=0)
Random Forest (n_estimators=15, random state=0)
Random Forest (n_estimators=50, random state=0)
Random Forest (n_estimators=100, random state=0)
Evaluation, Comparison
Evaluation, Comparison
Movie Recommendation using the best model
Conclusion

Figure 1. Research Method using Decision Tree and Random Forest Algorithm

2.1. Data Collection

This research uses a dataset from Kaggle public repository with 5850 data records and 15 fields based on the Netflix database [21]. Includes id, title, type, description, release year, age certification, runtime, genres, production countries, season, and even the imdb and tmdb information. Some data fields will be used in this research, including title as recommendation keyword based on its release year, age certification, genres, and production countries. Besides, field data type that contains two credits (movie and show) will be used as its target.

2.2. Data Preparation

Before data processing, there are several steps in preprocessing data. Such as data cleaning to filtrate data that are not included in the classification process, and data transformation to simplify the dataset. Preprocessing data steps are needed to clean the dataset from noise and missing values by removing them or initializing them with a new value [8][9]. This phase makes data more efficient, so it can develop better results. Fields with the largest null data, such as seasons data need to be removed. Another record with a null value on the title was also removed. The missing value on age certification will be initialized with "RP" value as Rating Pending, which means data has not yet received information about movie ratings. So, 5849 pieces of data will be used in this research. Target data is also transformed using label encoding to simplify the classification process. This classification process to make a movie recommendation using Random Forest will use release_year, age_certification, genres, and production_countries as its features and type as its target. It will be divided into data train and data test for accuracy testing.

2.3. Data Model

Data classification in this research will use Random Forest and Decision Tree as its methods. Decision Tree is a classification method using tree structure to analyze classification based on data train sample to classifier the rest data [10]. Besides, Random Forest Classification is an ensemble learning algorithm of decision tree by merging several trees and take the most votes to get optimal results [11][12]. By assembling predicted results from each tree to take majority votes, Random Forest could minimize misprediction and increase accuracy [13][14].

2.4. Feature Selection

Feature selection is needed to sort data relevance and relation to get better and more efficient predictions [15]. It will be used to find out which features impact the most on the classification process [16]. And according to its results, irrelevant features could be eliminated. By using feature importance, the performance of classification will be more stable and efficient [17].

2.5. Evaluation

Data evaluation is needed to compare each classification and get the best one [18]. And it will be using confusion matrix and accuracy score from scikit learn python library. Data accuracy considers the overall validation and exactness of data.

$$Accuracy = \frac{Total\ correct\ prediction}{Total\ data\ test} \tag{1}$$

Precision will be used to calculate correct data prediction positive of overall correct prediction.

$$Precision = \frac{True\ Positive}{True\ Positive + True\ Negative} \tag{2}$$

Recall data will be used to calculate correct data prediction positive of overall data positive that should be correct.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3}$$

F1-score will be used to calculate how well the classification, based on the average of precision and recall.

