Comparison of Feature Selection with Information Gain Method in Decision Tree, Regression Logistic and Random Forest Algorithms

by - -

Submission date: 13-Dec-2024 03:20AM (UTC+0000)
Submission ID: 238328618
File name: 2._usulan_JABT.docx (208.93K)
Word count: 4000
Character count: 22861

Journal of Applied Business and Technology

Comparison of Feature Selection with Information Gain Method in Decision Tree, Regression Logistic and Random Forest Algorithms

Muhammad Sholeh[a], Uning Lestari[a], Dina Andayati[b]

[a]Informatics Study Program, Faculty of Science and Information Technology, Universitas AKPRIND Indonesia
[b]Retail Management study program, Faculty of Communication and Business, Universitas AKPRIND Indonesia

Article History
Received
7 December 2023
Received in revised form
10 September 2024
Accepted
12 September 2024
Published Online
30 September 2024

*Corresponding author
muhash@akprind.ac.id

Abstract

One of the approaches that can be done is to perform feature selection. Feature selection is done by identifying the most informative features and not using features that do not directly contribute to the target feature. The purpose of feature selection is to increase the accuracy of the model. The research was conducted by comparing the performance of the model by comparing the accuracy results of the model without any feature selection with the model that has done feature selection. The process is done by comparing the accuracy results with decision tree, random forest and SVM algorithms. In the research method of feature selection on science data, the steps include understanding the domain and dataset, exploratory analysis, data cleaning, measuring feature relevance with criteria such as Information Gain, and feature ranking. The results are evaluated and validated using model performance metrics before and after feature selection. The process ensures selection of relevant features, improving accuracy. The research process used the Lung Cancer Prediction datasheet which consists of 306 rows and 16 attributes. The results show that feature selection can improve the performance of the classification model by reducing features that do not contribute to the target. Comparison results using decision tree, Regression Logistic and random forest classification model algorithms and feature selection resulted in a high accuracy value of 0.968 in the Regression Logistic algorithm with a feature selection of 5.

Keywords: Feature Selection, Classification Model, science data
DOI: https://doi.org/10.35145/jabt.v5i3.155
SDGs: Quality Education (4); Industry, Innovation and Infrastructure (9); Peace, Justice and Strong Institutions (16)

1.0 INTRODUCTION

Feature selection in the process of modelling science data can affect the accuracy of the results. The feature selection process supports in improving model accuracy or performance by focusing attention on the most relevant and informative features. By selecting the most influential features, the model can provide more accurate and effective predictions. In addition, feature selection brings computational efficiency by reducing the number of features in the dataset, optimizing model training time, and improving data analysis efficiency, especially on large datasets.

Improving accuracy can be done by feature selection and can provide solutions to data analysis by reducing the number of features without losing important information in a datasheet. Feature selection also allows the evaluation of the attributes or variables that are most important in making decisions, providing valuable insights into the factors that most influence the analysis results. Feature selection results in a subset of features that are considered the most relevant and informative for use in model development. The feature selection process involves various algorithms and methods, each designed to identify the most relevant and informative features of a dataset [1], [2], [3].

2.0 LITERATURE REVIEW

Feature selection is an important stage in data processing and machine learning model development. In many cases, a dataset may contain many features, and not all of them may contribute significantly to the understanding or prediction of the target variable. Therefore, feature selection techniques are crucial to retain the most informative features, optimize model performance, and reduce complexity [4], [5], [6], [7].

Research themed on the use of feature selection has been conducted with datasheets according to the research objectives and by using various methods and models. Feature selection with information gain is done [8], [9], [10] using Relief-F feature selection [11], using the classification method with the C4.5 algorithm based on forward selection [12], using feature selection correlation-based [13], using the Decision tree algorithm done [14], using the K-Nearest Neighbor model [15].

