Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Random Forest Analysis In Classifying Orange Quality Data

[1]Suci Ramadhani, [2]Muslimin B, [3]Ida Maratul Khamidah
[1,2,3]Politeknik Pertanian Negeri Samarinda, Indonesia

Article Info

Keywords:
Agriculture,
Classification,
Orange Quality,
Random Forest.

ABSTRACT

The quality of oranges is important to determine selling value. However, citrus quality assessments are often subjective and inconsistent, which can impact consumer satisfaction and market efficiency. In the agricultural industry, especially in citrus commodities, there are difficulties in classifying fruit quality accurately and efficiently, which has an impact on the assessment and determination of market prices. Given the importance of citrus quality in the agricultural and food industries, there is an urgent need for objective and efficient methods for classifying citrus quality. Inappropriate classification can cause economic losses for farmers and distributors, as well as reduce consumer satisfaction with product quality. As a solution, this research proposes the use of the Random Forest method to classify orange quality data. The method used in this research involved collecting orange quality data, including characteristics such as color, texture, and size. This data is then analyzed using the Random Forest algorithm. The Random Forest method is used to process orange quality data, by utilizing features such as color, size and skin texture. This model is trained using historical data to predict fruit quality. The research results show that the Random Forest method successfully classifies citrus quality data with high accuracy, demonstrating its potential as an effective tool for future citrus quality assessment by proving its effectiveness in supporting decisions in the agricultural sector.

Corresponding Author:
Suci Ramadhani
Politeknik Pertanian Negeri Samarinda, Indonesia
Suciramadhani@politanisamarinda.ac.id

INTRODUCTION

Previous studies have used the Naïve Bayes algorithm to classify citrus fruit freshness, but it was found that the process of sorting quality citrus fruit still takes a long time if done manually (Firmansyah, 2023). The Random Forest method has better prediction results compared to other statistical methods such as Logistic Regression and Naïve Bayes Classification, especially in terms of accuracy (Kalumbang, 2021). The application of Random Forest can increase effectiveness and efficiency in orange quality classification, which is very important for the agricultural industry to ensure product quality and consumer satisfaction (Pratama, 2019).

Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It's known for its high accuracy, robustness, and ease of use. Random Forest is known to have a high level of accuracy in classification and regression modeling, as well as being tolerant of outliers and unbalanced data. This method is also efficient in dealing with overfitting and increasing independence between decision trees, which contributes to more accurate predictions (Arnita, 2019).

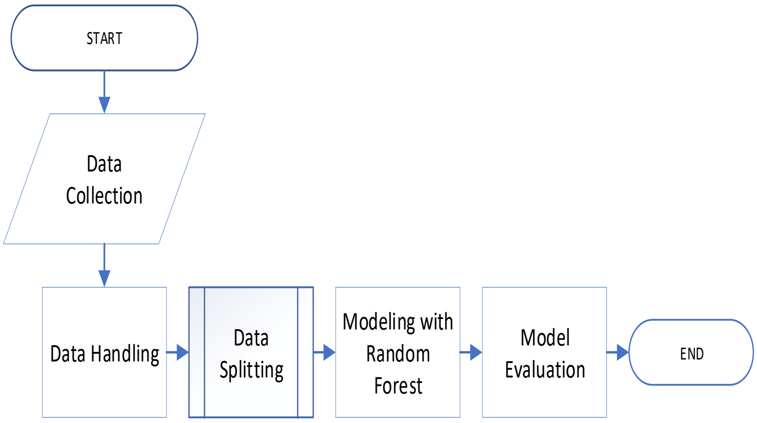
According to research (Amiri, 2019) the combination of models in trees that use random vectors that occur separately in the input vector, and each tree produces a class that is popular in classifying the input vector is the definition of Random Forest. This model has a random feature in combining each node to build a decision tree. Oranges are one of the horticulture products that have high economic value and market demand. Oranges have taste, aroma, freshness and are a source of vitamins for the body, so they are very popular and have become the family's favorite fruit. However, the quality of local oranges is still far behind imported oranges in terms of taste, color and size. Therefore, research on the quality of oranges is important.