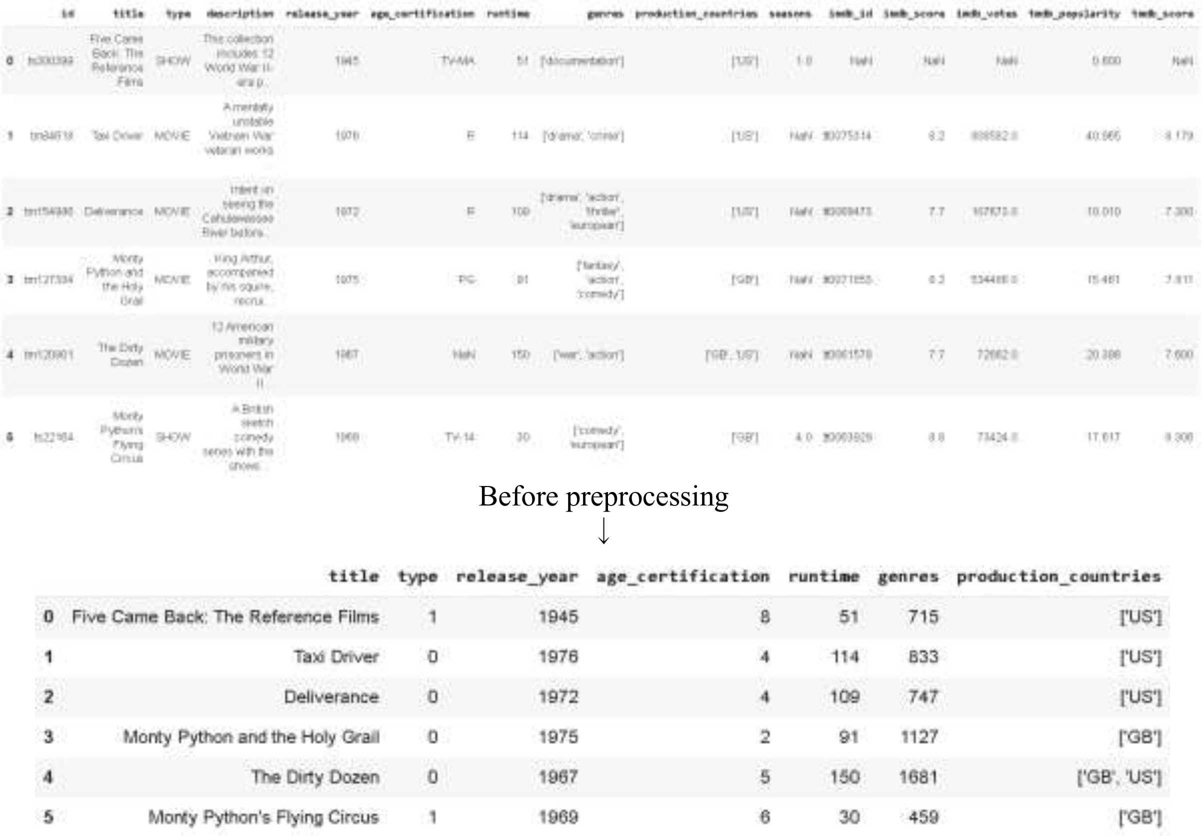
$$F1 - score = \frac{2(Precision \times Recall)}{Precision + Recall} \tag{4}$$

3. Results and Discussion

The classification process uses 4 attributes as variables such as release_year, age_certification as movie rating, and movie genres. Data type attribute is used as target data with two values, MOVIE and SHOW. The dataset is divided into data train 80% or 4679 data and data test 20% or 1170 data. Data train will be used for the classification process with Random Forest and Decision Tree algorithms, while data tests will be used for testing classification accuracy. Thereafter, the title attribute will be used to initialize the recommendation system based on the best classification.

3.1. Preprocessing Data

Before the data process, the dataset will be prepared by removing several attributes that are not included in the classification process such as id, imdb_id, and season attribute with the most missing values. Data records with a null value on the title attribute were also removed. Besides, null values in the age_certification attribute will be initialized with RP or Rating Pending. Label encoder is used to initialize data sorted by alphabetical order that will be included in the classification process [19]. Type attribute will be initialized with 0 as the movie value, and 1 as the show value. The results of preprocessing data can be seen in Figure 2.



	id	title	type	description	release_year	age_certification	runtime	genres	production_countries	seasons	imdb_id	imdb_score	imdb_votes	tmdb_popularity	tmdb_score
0	ts300399	Five Came Back: The Reference Films	SHOW	This collection includes 12 World War II-era p...	1945	TV-MA	51	['documentation']	['US']	1.0	NaN	NaN	NaN	0.600	NaN
1	tm84618	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works...	1976	R	114	['drama', 'crime']	['US']	NaN	tt0075314	8.2	808582.0	40.965	8.179
2	tm154986	Deliverance	MOVIE	Intent on seeing the Cahulawassee River before...	1972	R	109	['drama', 'action', 'thriller', 'european']	['US']	NaN	tt0068473	7.7	107673.0	10.010	7.300
3	tm127384	Monty Python and the Holy Grail	MOVIE	King Arthur, accompanied by his squire, recrui...	1975	PG	91	['fantasy', 'action', 'comedy']	['GB']	NaN	tt0071853	8.2	534486.0	15.461	7.811
4	tm120801	The Dirty Dozen	MOVIE	12 American military prisoners in World War II...	1967	NaN	150	['war', 'action']	['GB', 'US']	NaN	tt0061578	7.7	72662.0	20.398	7.600
5	ts22164	Monty Python's Flying Circus	SHOW	A British sketch comedy series with the shows...	1969	TV-14	30	['comedy', 'european']	['GB']	4.0	tt0063929	8.8	73424.0	17.617	8.306

