Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 125~133
e-ISSN: 2548-6861

Classification Analysis of Single Tuition Fees Using the Random Forest Method with K-Fold Cross Validation

Al Khaidar [1]*, Nurdin [2]*, Fajriana [3]*, Taufiq [4]*, Defry Hamdhana [5]*
* Master of Information Technology Study Program, Malikussaleh University, Lhokseumawe, Indonesia
alkhaidarkutablang@gmail.com [1], nurdin@unimal.ac.id [2], fajriana@unimal.ac.id [3], taufq.te@unimal.ac.id [4], defryhamdhana@unimal.ac.id [5]

Article Info

Article history:

Received 2025-11-18
Revised 2025-12-25
Accepted 2026-01-07

Keyword:

K-Fold Cross Validation,
Random Forest,
Classification,
Single Tuition Fee.

ABSTRACT

Classification is the process of grouping data into specific categories based on their characteristics or features, which plays a crucial role in the analysis, decision-making, and prediction of new data. In academic settings, classification is used to determine the Single Tuition Fee to place students according to their economic ability. Lhokseumawe State Polytechnic has implemented the UKT system since 2020 with eight categories, but some students are still placed in UKT groups that do not match the results of the manual process, which has limited accuracy. This study uses the Random Forest method as a technology-based solution to improve the accuracy and objectivity of UKT classification. The dataset used consists of 10,000 student data with 10 variables, covering economic and social information. The research process includes data preprocessing, Random Forest model training, performance evaluation using accuracy, precision, recall, and F1-score, and model stability testing through 10-fold K-Fold Cross Validation. The results show that Random Forest is able to classify most UKT classes well, especially classes 0–5 and 7. Class 6 has lower performance with a recall of 0.39 and an F1-score of 0.56 due to the limited number of samples. The overall accuracy of the model reaches 96%, while K-Fold Cross Validation produces an average accuracy of 95.50% with a standard deviation of 0.66%, indicating the model is stable and able to generalize to new data. This study proves that Random Forest is effective in UKT classification, producing an objective, fair, and efficient system. This implementation model supports data-driven decision-making in higher education and increases transparency in UKT determination.

I. INTRODUCTION

Classification is the process of grouping data or objects into certain categories or classes based on their characteristics or features [1],[2]. The main purpose of classification is to facilitate analysis, decision making, or prediction of new data based on patterns that have been recognized from previous data [3],[4]. Classification machine learning is a supervised learning technique used to map input data into predetermined output labels or categories [5],[6]. Classification models are trained using labeled datasets so they can recognize patterns and make predictions about new data [7],[8].

Classification is not only used in the industrial, business, and health sectors, but also has an important role in the academic world, one of which is in determining the Single Tuition Fee [9] . The Single Tuition Fee is an education fee paid by students every semester, where all cost components such as building fees, tuition fees, practicums, and others have been combined into one fixed payment. UKT is implemented in state universities as part of the Indonesian government's policy through Permendikbud No. 55 of 2013 concerning Single Tuition Fees and Single Tuition Fees [10].

Lhokseumawe State Polytechnic (PNL) is a state university under the auspices of the Ministry of Research, Technology, and Higher Education, and is located in Lhokseumawe City,

Aceh Province. The determination of the Single Tuition Fee (UKT) amount at PNL refers to the campus' internal Circular Letter based on Regulation Number 25 of 2020 [11]. In addition, this policy also takes into account various other supporting regulations, such as Regulation Number 55 of 2013, changes contained in Regulation Number 73 of 2014, Regulation Number 22 of 2015, Regulation Number 39 of 2016, and the latest Regulation Number 2 of 2024. The UKT payment system has officially been implemented at Lhokseumawe State Polytechnic since October 22, 2020. In its implementation, students are grouped into eight UKT categories, starting from Group I to Group VIII, which are adjusted to the economic conditions of each student.

Problems identified through interviews with academic staff include the fact that some students from low-income backgrounds are still placed in inappropriate tuition groups, and the accuracy of the UKT classification process remains limited due to the manual process. These issues require a technology-based solution to improve the accuracy and efficiency of the Single Tuition Fee (UKT) classification.

The Random Forest method was chosen because it is capable of handling categorical and numerical data. The results of research conducted by Huynh-Cam focused on the development and evaluation of several student academic performance prediction models using the CART, C5.0, Random Forest, and MLP algorithms. The study found that CART obtained the highest accuracy of 80.00% compared to C5.0 (74.59%), Random Forest (79.99%), and MLP (69.02%) [12]. The results of previous research also conducted by Chen showed that the combination of Random Forest and Genetic Algorithm was able to achieve an average accuracy of 93.11% with a minimum increase of 2.25% compared to the baseline method, so this approach is effective for increasing the reliability of student performance predictions as a basis for decision making in the UKT classification system [13]. Based on these two previous studies, the decision tree-based model is considered superior and relevant to be applied to educational data-based classification systems such as determining UKT groups.