The availability of imported oranges in Indonesia has led to an increase in imports of citrus fruit. Local oranges must have advantages in order to win competition in the national market. However, local orange production tends to decline, and the quality has not shown its superiority compared to imported oranges. When classifying orange quality, we deal with various features such as size, color, texture, and sugar content. Random Forest is adept at handling such multi-dimensional data and can easily capture the complex patterns that distinguish different quality levels of oranges. It starts with decision trees, which are models that make decisions based on the data's features. Randomly selects subsets of the dataset for each tree, ensuring diversity among the trees. Aggregates the predictions of individual trees to produce a more accurate and stable prediction.

In research using Random Forest, the tools that are often used are programming software such as Python with the Scikit-Learn library, which provides an implementation of the Random Forest algorithm. Extraction features such as LBP (Local Binary Pattern) and HSV (Hue, Saturation, Value) are also used to identify fruit quality characteristics. The data used for analysis comes from a public dataset which includes images of fresh and rotten fruit, which are then processed for feature extraction and classification. Other datasets include parameters such as fruit diameter, fruit weight, and skin texture measured from citrus fruit samples.

METHODS

Data preprocessing is carried out in 3 processes, namely the first is a missing value which replaces the numerical value with the mean value of similar attributes and columns. (Dou, 2019). Then the second is the process of deleting duplicate data. The score has a standard value, x is the dataset used in the analysis process for improving classification performance, μ is the average value (mean) then σ is the standard deviation for each data variable. So, the process of the Z-score is μ is 0 and σ is the number 1. In this process, it is actually not a problem if a small amount of the entire data is missing, but the percentage of missing data is small, such as only 1% of the entire dataset. If the missing data reaches a large amount, it is necessary to retest the data to see whether the data is suitable for further processing or not. The following is a Random Forest flowchart for classifying orange quality data:



Picture 1. Flowchart Research

a. Data Collection, Select and collect relevant citrus quality data
b. Data Handling, Cleaning and preparing data for analysis
c. Data Splitting, Building a classification model using the Random Forest algorithm
d. Modeling with Random Forest, Building a classification model using the Random Forest algorithm
e. Model Evaluation, Measure model performance with metrics such as accuracy, precision, and recall

Data collected from reliable sources, including physical measurements of oranges such as diameter, weight, and skin texture. Data set details may include images of oranges, numerical data from measurements, and expert-defined quality labels. Random Forest is an ensemble method consisting of many decision trees. Classification predictions are made based on the majority of votes from these trees.

Random Forest Method Stages

a. Bootstrap Aggregating, Uses a random sample of the data to build each tree
b. Random Feature Selection, In each split, only some features are considered
c. Tree Building, Build decision trees independently
d. Majority Voting, Combines predictions from all trees to make a final decision

The observation data will be reduced by the mean of each variable and divided by the standard deviation. (Liu, 2019).

a. Gather Data, Collect data on various features of oranges like size, color, weight, sugar content, etc.
b. Clean Data, Remove any outliers or erroneous entries that could skew the results.
c. Feature Selection, Choose the most relevant features that contribute to the quality of oranges.

Model Building

a. Split the Data, Divide your data into training and testing sets.
b. Initialize the Random Forest, Set up the Random Forest classifier with a specific number of trees.
c. Train the Model, Fit the Random Forest model to the training data.

Model Evaluation

a. Cross-Validation, Use cross-validation to assess the model's performance.
b. Performance Metrics, Evaluate the model using metrics like accuracy, precision, recall, and F1-score.
c. Feature Importance, Analyze which features are most important in determining orange quality.

Model Optimization

a. Hyperparameter Tuning, Adjust the model parameters like the number of trees, depth of trees, etc., to improve performance.
b. Pruning, Remove parts of the trees that do not provide power to the classification.

Deployment

a. Integrate the Model, Implement the model into a production environment where it can classify orange quality in real-time.
b. Monitor Performance, Continuously monitor the model's performance and make adjustments as necessary.