In the research conducted [16], the modelling process compares raw data without pre-processing, and data that has been pre-processed using feature selection based on correlation and data that has been pre-processed using information gain. The datasheet used is data regarding student learning activities in the E-learning management system. The data testing process is carried out using 10 folds cross validation using the C4.5 method, and data evaluation is carried out using confusion matrix. The test results show that the use of the C4.5 algorithm combined with feature selection based on correlation produces the highest accuracy, which is 76.92%. Meanwhile, the test results on raw data without feature selection and data selected using Information Gain have the same accuracy, which is 76.19%.

Research [17], used data mining feature selection techniques to evaluate the impact of courses on student study duration. The feature selection techniques used include Correlation Based, Information Gain Based, and Learner Based. The accuracy of each feature selection method was measured using the Naive Bayes classification algorithm. Experimental results show that the application of feature selection techniques can improve the classification accuracy of the Naive Bayes algorithm. Experiments on a dataset of student grades show that the Learner Based technique with the Wrapper model produces the highest accuracy, while the Information Gain technique provides the lowest accuracy.

[18], This research aims to identify indicators or attributes that have influence by using the correlation-based feature selection method. Furthermore, this study evaluates the performance of the Random Forest Classifier algorithm to forecast the academic performance students in online learning based on the Learning Management System (LMS) Open Learning. The data for this study was obtained from the academic administration and LMS Open Learning with a total of 2,663 data. Research [18], aims to optimize the use of the Naive Bayes algorithm by applying the Univariate Selection method to the UNSW-NB 15 data set. Only 40 features with the best relevance were selected for analysis. Furthermore, the data set was divided into two parts, namely test data and training data, with variations in the ratio of 10%:90%, 20%:80%, 30%:70%, 40%:60%, and 50%:50%. From the experimental results, it can be concluded that feature selection has a significant impact on the accuracy value obtained. The highest accuracy is achieved when the data set is divided into 40%:60%, either with or without feature selection. However, Naive Bayes with unselected features achieved the highest accuracy value of 91.43%, while with feature selection, the accuracy value increased to 91.62%. Therefore, it can be revealed that the use of feature selection method can increase the accuracy value of Naive Bayes by 0.19%.

One of the efforts to improve accuracy is by using feature selection and one of the methods used in feature selection is Information Gain. Information Gain is done by measuring how much uncertainty or entropy of the target variable can be removed by knowing the value of a feature. Information Gain is calculated by comparing the entropy of the target variable before and after feature selection. The entropy calculation process is done to measure the level of uncertainty in the dataset. Information Gain is calculated as the difference between the initial entropy and the entropy after feature selection, with the aim of maximizing uncertainty reduction [19], [20].

Based on the introduction, the research process includes comparing various feature selection methods and selecting classification algorithms to obtain high accuracy values. The feature selection method used is information gain with lung cancer datasheet which is a public datasheet. The classification algorithms used are decision tree, random forest and KNN.

3.0 METHODOLOGY

Data Analysis Technique

In conducting this research, data analysis was carried out using the Knowledge Discovery in Databases (KDD) method. KDD is a set of activities that includes collecting and using historical data to identify regularities, patterns, or interactions in large datasets [21]. The results of the data mining process are used to support decision making.

Figure 1 illustrates the KDD process. This process includes several steps, such as data collection, data cleaning, data integration, data selection, data transformation, data mining, data mining evaluation, knowledge presentation, and knowledge utilization. Each of these steps is important in generating valuable information from big data to support decision-making and understand patterns or regularities in the data.

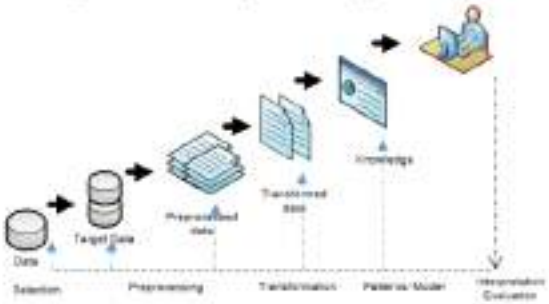

Figure 1. Stages in Knowledge Discovery method

Datasheet

The datasheet processed is a datasheet obtained from https://archive.ics.uci.edu. The datasheet used is a datasheet that contains data related to lung cancer and is stored in a CSV file (lung cancer.csv). The lung cancer datasheet can be downloaded at (https://archive.ics.uci.edu/dataset/62/lung+cancer). The datasheet consists of 309 data and consists of 16 columns.