Before preprocessing

↓

	title	type	release_year	age_certification	runtime	genres	production_countries
0	Five Came Back: The Reference Films	1	1945	8	51	715	['US']
1	Taxi Driver	0	1976	4	114	833	['US']
2	Deliverance	0	1972	4	109	747	['US']
3	Monty Python and the Holy Grail	0	1975	2	91	1127	['GB']
4	The Dirty Dozen	0	1967	5	150	1681	['GB', 'US']
5	Monty Python's Flying Circus	1	1969	6	30	459	['GB']

Figure 2. After Preprocessing Data

3.2. Model Implementation

The dataset is divided into training data for use in the classification process, and testing data to test the classification accuracy. Data classification is done using two algorithm-based methods, Decision Tree, and Random Forest. For Random Forest itself, experiments will be conducted with several n_estimators. N_estimator in Random Forest is an estimate of trees assembled in the classification process [11][20]. The more trees assemble, the more complex the classification process. The results from each method based on the classification report and accuracy score can be seen in Table 1.

Table 1. Accuracy Comparison before Feature Selection

Classification	Accuracy	Correct Prediction	Misprediction
Decision Tree	0.9564	1119	51
Random Forest (n_estimators=10)	0.9624	1132	38
Random Forest (n_estimators=50)	0.9573	1133	37
Random Forest (n_estimators=100)	0.9581	1132	38
Random Forest (n_estimators=150)	0.9675	1132	38

Based on Table 1, Random Forest with 50 n_estimator has the best accuracy than other classification methods. It has 96.84% accuracy with 1133 correct predictions of 1170 data tests. The method that has the most misprediction is the Decision Tree Classification with 51 data errors, besides Random Forest's average misprediction is only 38 data. Depending on the first experiment, Random Forest Classification relatively has better accuracy than using the Decision Tree method. To improve better classification, feature selection using feature importance is needed to consider the relevance of each feature [22].

Table 2. Accuracy Comparison after Feature Selection

Classification	Accuracy	Correct Prediction	Misprediction
Decision Tree	0.9607	1124	46
Random Forest (n_estimators=10)	0.9658	1130	40
Random Forest (n_estimators=50)	0.9684	1134	36
Random Forest (n_estimators=100)	0.9692	1133	37
Random Forest (n_estimators=150)	0.9667	1131	39

Table 2 shows that the highest accuracy still owned by Random Forest with 50 trees assembled. Compared with Table 1, Random Forest with 50 trees increases its accuracy from 0.9684 to 0.9692. Besides, other classification accuracies also changed. Such as Decision Tree Classification which had 0.9564 before feature selection and increased to 0.9607, but Random Forest Classification with 150 trees assembled has decreased its accuracy from 0.9675 to 0.9667. This indicates that increasing numbers of trees assembled do not always give better results. Comparing to Table 1, the Decision Tree Classification has increased its correct prediction. The misprediction before it gets feature selection is 51 data errors, and it decreases to 46 data errors. Aside, Random Forest also changed its accuracy with three of them having better predictions. andom Forest Classification with 50 trees assembled will be used to implement a movie recommendation system based on the Netflix dataset. The results of the movie recommendation with “Love Alarm” as its title can be seen in Table 3.

Based on the results, Classification using Random Forest Algorithm has better results with 96.76% average accuracy and it increases by eliminating the release_year feature to 96.77% average accuracy. Compared to Decision Tree Algorithm which has 95.64% accuracy and increases to 96.07% accuracy, Random Forest proves better accuracy for every experiment with different tree estimators by having higher accuracy values. Compared to the previous research, this research has increased the accuracy to make a better movie recommendation system. The previous research used Decision Tree Classification and it has only 39% average accuracy [1], another research that used kNN algorithm to make a recommendation system increased its accuracy to 45.4% [5]. This research also increased the accuracy to 96% average accuracy by using different data fields as predictor variables. It gets better after re-selecting the features by using feature importance to determine the relevance of each feature.