This research is able to produce a more accurate and fair UKT classification system, so that students from low-income families can be placed in cost groups according to their actual economic conditions. Decisions on UKT determination become more objective, data-based, and no longer rely on manual subjectivity, while also increasing the efficiency of higher education administration in the verification process. The novelty of this research lies in the use of 10 variables that are more comprehensive than previous research and the utilization of valid data obtained directly from institutional sources, so that the resulting classification model is more accurate, representative, and reflects the real conditions of students. The practical impact is the realization of a transparent and accountable UKT determination policy, increasing fair access to education for students, and providing a strong foundation for improving UKT policies in the future.

II. METHODOLOGY

A. Research Stages

This research stage is structured to provide a systematic overview of the research process, from problem identification, data collection and pre-processing, application of machine learning algorithms, to evaluation of classification results. These stages are designed to ensure the research is structured and scientifically accountable. The research stage flow is shown in Figure 1.

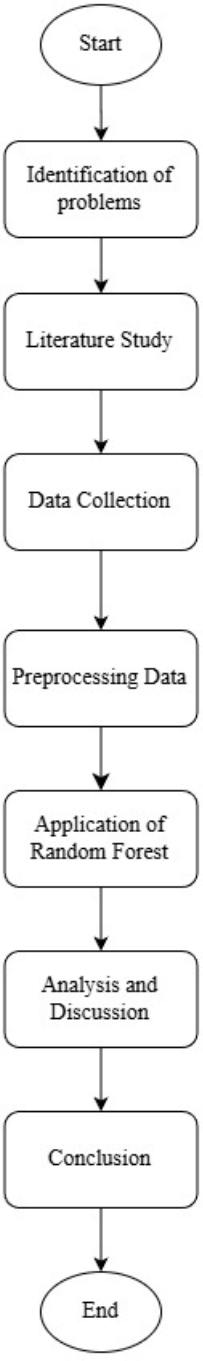

Figure 1. Research Stages

Figure 1 shows the overall research flow, which begins with problem identification as the initial step in determining the main issues related to UKT management and grouping. Once the problem is clearly formulated, the research continues with a literature review, which examines various theories, methods, and relevant previous research as a basis for model development. The next stage is data collection, where UKT data is obtained through internal institutional sources and prepared for further analysis. The collected data then enters a preprocessing stage, which includes data cleaning, handling missing data, normalization, and encoding to prepare the data for use by machine learning algorithms. The processed data is then applied to a machine learning model to classify UKT, with the aim of producing more objective and accurate UKT groupings or category predictions.

The model output is analyzed and discussed to assess its performance, accuracy, and relevance to the research problem. Based on this analysis, the research then formulates conclusions that describe the main findings and provide suggestions that can be used for further research development and system implementation within an institutional environment.

B. Data Collection

During data collection, researchers obtained primary information through interviews with relevant parties who are familiar with the UKT determination mechanism at Politeknik Negeri Lhokseumawe. The interview results were analyzed to identify and define relevant variables in the UKT classification process. Based on these findings, 10 primary variables were determined to serve as the basis for the modeling. Furthermore, this study utilized secondary data comprising 10,000 student records collected from the period of 2020 to 2025, which had been processed and adapted to suit the analysis needs. The variables used in this study can be seen in Table 1.

TABLE I.
VARIABLE DATA

No	Variable	Description	Measurement Scale
1	Father's Occupation	Type of father's occupation categorized into not working, farmer, fisherman, civil servant (PNS), entrepreneur, private employee, TNI/POLRI, and others	Categorical (Nominal)
2	Father's Income	Father's monthly income grouped into income intervals ranging from no income to Rp. 5,750,000	Numerical (Ordinal)
3	Father's Status	Father's marital and life status categorized as deceased, divorced, or alive	Categorical (Nominal)
4	Mother's Occupation	Type of mother's occupation categorized into not working, farmer, entrepreneur, private employee, civil servant (PNS), and others	Categorical (Nominal)
5	Mother's Income	Mother's monthly income grouped into income intervals ranging from no income to Rp. 5,500,000	Numerical (Ordinal)
6	Mother's Status	Mother's marital and life status categorized as deceased, divorced, or alive	Categorical (Nominal)
7	Number of Dependents	Total number of family members financially supported by parents, ranging from 1 to 20 dependents	Numerical (Discrete)
8	Water Source	Status of residence categorized as no ownership, living temporarily, monthly rent, yearly rent, or owned	Categorical (Ordinal)
9	Home Ownership	Main water source used by the household, including well, river/spring, PDAM, and bottled water	Categorical (Nominal)
10	Home Condition	Physical condition of the house categorized as poor, moderate, or good	Categorical (Ordinal)