RESULTS AND DISCUSSION

Data has been analyzed so that have type column for identified. During data analysis, column final considered as target and column other will considered as input field. Shared dataset into training sets, validation sets and test sets. Data has been analyzed for have type column for identified. Moment analyze data, column final treated as target and column other will enforced as input field. Shared dataset into training data, validation sets and testing data. Following is data table used:

Table 1 Orange Dataset	Size (cm)	Weight (g)	Brix (Sweetness)	pH (Acidity)	Softness (1-5)	Harvest Time (days)	Ripeness (1-5)	Color	Variety	Blemishes (Y/N)	Quality (1-5)
0	7.5	180	12	3.2	2	10	4	Orange	Valencia	N	4
1	8.2	220	10.5	3.4	3	14	4.5	Deep Orange	Navel	N	4.5
2	6.8	150	14	3	1	7	5	Light Orange	Cara Cara	N	5
3	9	250	8.5	3.8	4	21	3.5	Orange-Red	Blood Orange	N	3.5
4	8.5	210	11.5	3.3	2.5	12	5	Orange	Hamlin	Y (Minor)	4.5

A Correlation Matrix is a statistical tool that displays the correlation coefficients between multiple variables. Each cell in the matrix represents the correlation between two variables. The values range from -1 to 1, where:

Tables 2. Correlation Matrix

	Size (cm)	Weight (g)	Brix (Sweetness)	pH (Acidity)
Size (cm)	1.000000	0.305348	-0.305865	0.330487
Weight (g)	0.305348	1.000000	-0.239659	0.303806
Brix (Sweetness)	-0.305865	-0.239659	1.000000	-0.300080
pH (Acidity)	0.330487	0.303806	-0.300080	1.000000

	Size (cm)	Weight (g)	Brix (Sweetness)	pH (Acidity)
Softness (1-5)	0.236863	0.316806	-0.344376	0.360222
HarvestTime (days)	0.313330	0.374670	-0.329862	0.262461
Ripeness (1-5)	-0.256665	-0.260454	0.144666	-0.232627
Quality (1-5)	-0.243113	-0.330268	0.631343	-0.321942

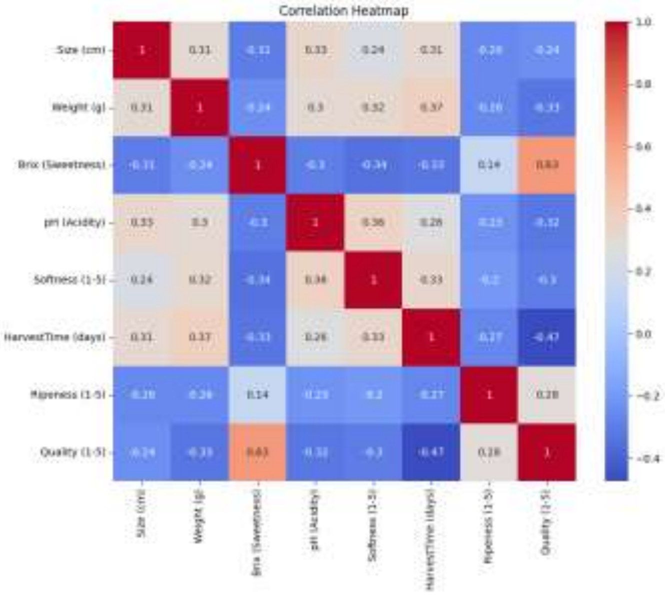
A correlation heatmap is a graphical representation of the correlation coefficients between different attributes (features) in a dataset. It provides insights into how strongly related each pair of attributes is. Here's how to interpret a correlation heatmap:

a. Color Representation:
Darker colors indicate stronger correlations, while lighter colors represent weaker correlations. Positive correlations (when one variable increases, the other tends to increase) are usually shown in warm colors like red or orange. Negative correlations (when one variable increases, the other tends to decrease) are typically displayed in cooler colors like blue.

b. Heatmap Cells:
Each cell in the heatmap corresponds to the correlation between two attributes. The value in the cell represents the correlation coefficient (usually Pearson or Spearman) between the pair of attributes.