Information Gain Algorithm

The Information Gain algorithm is used in the context of developing a decision tree for feature selection. The steps of the Information Gain algorithm:

1. Calculate Source Entropy (Entropy(S)):
   Measures the uncertainty or vagueness in the datasheet [22].
   The formula is:

$$Entropy(S) = -\sum_{i=1}^{n} p_i . \log_2(p_i) \dots\dots\dots\dots\dots\dots (1)$$

   where n is the number of classes, and pi is the proportion of samples belonging to class i.
2. Calculate Entropy After Feature Selection (Entropy_S_F):
   Calculate the entropy after feature selection based on a particular feature value. This is done by calculating the entropy for each unique value of the feature and then taking the average of the values. The average value is each value weighted according to its proportion of frequency in a data set based on the frequency of each feature value [22].
   The formula is:

$$Entropi\,SF = \sum_{j=1}^{m} \left(\frac{N_j}{N} . Entropy\,(S_j)\right) \dots\dots\dots\dots\dots\dots (2)$$

   where m is the number of unique values of the feature, Nj is the number of samples having feature value j, N is the total number of samples, and Entropy (Sj) is the entropy of the subclass having feature value j.
3. Calculate the Information Gain
   Information Gain is the difference between the entropy before and after feature selection [22]. The formula is:

$$Information\ Gain = Entropi(S) - Entropy_{SF} \dots\dots\dots\dots\dots (3)$$

4. Select the Feature with the Highest Information Gain:
   Repeat steps 2 and 3 for each available feature and select the feature that has the highest Information Gain. This feature will be the best feature to split the dataset in the next steps of decision tree construction.

Model Evaluation

Model evaluation is a critical step in understanding the extent to which the model that has been built is able to make good predictions. A common evaluation method involves comparing the model's predictions with the actual values in the test dataset.

In classification model evaluation, accuracy is one of the commonly used metrics to measure the extent to which the model can correctly predict the class or label. Accuracy is calculated as the ratio between the number of correct predictions (True Positive and True Negative) and the total number of observations [19]. The accuracy formula can be formulated as follows:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Samples} \quad \ldots\ldots (4)$$

Precision is one of the evaluation metrics in classification models that measures how many of the positive instances predicted by the model are actually positive [19]..

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \quad \ldots\ldots (5)$$

Recall, also known as Sensitivity or True Positive Rate, is an evaluation metric in classification models that measures how many of the overall positive instances are successfully predicted by the model [19]., Recall can be calculated with the following formula:

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negatives} \quad \ldots\ldots (6)$$

F1-Score is an evaluation metric in classification models that includes a balance between precision and recall. It provides a single value that combines both metrics and is useful when there is a need to comprehensively assess model performance [19].. F1-Score is calculated by the following formula:

$$F1-Score = 2\,x\frac{Precision\,c\,Recall}{Precision+Recall} \quad \ldots\ldots (7)$$

Confusion matrix is a classification model performance evaluation tool that presents a summary of the model's predicted results against the actual known test data. This matrix consists of four main elements: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Confusion matrix) is a classification model performance evaluation tool that presents a summary of the model's prediction results against actual known test data [19].. This matrix consists of four main elements:

1. True Positive (TP):
   Represents the number of positive observations correctly predicted by the model.
2. True Negative (TN):
   Represents the number of negative observations correctly predicted by the model.
3. False Positive (FP):
   Represents the number of negative observations that were incorrectly predicted as positive by the model (False Alarm or Type I Error).
4. False Negative (FN):
   Represents the number of positive observations that were incorrectly predicted as negative by the model (Miss or Type II Error).