Table 1 shows the ten economic and social variables used in this study, each accompanied by a clear description and measurement scale. The variables of occupation and parental status are presented on a categorical scale to describe the conditions of employment and the continuity of economic roles within the family, while the variables of father's and mother's income are arranged in income intervals with an ordinal scale to reflect the level of economic capacity in stages. The number of dependents is measured using a discrete numeric scale because it directly indicates the magnitude of the family's economic burden. In addition, the variables of air source, home ownership, and housing condition are used as indicators of family welfare and quality of life, expressed on a nominal or ordinal categorical scale according to the characteristics of the data. Presenting these detailed descriptions and measurement scales is important for assessing the validity of the features used, as it ensures that each variable truly represents the socioeconomic conditions of students accurately and can be processed consistently in the UKT classification model.

C. Random Forest

Random Forest, introduced by Leo Breiman in 2001, is a leading Ensemble Learning technique [14]. This method is an evolution of the Classification and Regression Tree (CART) algorithm, where it integrates two main techniques: Bootstrap Aggregating (Bagging) and Random Feature Selection [15],[16].

The Random Forest model offers several significant advantages: it is capable of providing accurate classification results (or high predictive accuracy), produces a minimal error rate, and can process massive training datasets efficiently.

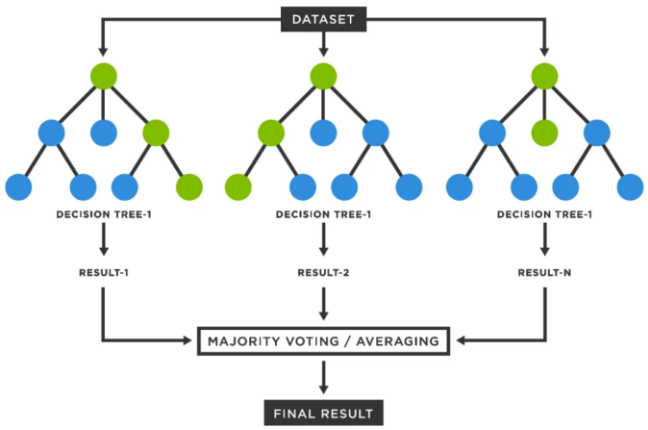

Figure 2. Random Forest Performance Scheme

Random Forest works by generating a set of diverse decision trees. To determine the final classification result, this method adopts the principle of majority voting; that is, the class most frequently predicted (mode) by all trees in the set will be selected as the final output [17], [18].

1. Perform random sampling with replacement (bootstrapping) from the initial dataset to form training data for each tree.
2. Randomly select a number of features (random feature selection) at each node to determine the best solver.
3. Calculate Entropy and Information Gain values to determine the optimal solver features.
4. Set the attribute with the highest Information Gain as the Root Node and build the tree recursively until the data is pure.
5. Determine Leaf Nodes as the final result if the data subset is pure.
6. Repeat the tree building process until k different decision trees are formed.
7. Combine the classification results from all trees using a majority voting mechanism to obtain a more accurate final prediction.

The procedure for constructing a Decision Tree begins by calculating two key metrics: the Entropy value and the Information Gain value [19],[20]. Entropy serves as an indicator to measure the level of impurity of an attribute. Entropy calculations can be performed for one attribute (using formula 1) or for two attributes involving a frequency table (using formula 2). After that, the Information Gain value is calculated using equation 3 to identify the most optimal solving feature.

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2\, p_i \quad (1)$$

Where:

$Entropy(S)$ = Entropy value of dataset S
S = Dataset set
C = Number of classes
p_i = Probability of class i in dataset

$$Entropy(T,X) = \sum_{c \epsilon x} P(c)E(c) \quad (2)$$

Where :

$Entropy(T,X)$ = Entropy after attribute T is separated based on attribute X
$P(c)$ = Probability of attribute class
$E(c)$ = Entropy value of attribute class

$$Gain(A) = Entropy(S) - \sum_{i=1}^{k} \frac{[S_i]}{[S]}\, x\, Entropy(S_i \quad (3)$$

Where :

$Gain\ (A)$ = Information Gain of attribute A
$Entropy\ (S)$ = Entropy value of dataset S
$[S_i]$ = Number of samples for value i
$[S]$ = Total number of data samples
$Entropy\ (S_i)$ = Entropy of the i-th subset after splitting

D. Implementation of the Random Forest Method

In this stage, the Random Forest method is applied to classify UKT categories based on the variables determined in the previous process. The modeling process begins with the formation of several decision trees trained using different data subsets, ensuring each tree has its own characteristics and prediction patterns. Next, all trees are combined through a majority voting mechanism to produce a more stable and accurate final prediction. This approach was chosen because Random Forest is able to overcome overfitting and provides

good performance on large data sets. The complete implementation of this method can be seen in Figure 3.