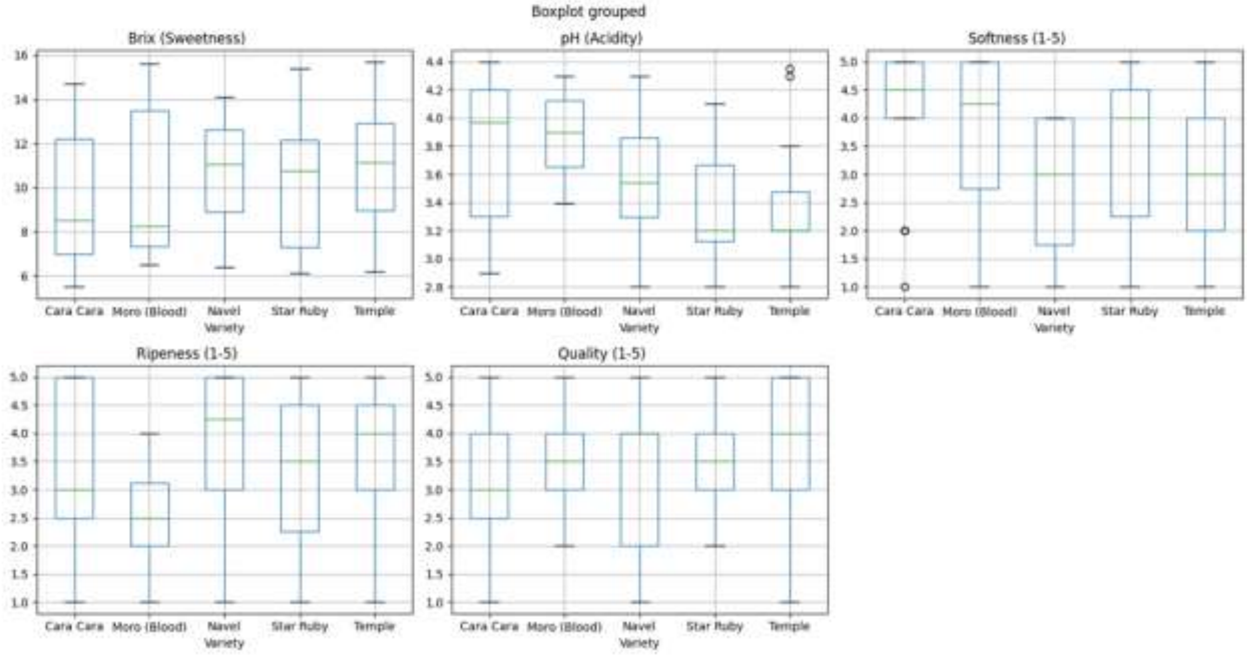
c. Interpretation: Look at the heatmap to identify patterns:
Strong positive correlations: Dark red cells.
Strong negative correlations: Dark blue cells.
No correlation: Lighter-colored cells near zero.

d. Usefulness:
Correlation heatmaps help you understand which attributes are related and can guide feature selection, model building, and data exploration.
When analyzing orange quality data, a correlation heatmap can reveal relationships between features such as size, color, texture, and sugar content.



Picture 1. Correlation Heatmap Orange Dataset

The correlation between Size and Weight is positive (0.305348), indicating that larger oranges tend to weigh more, which is expected. There is a relatively strong positive correlation (0.631343) between Brix (Sweetness) and Quality rating. This suggests that sweeter oranges are generally rated as having higher quality. There is a moderate negative correlation (-0.321942) between pH (Acidity) and Quality rating. This implies that oranges with lower acidity levels (higher pH) tend to have higher quality ratings. There is a moderate negative correlation (-0.474754) between HarvestTime (days since harvest) and Quality rating. This indicates that oranges harvested more recently tend to have higher quality ratings, which is expected as they are likely fresher. There is a positive correlation (0.280764) between Ripeness rating and Quality rating, suggesting that riper oranges tend to have higher quality ratings. There is a moderate negative correlation (-0.302732) between Softness rating and Quality rating. This could imply that softer oranges (higher Softness rating) are perceived as having lower quality. There is a moderate negative correlation (-0.300080) between Brix (Sweetness) and pH (Acidity), which means that sweeter oranges tend to have lower acidity levels. There is a weak negative correlation between Size and Brix (-0.305865), and a weak positive correlation between Size and pH (0.330487). This suggests that larger oranges tend to be slightly less sweet and have slightly higher acidity levels.