Table 3. Movie Recommendation based on "Love Alarm"

Title
My Liberation Notes
All Hail King Julien: Exiled
Dracula
DreamWorks Shrek's Swamp Stories
Better Call Saul
Uncle From Another World
iCarly
Selling Tampa
Crashing
Hache

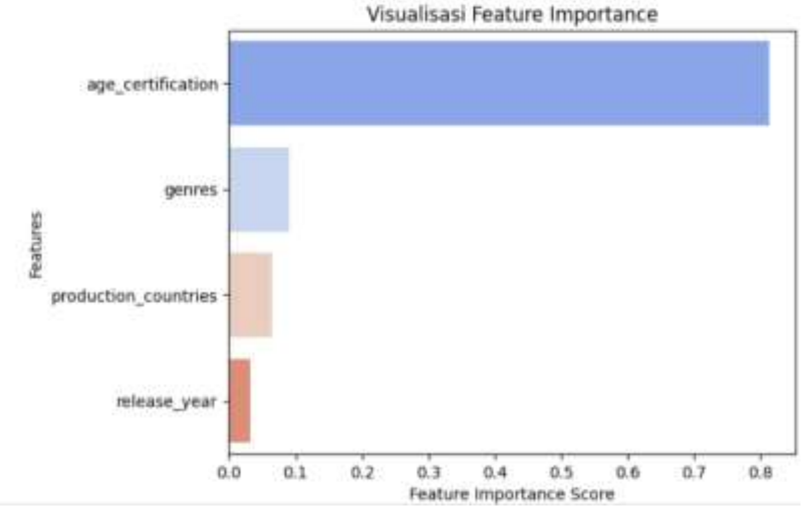


Figure 3. Feature Importance Visualization

According to Figure 3, the age_certification feature has a huge impact on this research with 0.8130 points, while the release_year feature has the lowest relevance score to the other field with 0.0315 points. Based on the result, the release_year feature needs to be eliminated to recompare its accuracy score.

4. Conclusion

Based on this research, Random Forest Classification proves that it has better results overall with 96.76% average accuracy compared to the Decision Tree method which only has 95.98% accuracy. According to this study, feature selection makes the better accuracy classification. The relevance between the features is filtered back to get better results. By using feature selection, the accuracy of Decision Tree had been increases to 96.07% and Random Forest algorithm has 96.77% accuracy. Classification with the best accuracy in this research is owned by Random Forest with 50 trees assembled. With 96.84% accuracy before feature selection, it increases to 96.92% accuracy after feature selection. Based on this paper, the more trees assembled in the classification process influence the accuracy score. However, the accuracy score is not always linearly related to the number of trees built.

Increasing the number of trees assembled will increase the classification process complexity. This research still can be developed to test the influence of another parameter in Random Forest classification such as random state or max features to get better accuracy. It also can be developed to compare other ensemble classifications, such as Adaboost, to get a better and efficient algorithm.

References

[1] Mukhsinin, D. A., Rafliansyah, M., Ibrahim, S. A., Rahmaddeni, R., & Wulandari, D. (2024). Implementasi Algoritma Decision Tree untuk Rekomendasi Film dan Klasifikasi Rating pada Platform Netflix. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4(2), 570–579. https://doi.org/10.57152/malcom.v4i2.1255

[2] Maulidah, M., Gata, W., Aulianita, R., Agustyaningrum, C. I., Studi, P., Komputer, I., & Mandiri, N. (2020). Algoritma Klasifikasi Decision Tree Untuk Rekomendasi Buku Berdasarkan Kategori Buku. 13(2), 89–96. https://doi.org/10.51903/e-bisnis.v13i2.251

[3] Setiawan, D., Alfiyani, L., Sulistio, J., & Qurtubi, Q. (2024). Utilizing Data Mining Techniques to Analysis Changes in Purchase Behavior of Batik's Customers. Advance Sustainable Science, Engineering and Technology, 6(2), 02402015. https://doi.org/10.26877/asset.v6i2.18506