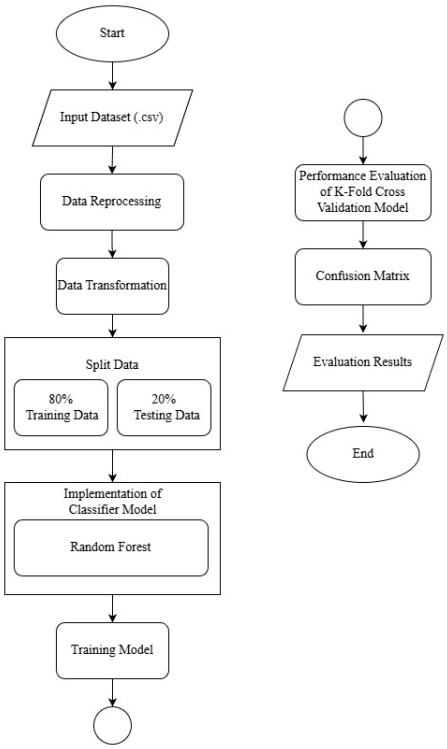

Figure 3. Implementation of the Random Forest Method

Figure 3 shows the process of implementing the Random Forest method in this study, starting with inputting a .csv dataset. Data preprocessing was then performed to ensure data quality and consistency before being used in modeling. Next, the data underwent a transformation process to adjust the format and variable scale required by the algorithm. The data was then divided into two parts: 80% as training data and 20% as test data. After that, the Random Forest classification model was implemented and the model was trained using the training data. The resulting model was then evaluated using the K-Fold Cross Validation method to test its performance consistency, and analyzed using a Confusion Matrix to obtain evaluation results covering accuracy, precision, recall, and other metrics.

E. Preprocessing Data

The data preprocessing stage is a crucial initial step in this research before the data is used in the modeling process. This stage aims to ensure that the data used is clean, consistent, and ready to be processed, thereby improving the accuracy and reliability of the classification model being developed. The stages are shown in Figure 4.

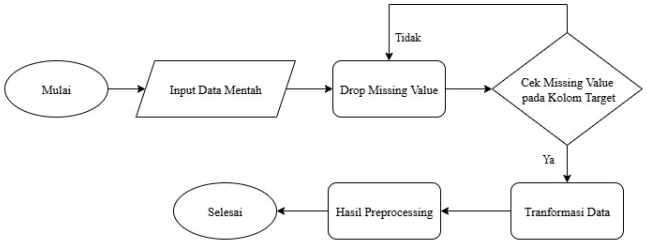

Figure 4. Data Preprocessing Stage

Figure 4 shows the data preprocessing process, starting with raw data input, followed by checking and handling missing values, particularly in the target column. Missing data are removed to maintain the validity of the model training process. Once the data is cleared of crucial missing values, a data transformation process is performed to meet the algorithm's requirements, resulting in ready-to-use preprocessed data. However, in addition to handling missing values, the preprocessing stage ideally also includes outlier handling to mitigate the influence of extreme values that can disrupt model performance, normalization or standardization of numeric variables to prevent scale differences from impacting the learning process, and encoding of categorical variables to convert non-numeric data into a numeric form that can be processed by the Random Forest algorithm, thus making the overall data more consistent and representative, and supporting improved classification accuracy.

F. Model Evaluation

Model performance evaluation was conducted to measure the accuracy and ability of the model to perform consistent classification. In this study, the evaluation process utilized two approaches: a Confusion Matrix to analyze predictive performance based on accuracy, precision, recall, and f1-score values, and K-Fold Cross Validation to assess model stability and generalization across different data sets. These two methods allow for a more comprehensive assessment of model quality.

III. RESULT AND DISCUSSION

This section presents the results and discussion of the application of the Random Forest model to classify UKT categories using 10,000 student data samples consisting of 10 variables. Analysis was conducted to evaluate the model's performance in recognizing patterns in each UKT category by utilizing metrics such as accuracy, precision, recall, F1-score, confusion matrix, and K-Fold Cross Validation. The discussion also includes the interpretation of the results obtained from each metric, thus providing a comprehensive understanding of the model's predictive capabilities, demonstrated advantages, and potential limitations in certain classes.

A. Data Distribution

Data distribution analysis was performed to identify the number of each label in the entire dataset, consisting of 10,000 entries. Examination of this distribution aims to obtain an overview of the proportion of each class, thus identifying the level of balance or potential imbalance in the data that could impact model performance. This information serves as an important basis for ensuring that the modeling and evaluation processes are comprehensive and representative. The results of the data distribution can be seen in Figure 3.