Pictures 2. Boxplot Grouped

Here are some potential insights:

Brix (Sweetness): The 'Temple' variety has the highest mean Brix (11.02), suggesting it is the sweetest among the top varieties. The 'Navel' variety also has a relatively high mean Brix of 10.71. The 'Cara Cara' variety has the lowest mean Brix (9.37) among the top 5 varieties, indicating it may be less sweet on average. pH (Acidity): The 'Moro (Blood)' variety has the highest mean pH (3.89), suggesting it may have slightly lower acidity levels compared to other varieties. The 'Star Ruby' variety has the lowest mean pH (3.34), indicating it may be more acidic on average. Softness (1-5): The 'Cara Cara' variety has the highest mean Softness rating (4.02), suggesting it may be perceived as softer compared to other varieties. The 'Navel' variety has the lowest mean Softness rating (2.63), indicating it may be perceived as firmer on average. Ripeness (1-5): The 'Cara Cara' variety has the highest mean Ripeness rating (3.36), suggesting it may be perceived as riper compared to other varieties. The 'Moro (Blood)' variety has the lowest mean Ripeness rating (2.59), indicating it may be perceived as less ripe on average. Quality (1-5): The 'Temple' and 'Star Ruby' varieties have the highest mean Quality ratings (3.64 and 3.61, respectively), suggesting they may be perceived as having the highest overall quality among the top varieties. The 'Cara Cara' variety has the lowest mean Quality rating (3.19), indicating it may be perceived as having lower overall quality compared to other top varieties.

The MSE value of 0.39 represents the average squared difference between the predicted and actual quality ratings. Since the quality ratings are on a scale of 1 to 5, an MSE of 0.39 is relatively small, indicating that the model's predictions are reasonably close to the actual values. Root Mean Squared Error (RMSE): The RMSE value of 0.63 is the square root of the MSE and provides an estimate of the average magnitude of the errors in the same units as the target variable (quality rating). An RMSE of 0.63 on a scale of 1 to 5 is not too large, suggesting that the model's predictions are reasonably accurate. R-squared (R^2): The R^2 value of 0.47 indicates that the model can explain approximately 47% of the variance in the quality ratings based on the provided features. An R^2 value of 1 would represent a perfect fit, while 0 would indicate that the model is not able to explain any of the variance. Based on these feature importances, it appears that the primary drivers of orange quality, according to the Random Forest Regression model, are the sweetness level (Brix) and the softness or texture (Softness rating). Other factors like ripeness, acidity, harvest time, and the presence of blemishes play a secondary role, while the physical characteristics like size and weight, as well as the color and variety, have a relatively minor impact on the quality prediction.

CONCLUSION

This research provides new insights into the use of the Random Forest algorithm for citrus quality classification, which is a step forward from traditional classification methods. The research results show that Random Forest can improve the accuracy of citrus classification, which is important for quality assessment and fruit distribution processes. This research also contributes to the development of smart agricultural technology, which can help farmers and distributors manage the quality of their products more efficiently. Future research could focus on improving the model by integrating other sensor data such as aroma or humidity for more comprehensive classification. Applying other machine learning techniques such as Deep Learning to compare performance with Random Forest in citrus quality classification. Conduct field studies to test the practical application of the Random Forest model in a real agricultural environment. Investigate the use of Random Forest in other fruit quality classifications to evaluate the universality of this method. This research is important because it can help in increasing efficiency and effectiveness in citrus quality classification, which can ultimately increase productivity and economic profits for farmers and distributors. In addition, the results of this research can be applied to support decisions in the agricultural industry and ensure better product quality for consumers.

REFERENCE

A. Primajaya et al. (2018). Random Forest Algorithm for Prediction of Precipitation," IJAIDM (Indonesian J. Artif. Intell. Data Mining), vol. 1, no. 1, pp. 27–31.

Amiri, M., Pourghasemi, H. R., Ghanbarian, G. A., & Afzali, S. F. (2019). Assessment of the Importance of Gully Erosion Effective Factors using Boruta Algorithm and Its Spatial Modeling and Mapping using Three Machine Learning Algorithms. Elsevier, 340, 55–69. https://doi.org/10.1016/j.geoderma.2018.12.042

Arinta, R. R., & Emanuel, A. W. R. (2019). Natural Disaster Application on Big Data and Machine Learning: A Review.