Using these symbols, the Confusion matrix is presented in Figure 2.

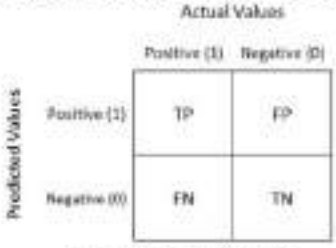

Figure 2. Confusion matrix

4.0 RESULTS AND DISCUSSION

Data Cleaning and Integration

The data cleaning process is done by checking the datasheet used. This process is done by checking empty data, double data, outlier data and consistency of data content [23], [24].

- Blank data checking
- Checking empty data is done by looking at the missing value column on the datasheet.
- Removing duplicate data
- Duplicate data can cause errors in understanding the distribution and characteristics of the data. Statistical analyses performed on unclean datasets may result in inaccurate estimates.
- Checking for outlier data

- Removal of outliers from a dataset is an important consideration in statistical analysis and data modelling. Outliers, or extreme points that differ significantly from the majority of the data, can affect various aspects of statistical analysis
- Data consistency checking
- Data consistency checking is the process of verification and evaluation to ensure that the data in a dataset or database meets predefined standards and rules. The process of checking categorical data is shown in Figure 3.

```
for column in df.columns:
    unique_values = df[column].unique()
    print(f"Unique Value In Value Column   '{column}':")
    print(unique_values)
    print("\n")
```

Figure 3. Data consistency checking process

The result of checking, each feature already has a consistent value. The GENDER feature only contains M and F data and other features are in accordance with consistent data content. The checking results are in Figure 4

```
Unique Value In Value Column   'GENDER':
['M' 'F']

Unique Value In Value Column   'AGE':
[69 74 59 63 75 62 51 60 55 81 71 66 58 48 57 44 64 23 81 55 62 56 67 77
 70 54 46 73 47 71 66 76 78 81 79 36 39 87 46]

Unique Value In Value Column   'SMOKING':
[1 2]

Unique Value In Value Column   'YELLOW_FINGERS':
[2 1]

Unique Value In Value Column   'ANXIETY':
[2 1]

Unique Value In Value Column   'PEER_PRESSURE':
[1 2]
```

Figure 4. Data consistency checking results

Data Selection and Transformation

The data selection and transformation process is an important stage in the data analysis cycle that aims to improve the quality and relevance of the information contained in the dataset.

The ultimate goal of data selection and transformation is to form datasets that are better prepared for analysis or modelling, improve the interpretation of results, and support more effective decision-making. By optimizing data representation, reducing noise, and improving computation efficiency, this process becomes a critical step in bringing added value to the information contained in the dataset. Feature selection is an effective way to reduce dimensionality by removing redundant and irrelevant data [25], [26].

Data Mining

In the context of feature selection, Information Gain is used to assess how well a feature can separate a dataset into different classes. The main principle is that features that provide more information about the target variable have higher Information Gain. Features that provide the highest Information Gain or the most in the separation of the dataset. Features that have high Information Gain are considered more informative and have a significant impact on the target variable. Figure 6, feature selection results on the datasheet.

Figure 5. Feature selection results

Based on Figure 5, 5 features that have a high gain ratio were determined. The selected features are allergy, alcohol consuming, wheezing, coughing, swallowing difficulty.

Model Evaluation

Model evaluation is used to measure the accuracy obtained from the model created. Model evaluation is done by comparing the results of model building with decision tree, Regression Logistic and Random Forest algorithms. Modelling is done by comparing the use of all existing features and the results of feature selection. After feature selection, the results tend to be more efficient and there is an increase in accuracy value. A smaller number of features can increase the speed of model training, reduce complexity, and make interpretation easier. Although the results may vary depending on the feature selection method used, in general, a good feature selection can improve the accuracy of the model by maintaining or even improving the prediction performance against the target variable [16], [15], [27]. The accuracy results of the 6 experiments are presented in table 1.