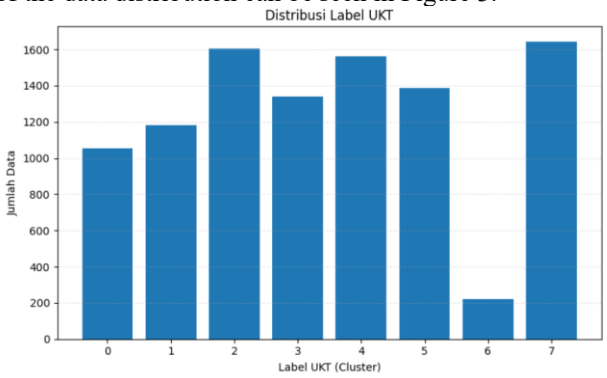

Figure 5. Label Distribusi

Figure 5 shows the distribution of UKT labels in the dataset, showing the varying amounts of data for each cluster. The clusters with the largest amounts of data are clusters 2, 4, and 7, with 1,604, 1,564, and 1,645 entries, respectively. This indicates that these three clusters have a more dominant representation within the overall dataset. Meanwhile, clusters 0, 1, 3, and 5 have relatively balanced amounts of data, ranging from 1,054 to 1,389 entries, thus still contributing significantly to the data composition.

Conversely, cluster 6 has the least amount of data, with only 222 entries. This difference in numbers between clusters illustrates the tendency of student data characteristics to be more clustered in certain categories than others. Understanding this distribution is important as a basis for understanding the representation of each label in the dataset, allowing interpretation of analysis and modeling results according to the proportions of the available data.

Correlation matrix analysis is performed to identify the relationships between variables in a dataset, both in terms of direction and strength of correlation. This matrix demonstrates the extent to which each variable influences each other, providing a deeper understanding of the data structure and the relationships between features used in the modeling process. The correlation matrix can be seen in Figure 6.

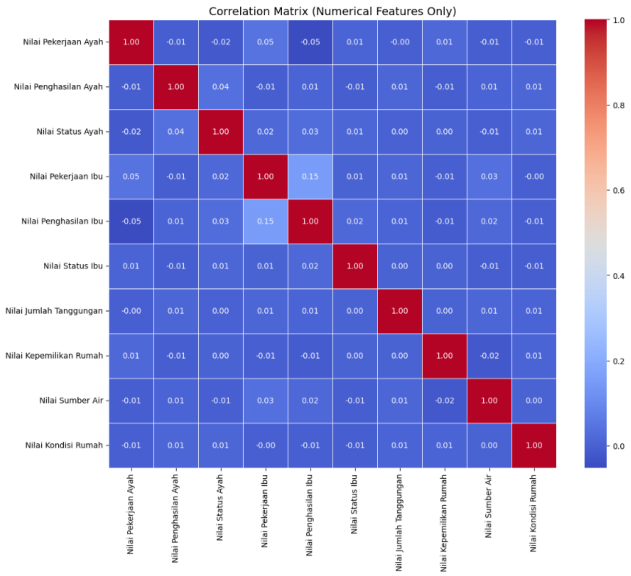

Figure 6. Correlation matrix

Figure 5 shows the correlation values in the table. All pairs of variables exhibit a very weak linear relationship, with values ranging from –0.05 to 0.03, and none exceed the generally low correlation threshold (|0.1|). The highest correlation value is found between Mother's Occupation Value and Mother's Income Value, at 0.15. However, this is still considered very weak and does not indicate a significant linear relationship. The largest negative correlation is between Father's Income Value and Mother's Occupation Value, at –0.05, which also falls within the very low correlation category. Overall, the numerical values in the table indicate that none of the variables have a strong statistical relationship with each other. The majority of values range from –0.01 to 0.02, indicating a minimal or almost non-existent relationship.

B. K- Fold Cross Validation

The K-Fold Cross Validation technique aims to evaluate model performance more comprehensively by dividing the dataset into several parts (folds) that are used alternately as training and testing data. This method helps reduce evaluation bias and ensures the model has good generalization capabilities. This study used 10 iterations (10-Fold Cross Validation) so that each data subset plays an equal role in the training and testing processes. The results of the K-fold cross validation test can be seen in Figure 6.