Dou, J., Yunus, A. P., Tien Bui, D., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.-W., Khosravi, K., Yang, Y., & Pham, B. T. (2019). Assessment of Advanced Random Forest and Decision Tree Algorithms for Modeling Rainfall-Induced Landslide Susceptibility in The Izu-Oshima Volcanic Island, Japan. Elsevier, 662, 332–346. https://doi.org/10.1016/j.scitotenv.2019.01.221

Firmansyah, Guntur & Hermawan, Arief. (2023). Implementasi Algortima Naive Bayes Untuk Klasifikasi Kesegaran Buah Jeruk. Jurnal Informatika. 10. 180-184. 10.31294/inf.v10i2.16115.

Fitri, V.A., Andreswari, R., and Hasibuan, M.A. (2019). Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm. Procedia Computer Science, 161: 765-772.

Iksan, N., Widodo, D.A., Sunarko, B., Udayanti, E.D. and Kartikadharma, E. (2021). Sentiment analysis of public reaction to COVID19 in twitter media using naïve Bayes classifier. In 2021 IEEE International Conference on Health, Instrumentation & Measurement, and Natural Sciences (InHeNce) (pp. 1-4). IEEE.

Kalumbang, Sri Wahyuni., Subanar. (2021). Comparison The Logistic Regression, Naive Bayes Classification, And Random Forest. Jurnal Matematika Thales (JMT): 2021 Vol. 03 No. 02

Khanvilkar, G., and Vora, D. (2019). Product Recommendation using 96 Sentiment Analysis of Reviews: A Random Forest Approach. International Journal of Engineering and Advanced Technology (IJEAT), 8: 2249-8958.

Khosravi, K., Pham, B. T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., & Tien Bui, D. (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Elsevier, 627, 744–755. https://doi.org/10.1016/j.scitotenv.2018.01.266

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

M. A. Ghani and A. Subekti. (2018). Email Spam Filtering Dengan Algoritma Random Forest," IJCIT (Indonesian J. Comput. Inf. Technol., vol. 3, no. 2, pp. 216–221.

Mustaqim, M., Warsito, B dan Surarso, B. (2019). Kombinasi Synthetic Minority Oversampling Technique (SMOTE) Dan Neural Network Backpropagation Untuk Menangani Data Tidak Seimbang Pada Prediksi Pemakaian Alat Kontrasepsi Implan, Jurnal Ilmiah Teknologi Sistem Informasi 5 (34), 116– 127.

Pandey, V. K., Sharma, K. K., Pourghasemi, H. R., & Bandooni, S. K. (2019). Sedimentological characteristics and application of machine learning techniques for landslide susceptibility modelling along the highway corridor Nahan to Rajgarh (Himachal Pradesh), India. Elsevier, 182, 104150. https://doi.org/10.1016/j.catena.2019.104150

Pangasturi, S.S. (2018). Perbandingan Metode Ensemble Random Forst Dengan Smote-Boosting Dan Smote-Bagging Pada Klasifikasi Data Mining Untuk Kelas Imbalance, Tesis., Surabaya

Parmar, A., Kataruya, R dan Petal, V. (2019). A Review on Random Forest: An Ensemble Classifier, Lecture Notes on Data Engineering and Communications Technologies 26, 758–763.

Tanyu, B.F., Abbaspour, A., Alimohammadlou, Y & Tecuci, G. (2021). Landslide susceptibility analyses using Random Forest, C4.5, and C5.0 with balanced and unbalanced datasets, Elsevier., USA

Ustyannie, W & Suprapto. (2020). Oversampling Method To Handling Imbalanced Datasets Problem In Binary Logistic Regression Algorithm, IJCCS., Yogyakarta

Utari, M., Warsito, B., Kusumaningrum, R. (2020). Implementation of Data Mining for Drop-Out Prediction Using Random Forest Method. In 2020 8th International Conference on Information and Communication Technology (ICoICT), IEEE, 1–5.

X. Gao and J. Wen. (2019). An Improved Random Forest Algorithm for Predicting Employee Turnover," Semant. Sch., vol. 2019.