[4] Amini, N., Saragih, T. H., Faisal, M. R., Farmadi, A., Abadi, F., Komputer, I., Dan Ilmu, M., Alam, P., Lambung, U., Jalan, M., Ahmad, J., Km, Y., & Selatan, K. (n.d.). JIP (Jurnal Informatika Polinema) Implementasi Algoritma Genetika Untuk Seleksi Fitur Pada Klasifikasi Genre Musik Menggunakan Metode Random Forest. https://doi.org/10.33795/jip.v9i1.1028

[5] Fanani, N. M. A. (2024). Sistem Rekomendasi Film Menggunakan Metode K-NN. Jurnal Ilmiah Sistem Informasi Dan Ilmu Komputer, 4(1), 178–185. https://doi.org/10.55606/juisik.v4i1.760

[6] Alam, L. (2024). Implementation of the Adaboost Method to Increase the Accuracy of Early Diabetes Predictions to Prevent Death Decision Tree-Based. Advance Sustainable Science, Engineering and Technology, 6(2), 0240207. https://doi.org/10.26877/asset.v6i2.18342

[7] Dwiyani, L. K. D., Suarjaya, I. M. A. D., & Rusjayanthi, N. K. D. (2023). Classification of Explicit Songs Based on Lyrics Using Random Forest Algorithm. Journal of Information Systems and Informatics, 5(2), 550–567. https://doi.org/10.51519/journalisi.v5i2.491

[8] Çetin, V., & Yıldız, O. (2022). A comprehensive review on data preprocessing techniques in data analysis. Pamukkale University Journal of Engineering Sciences, 28(2), 299–312.

[9] Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. In Frontiers in Energy Research (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fenrg.2021.652801

[10] Lan, T., Hu, H., Jiang, C., Yang, G., & Zhao, Z. (2020). A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification. Advances in Space Research, 65(8), 2052–2061.

[11] Wang, H. (n.d.). Research on the Application of Random Forest-based Feature Selection Algorithm in Data Mining Experiments. In IJACSA) International Journal of Advanced Computer Science and Applications (Vol. 14, Issue 10).

[12] Navisa, S., Hakim, L., Nabilah, A., Informasi, S., Sains, F., Teknologi, D., Sunan, U., Uin, A., & Ampel, S. (2021). Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM. In Jurnal Sistem Cerdas.

[13] Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, 220.

[14] Chen, D. (2024). Walmart sales prediction based on random forest model and application of feature importance. Applied and Computational Engineering, 53(1), 264–273. https://doi.org/10.54254/2755-2721/53/20241461

[15] Pehlivan S, İşler Y. Detection of heart disease risk utilizing correlation matrix, random forest and

permutation feature importance approaches. Akıllı Sistemler ve Uygulamaları Dergisi (Journal of Intelligent Systems with Applications) 2020; 3(1): 29-34.Sandag, G. A. (2020).

[16] Ratnasingam, S., & Muñoz-Lopez, J. (2023). Distance Correlation-Based Feature Selection in Random Forest. Entropy, 25(9), 1250. https://doi.org/10.3390/e25091250

[17] Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. In Journal of King Saud University - Computer and Information Sciences (Vol. 34, Issue 4, pp. 1060–1073). King Saud bin Abdulaziz University.

[18] Kamila, I. P., Sari, C. A., Rachmawanto, E. H., & Cahyo, N. R. D. (2023). A Good Evaluation Based on Confusion Matrix for Lung Diseases Classification using Convolutional Neural Networks. Advance Sustainable Science, Engineering and Technology, 6(1), 0240102. https://doi.org/10.26877/asset.v6i1.17330

[19] Disha, R.A., Waheed, S. Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. Cybersecurity 5, 1 (2022). https://doi.org/10.1186/s42400-021-00103-8

[20] Sage, A. J., Genschel, U., & Nettleton, D. (2020). Tree aggregation for random forest class probability estimation. Statistical Analysis and Data Mining, 13(2), 134–150. https://doi.org/10.1002/sam.11446

[21] Netflix TV Shows and Movies. (2022, July 26). Kaggle. https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies

[22] Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. SN Applied Sciences, 3(2). https://doi.org/10.1007/s42452-021-04148-9