Table 1. Comparison of Accuracy Results

Evaluation	Before feature selection			After feature selection		
	DT	LR	RF	DT	LR	RF
Accuracy	0.947	0.948	0.947	0.958	0.968	0.958
Recall	0.947	0.958	0.947	0.958	0.968	0.958
Precision	0.949	0.959	0.949	0.958	0.968	0.958
F1 Score	0.947	0.9658	0.947	0.958	0.96 8	0.958
Confusion Matrix	[[45 1]	[[45 1]	[[45 1]	[[50 2]	[[51 1]	[[50 2]
	[4 46]]	[3 47]]	[4 46]]	[2 42]]	[2 42]]	[2 42]

The results of 6 model building experiments show that the classification model using the Regression Logistic algorithm with feature selection has a better accuracy rate. The results of the 6 trials of the accuracy value are not too far apart [28-32].

5.0 CONCLUSION

The feature selection process is an important step in the data analysis process that aims to improve model performance by selecting the most relevant subset of features. The feature selection process used is information gain. The advantages of using Information Gain include its ability to identify the most informative features, separate the dataset into more homogeneous groups, and increase the efficiency of the model by reducing the dimension of the dataset. Feature selection with Information Gain also helps overcome overfitting, maintain model interpretability, and improve understanding of influential factors [33-37].

The results showed that the accuracy value after feature selection resulted in higher accuracy. The number of features selected was 5, namely allergy, alcohol consuming, wheezing, coughing, swallowing difficulty. Accuracy with Regression Logistic algorithm produces a value before feature selection of 0.948 and after feature selection produces an accuracy value of 0.968.

References

[1] D. Dielen, A. D. B. Meysman, and M. Ali, *Introducing Data Science*. 2016.

[2] M. Arhami and M. Nasir, *Data Mining - Algoritma dan Implementasi*. Yogyakarta: Penerbit Andi, 2020.

[3] D. Jollyta, W. Ramdhan, and M. Zarlis, *Konsep Data Mining Dan Penerapan*. Yogyakarta: Deepublish Publisher, 2020.

[4] P. Mathur, *Machine Learning Applications Using Python*. 2019.

[5] M. Barlow, *Learning to Love Data Science*. Gravenstein Highway North, Sebastopol: O'Reilly Media, Inc, 2015.

[6] D. Sarkar, R. Bali, and T. Sharma, *Practical Machine Learning with Python*. Bangalore, Karnataka, India: Apress, 2018.

[7] A. Naif Alharbi and M. Dahab, "Comparative Study on Fast Feature Selection," *International Journal of Information Technology and Language Studies (IJITLS)*, vol. 2, no. 2, pp. 55–64, 2018.

[8] S. H. A. Aini, Y. A. Sari, and A. Arwan, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 9, pp. 2546–2554, 2018.

[9] M. Rijal *et al.*, "Perbandingan Kinerja Metode Seleksi Fitur untuk Mendeteksi Aktivitas Trojan Performance Comparison of Feature Selection Methods for Detecting Trojan Activity," *Jurnal_Pekommas_Vol._7_No*, vol. 2, no. april 2020, pp. 85–97, 2022.

[10] K. Kurniabudi, A. Harris, and A. Rahim, "Seleksi Fitur Dengan Information Gain Untuk Meningkatkan Deteksi Serangan DDoS menggunakan Random Forest," *Techno.Com*, vol. 19, no. 1, pp. 56–66, 2020. doi: 10.33633/tc.v19i1.2860.

[11] R. N. Yusra, O. S. Sitompul, and Sawaluddin, "Kombinasi K-Nearest Neighbor (KNN) dan Relief-F Untuk Meningkatkan Akurasi Pada Klasifikasi Data," *InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan*, vol. 1, pp. 0–5, 2021.