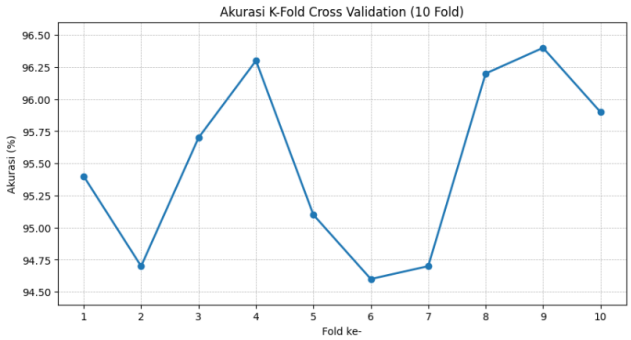

Figure 7. Result K-Fold Cross Validation

Figure 7 shows the results of K-Fold Cross Validation with 10 iterations, showing that the model accuracy for each fold varied: 0.954, 0.947, 0.957, 0.963, 0.951, 0.946, 0.947, 0.962, 0.964, and 0.959. This variation arises because each iteration uses a different portion of the test data, thus the complexity of the data in each fold also affects model performance. The sixth fold recorded the lowest accuracy of 0.946 because the test data portion in that iteration had more diverse patterns and did not completely resemble the training data. Conversely, the ninth fold achieved the highest accuracy of 0.964, indicating that the patterns in the test data in that iteration were more consistent with the patterns learned by the model.

Overall, the relatively small range of accuracy variation resulted in an average accuracy of 95.50% with a standard deviation of 0.66%, indicating that the model performed stably across all iterations. The low standard deviation shows that the model's performance did not fluctuate significantly even when tested on various data subsets. This indicates that the Random Forest model has good generalization capabilities and can be relied upon to consistently predict different data sets.

K-Fold Cross Validation shows that the model has good stability, as evidenced by the results of 10-fold K-Fold Cross Validation with a relatively small accuracy deviation between folds. This indicates that the model's performance tends to be consistent even when tested on different data subsets. However, the analysis presented is still minimal because it does not discuss in detail the performance distribution in each fold and the potential for data leakage, especially if there are student attributes that have strong correlations and have the potential to appear simultaneously in the training and test data, so this needs to be considered in the research.

C. Confusion Matrix

The Confusion Matrix is used to evaluate model performance in more detail by examining the number of correct and incorrect predictions for each class. This matrix reveals how the model differentiates between categories, including its ability to correctly identify the class and the types of prediction errors that occur. This evaluation helps understand the model's overall performance, not just based on a single accuracy value. The results of the confusion matrix can be seen in Figure 8.

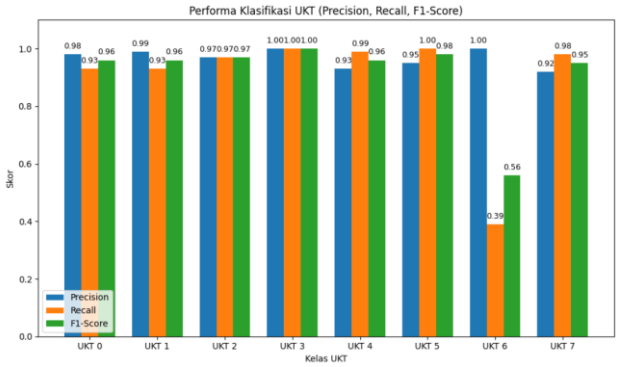

Figure 8. Classification Report

Figure 8 shows the model evaluation results, indicating that the Random Forest algorithm performed very well in classifying the UKT categories, achieving a total accuracy of 96%. In the UKT 0 and UKT 1 classes, the model achieved high precision of 0.98 and 0.99, a recall of 0.93 for both, and an F1-score of 0.96 each. The UKT 2 class achieved a precision of 0.97, a recall of 0.97, and an F1-score of 0.97. The UKT 3 class showed perfect performance with precision, recall, and F1-score all equal to 1.00. For UKT 4, the model produced a precision of 0.93, a recall of 0.99, and an F1-score of 0.96, while UKT 5 had a precision of 0.95, a recall of 1.00, and an F1-score of 0.98. Meanwhile, the UKT 6 class showed lower performance with a precision of 1.00, a recall of 0.39, and an F1-score of 0.56, indicating that some samples in this class were not recognized properly. The UKT 7 class performed very well with a precision of 0.92, a recall of 0.98, and an F1-score of 0.95. Overall, these results indicate that the Random Forest model works effectively for the majority of UKT classes and is suitable as a supporting model for determining UKT categories.

In terms of class-level interpretation, the high performance observed in UKT 0 and UKT 1 indicates that the model is able to reliably identify students who fall into the lowest tuition categories, which is crucial to ensuring that economically disadvantaged students are not mistakenly assigned to higher UKT levels. The strong results for UKT 2, UKT 3, UKT 4, and UKT 5 suggest that the model can accurately distinguish middle-range tuition categories, thereby supporting a proportional and consistent allocation of UKT based on students' economic conditions. However, special attention is required for the UKT 6 class, which exhibits a low recall value, indicating that several students who should belong to this category were misclassified into other UKT levels. This misclassification may lead to inequitable outcomes, such as students being assigned to inappropriate tuition categories that do not reflect their actual financial capacity. Meanwhile, the high performance of the UKT 7 class shows that students in the highest tuition category are generally identified correctly, reducing the risk of unfair subsidy allocation. Overall, this class-level analysis