[12] E. Nurlia and U. Enri, "Penerapan Fitur Seleksi Forward Selection Untuk Menentukan Kematian Akibat Gagal Jantung Menggunakan Algoritma C4.5," *Jurnal Teknik Informatika Musirawas) Elin Nurlia*, vol. 6, no. 1, p. 42, 2021.

[13] A. N. Puteri, A. Arizal, and A. D. Achmad, "Feature Selection Correlation-Based pada Prediksi Nasabah Bank Telemarketing untuk Deposito," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 2, pp. 335–342, 2021, doi: 10.30812/matrik.v20i2.1183.

[14] P. Riswanto, R. A. Aziz, and S. -, "Penerapan Decision Tree C4.5 Sebagai Seleksi Fitur Dan Support Vector Machine (Svm) Untuk Diagnosa Kanker Payudara," *Jurnal Informatika*, vol. 19, no. 1, pp. 54–61, 2019, doi: 10.30873/ji.v19i1.1442.

[15] A. Bode, "Seleksi Fitur Untuk Prediksi Rating Film Hollywood Menggunakan Model K-Nearest Neighbor," *JUPITER: Jurnal Penerapan Ilmu-Ilmu Komputer*, vol. 5, no. 1, 2019.

[16] A. S. B. Asmoro, W. S. G. Irianto, and U. Pujianto, "Perbandingan Kinerja Hasil Seleksi Fitur pada Prediksi Kinerja Akademik Siswa Berbasis Pohon Keputusan.," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 4, no. 2, p. 84, 2018.

[17] I. M. B. Adnyana, "Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa," *Jurnal Sistem dan Informatika*, vol. 13, no. 2, pp. 72–76, 2019.

[18] Harianto, A. Sunyoto, and S. Sudarmawan, "Optimasi Algoritma Naïve Bayes Classifier untuk Mendeteksi Anomaly dengan Univariate Fitur Selection," *Edumatic: Jurnal Pendidikan Informatika*, vol. 4, no. 2, pp. 40–49, 2020, doi: 10.29408/edumatic.v4i2.2433.

[19] M. Swamynathan, *Mastering Machine Learning with Python in Six Steps*. Bangalore, Karnataka, India: apress, 2017.

[20] S. Ozdemir, *Principles of Data Science*. Birmingham: Packt Publishing Ltd, 2017.

[21] S. Suraya, M. Sholeh, and D. Andayati, "Penerapan Metode Clustering Dengan Algoritma K-Means Pada Pengelompokan Indeks Prestasi Akademik Mahasiswa," *Skanika*, vol. 6, no. 1, pp. 51–60, 2023, doi: 10.36080/skanika.v6i1.2982.

[22] G. Bonaccorso, *Machine Learning Algorithm*. Birmingham: Packt Publishing Ltd, 2017.

[23] A. Fadli and M. I. Rosadi, "Klasifikasi Penyakit Jantung Koroner Menggunakan Seleksi Fitur dan Support Vector Machine," *Jurnal Explore IT*, vol. 10, no. 2, pp. 32–41, 2018.

[24] K. N. F. S. Dewi Fatmarani Surianto, "SELEKSI FITUR INFORMATION GAIN (IG) PADA KLASIFIKASI DATA OPINI SAHAM MENGGUNAKAN METODE NAÏVE BAYES," *Jurnal INSTEK (Informatika Sains dan Teknologi)*, vol. 8, no. 1, pp. 36–45, 2023.

[25] C. Kuzudisli, B. Bakir-Gungor, N. Bulut, B. Qaqish, and M. Yousef, "Review of feature selection approaches based on grouping of features," *PeerJ*, vol. 11, 2023, doi: 10.7717/peerj.15666.

[26] Y. Bouchlaghem, Y. Akhiat, and S. Amjad, "Feature Selection: A Review and Comparative Study," *E3S Web of Conferences*, vol. 351, pp. 1–6, 2022, doi: 10.1051/e3sconf/202235101046.
[27] A. Hermawan and A. P. Wibowo, "Implementasi Korelasi untuk Seleksi Fitur pada Klasifikasi Jamur Beracun Menggunakan Jaringan Syaraf Tiruan," *Jurnal INTEK*, vol. 5, no. 1, pp. 63–67, 2022.
[28] J. Angelyn and R. N. Putri, "Diagnosis System Design of Depression and Anxiety with NAÏVE BAYES Method," *J. Appl. Bus. Technol.*, vol. 2, no. 2, pp. 92–97, 2021.
[29] E. M. Nazara and D. Nasien, "Employee Attendance System Using Rapid Application Development Method Based on Location Based Service," *J. Appl. Bus. Technol.*, vol. 5, no. 2, pp. 96–104, 2024, doi: https://doi.org/10.35145/jabt.v5i2.148.
[30] C. Effendy and G. Gusrianty, "Application of Round Robin in Scheduling in Web-Based Wedding Organizers," *J. Appl. Bus. Technol.*, vol. 5, no. 2, pp. 90–95, 2024, doi: https://doi.org/10.35145/jabt.v5i2.150.
[31] E. Susanto, G. Gustientiedina, and M. Siddik, "Application of the Forward Chaining Method in Diagnosing Tomato Fever," *J. Appl. Bus. Technol.*, vol. 5, no. 1, pp. 41–50, 2024, doi: https://doi.org/10.35145/jabt.v5i1.143 1.0.
[32] S. R. Silva et al., "Extensive Sheep and Goat Production: The Role of Novel Technologies towards Sustainability and Animal Welfare," *Animals*, vol. 12, no. 885, pp. 1–28, 2022, doi: 10.3390/ani12070885.
[33] J. Chen and G. Gustientiedina, "Implementation of Fuzzy Expert System to Detect Parkinson's Disease Based on Mobile," *J. Appl. Bus. Technol.*, vol. 5, no. 2, pp. 72–81, 2024, doi: 10.35145/jabt.v5i2.145.
[34] Sudarno, N. Y. Putri, N. Renaldo, M. B. Hutahuruk, and Cecilia, "Leveraging Information Technology for Enhanced Information Quality and Managerial Performance," *J. Appl. Bus. Technol.*, vol. 3, no. 1, pp. 102–114, 2022, doi: https://doi.org/10.35145/jabt.v3i1.97.
[35] N. Renaldo, Sudarno, M. B. Hutahuruk, A. T. Junaedi, Andi, and Suhardjo, "The Effect of Entrepreneurship Characteristics, Business Capital, and Technological Sophistication on MSME Performance," *J. Appl. Bus. Technol.*, vol. 2, no. 2, pp. 109–117, 2021, doi: https://doi.org/10.35145/jabt.v2i2.74.
[36] N. Renaldo, Suhardjo, Suharti, Suyono, and Cecilia, "Benefits and Challenges of Technology and Information Systems on Performance," *J. Appl. Bus. Technol.*, vol. 3, no. 3, pp. 302–305, 2022, doi: https://doi.org/10.35145/jabt.v3i3.114.
[37] M. Irman, E. A. Suhendra, and H. Diana, "Work Experience, Professionalism, Independence and the Application of Information Technology on Auditor Performance in Order to Increasing Audit Quality at the Financial Audit Agency of the Republic Indonesia Representative of the Riau Province," *J. Appl. Bus. Technol.*, vol. 2, no. 3, pp. 206–222, 2021, doi: https://doi.org/10.35145/jabt.v2i3.78.