highlights that while the model performs well across most UKT categories, careful consideration of misclassification in specific classes is necessary to maintain fairness in tuition fee determination.

```
=== AKURASI RANDOM FOREST (TEST 20%) ===
Akurasi: 95.50%

=== CLASSIFICATION REPORT ===
              precision    recall  f1-score   support

           0       0.98      0.93      0.96       211
           1       1.00      0.92      0.96       236
           2       0.97      0.97      0.97       321
           3       1.00      1.00      1.00       268
           4       0.94      0.93      0.94       313
           5       0.95      1.00      0.98       278
           6       1.00      0.39      0.56        44
           7       0.88      1.00      0.93       329

    accuracy                           0.95      2000
   macro avg       0.97      0.89      0.91      2000
weighted avg       0.96      0.95      0.95      2000
```

Figure 9. Confusion Matrix Evaluation

Figure 9 shows the results of the Random Forest model performance evaluation on 20% of the test data, where an accuracy rate of 95.50% was obtained, indicating that the model was able to classify UKT data very well. The classification report shows that almost all classes have high precision, recall, and f1-score values, with some classes even achieving a perfect f1-score of 1.00, indicating the model's accuracy in distinguishing each UKT category. The weighted average values for precision, recall, and f1-score, which reached 0.96, 0.95, and 0.95, respectively, indicate that the model remains stable despite the differences in the amount of data in each class. However, in one class with a relatively small amount of data, a low recall value is still seen, indicating the model's limitations in recognizing all data in that class. Overall, these results indicate that the Random Forest model has optimal performance and is reliable in the UKT classification process.

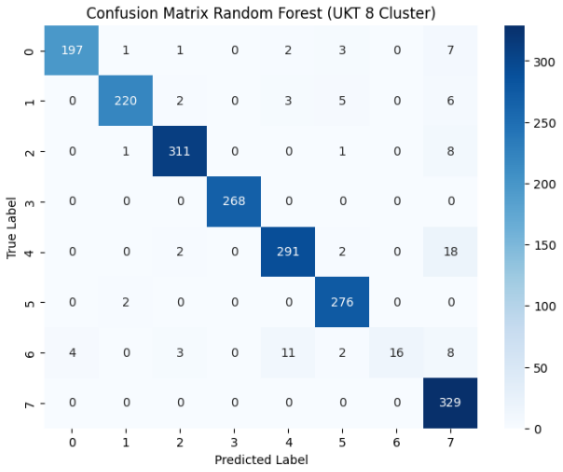

Figure 10. Confusion Matrix Random Forest

Figure 10 shows the confusion matrix data for each UKT class. The Random Forest model performed very well across most classes. Classes 0 to 5, and class 7, had high True Positive (TP) and low False Negative (FN) counts, indicating that the model was able to accurately recognize positive samples, although there were a few False Positive (FP) counts in some classes. Class 3 performed excellently with a TP of 268, a TN of 1732, an FP of 0, and an FN of 0. Conversely, class 6 performed poorly with a TP of only 16 and an FN of 28, indicating that most positive samples in this class were not detected. Overall, this confusion matrix indicates that the model was able to classify most UKT categories well, although it struggled with classes with very small sample distributions, such as class 6.

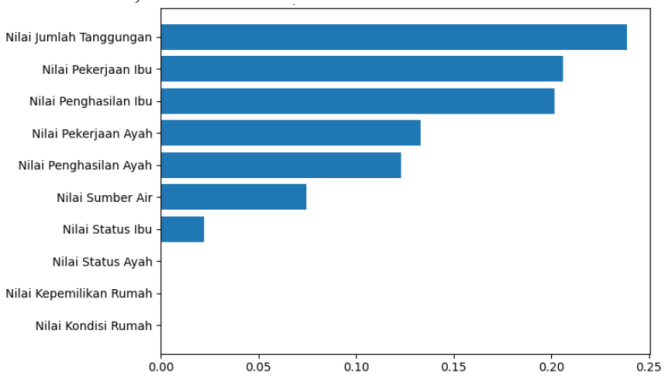

Figure 11. feature importance

Figure 11 shows that the number of dependents is the variable with the largest contribution in the UKT classification process, followed by maternal occupation and maternal income. This finding indicates that the family's economic burden and the mother's economic role have a significant influence in determining students' financial capabilities. Although the father's income does not rank first, this variable still makes a significant contribution along with the father's occupation, so that overall the parental income factor remains a major component in the UKT assessment. The water source variable shows a moderate influence as an indicator of welfare conditions, while the variables of parental status, home ownership, and housing conditions have a relatively small contribution of information, indicating that the contents of these variables are already represented by other economic variables. Thus, these results indicate that the UKT classification is more influenced by economic factors and the burden of family dependents than by physical factors of the residence.