Comparison of Feature Selection with Information Gain Method in Decision Tree, Regression Logistic and Random Forest Algorithms

ORIGINALITY REPORT

21%	17%	14%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	mdpi-res.com Internet Source	2%
2	e-journal.hamzanwadi.ac.id Internet Source	2%
3	Submitted to United International University Student Paper	2%
4	www.geeksforgeeks.org Internet Source	1%
5	Charu C. Aggarwal. "Data Classification - Algorithms and Applications", Chapman and Hall/CRC, 2019 Publication	1%
6	www.e-jabt.org Internet Source	1%
7	e-jabt.org Internet Source	1%
8	www.nature.com Internet Source	

		1 %
9	eprints.intimal.edu.my Internet Source	1 %
10	journal.universitasbumigora.ac.id Internet Source	1 %
11	Ivan Rifky Hendrawan, Ema Utami, Anggit Dwi Hartanto. "Comparison of Naïve Bayes Algorithm and XGBoost on Local Product Review Text Classification", Edumatic: Jurnal Pendidikan Informatika, 2022 Publication	1 %
12	Mespin Andayani, Fitri Marisa, Rangga Pahlevi Putra. "Sentiment Analysis of Indonesia 2024 Election with a Comparison of Naive Bayes and KNN Algorithms on Twitter", SAR Journal - Science and Research, 2024 Publication	1 %
13	Submitted to Sheffield Hallam University Student Paper	1 %
14	Submitted to University of Warwick Student Paper	1 %
15	inass.org Internet Source	1 %
16	Sigurdur Ólafsson, Jaekyung Yang. "Intelligent Partitioning for Feature Selection", INFORMS	<1 %

	Journal on Computing, 2005 Publication	
17	www.slideshare.net Internet Source	<1 %
18	peerj.com Internet Source	<1 %
19	www.ui.ac.id Internet Source	<1 %
20	Pengcheng Li, Baotian Dong, Sixian Li. "A Study of Adjacent Intersection Correlation Based on Temporal Graph Attention Network", Entropy, 2024 Publication	<1 %
21	jurnal.seaninstitute.or.id Internet Source	<1 %
22	Submitted to University of Technology, Sydney Student Paper	<1 %
23	data.fesb.unist.hr Internet Source	<1 %
24	jurnal.unai.edu Internet Source	<1 %
25	Fengjun Zhang, Lisheng Huang, Kai Shi, Shengjie Zhai, Yunhai Lan, Qinghua Li. "Intrusion detection based on hybrid	<1 %

	metaheuristic feature selection", The Computer Journal, 2024 Publication	
26	Sid Ahmed Mokeddem. "A fuzzy classification model for myocardial infarction risk assessment", Applied Intelligence, 2017 Publication	<1 %
27	arno.uvt.nl Internet Source	<1 %
28	theses.lib.polyu.edu.hk Internet Source	<1 %
29	Henny Pramoedyo, Danang Ariyanto, Novi Nur Aini. "COMPARISON OF RANDOM FOREST AND NAÏVE BAYES METHODS FOR CLASSIFYING AND FORECASTING SOIL TEXTURE IN THE AREA AROUND DAS KALIKONTO, EAST JAVA", BAREKENG: Jurnal Ilmu Matematika dan Terapan, 2022 Publication	<1 %
30	publikasi.dinus.ac.id Internet Source	<1 %
31	www.evidentlyai.com Internet Source	<1 %
32	www.mdpi.com Internet Source	<1 %

33	Putri ., Evy Sulistianingsih, Nurfitri Imro'ah, Naomi Nessyana Debataraja. "APPLICATION OF C4.5 ALGORITHM WITH FEATURE SELECTION IN CLASSIFICATION OF DISCHARGE STATUS OF HEAD INJURY PATIENTS", VARIANCE: Journal of Statistics and Its Applications, 2024 Publication	<1 %
34	jurnal.atmaluhur.ac.id Internet Source	<1 %
35	repository.wicida.ac.id Internet Source	<1 %
36	www.ijmra.in Internet Source	<1 %
37	www.researchsquare.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

Comparison of Feature Selection with Information Gain Method in Decision Tree, Regression Logistic and Random Forest Algorithms