IV. CONCLUSION

Based on the analysis results, the application of the Random Forest method to the UKT dataset with 10,000 samples and 10 variables effectively classified UKT categories, with an overall accuracy of 96%. Evaluation using precision, recall, and F1-score metrics showed high performance for most classes, particularly classes 0–5 and 7, which were successfully recognized. However, class 6 performed lower, with a recall of 0.39 and an F1-score of 0.56, indicating the model's difficulty recognizing samples in very small classes. This indicates that Random Forest is

capable of capturing dominant data patterns, but its performance declines for classes with limited sample distribution.

The results of the 10-fold K-Fold Cross Validation confirmed the model's stability and generalization capabilities, with an average accuracy of 95.50% and a standard deviation of 0.66%, indicating relatively small performance variations between data subsets. The low accuracy fluctuations indicate that the model remains consistent in predicting new data, even though each fold has varying data complexity. Overall, the combination of high accuracy results, stable performance, and good generalization capabilities confirms that Random Forest is a reliable and effective method for classifying UKT categories in this dataset, and can also be a reference for decision-making related to UKT management. As a recommendation, SMOTE can be applied to handle imbalanced UKT data by generating synthetic samples for minority classes. This helps the model better recognize underrepresented categories, improves recall and f1-score, and reduces bias toward majority classes. Overall, the use of SMOTE can enhance the stability and fairness of the Random Forest classification results.

REFERENCES

[1] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput Sci*, vol. 2, no. 3, 2021.

[2] U. I. Akpan, "Review of Classification Algorithms with Changing Inter-Class Distances," *J. King Saud*, 2021.

[3] N. Nurdin, "Analisa Data Mining Dalam Memprediksi Masyarakat Kurang Mampu Menggunakan Metode K-Nearest Neighbor," *J. Inform. Dan Tek. Elektro Terap.*, vol. 12, no. 2, 2025.

[4] A. Khaidar, "Analisis Sentimen Di Instagram Terhadap Menteri Keuangan Purbaya Yudhi Sadewa Menggunakan Metode Logistic Regression"," *Jitet*, Vol. 13, No. 3s1, 2025.

[5] K. Taha, "A Comprehensive Survey of Text Classification Techniques," *Expert Syst. Appl.*, vol. 202, pp. 117–134, 2024.

[6] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer. Behav Ther," 2021. doi: 10.1016/j.beth.2020.05.002.

[7] S. Kurnia and A. Khaidar, "Perbandingan Metode Machine Learning Menggunakan Metode Support Vector Machine Dan Artificial Neural Network Dalam Memprediksi Serangan Jantung," *J. Inform. Kaputama (Jik*, Vol. 9, No. 2, Pp. 87–94, 2025.

[8] N. Nurdin, M. Suhendri, Y. Afrilia, and R. Rizal, "Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (NBC," *Sist. J. Sist. Inf.*, vol. 10, no. 2, pp. 268–279.

[9] A. Khaidar, M. Arhami, and M. Abdi, "Application of the Random Forest Method for UKT Classification at Politeknik Negeri Lhokseumawe," *J. Artif. Intell. Softw. Eng.*, vol. 4, no. 2, pp. 94–103, 2024.

[10] K. P. Kebudayaan Republik Indonesia, "Peraturan Menteri Pendidikan dan Kebudayaan Nomor 2 Tahun 2024 tentang Biaya Kuliah Tunggal dan Uang Kuliah Tunggal pada Perguruan Tinggi Negeri di Lingkungan Kementerian Pendidikan dan Kebudayaan," 2024, *Jakarta*.

[11] K. P. Kebudayaan Republik Indonesia, "Peraturan Menteri Pendidikan dan Kebudayaan Nomor 25 Tahun 2020 tentang Standar Satuan Biaya Operasional Pendidikan Tinggi pada Perguruan Tinggi Negeri di Lingkungan Kementerian Pendidikan dan Kebudayaan," 2020, *Jakarta*. [Online]. Available: https://peraturan.bpk.go.id/Details/163756/permendikbud-no-25-tahun-2020

[12] T.-T. Huynh-Cam, L.-S. Chen, and H. Le, "Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance," *Algorithms*, vol. 14, no. 11, p. 318, 2021, doi: 10.3390/a14110318.

[13] M. Chen and Z. Liu, "Predicting performance of students by optimizing tree components of random forest using genetic algorithm," *Heliyon*, vol. 10, no. 12, p. e32570, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e32570.

[14] J. Gao and Y. Liu, "Prediction and the influencing factor study of colorectal cancer hospitalization costs in China based on machine learning-random forest and support vector regression: a retrospective study," *Front Public Heal.*, vol. 12, no. 1211220, 2024, doi: 10.3389/fpubh.2024.1211220.

[15] E. Widhiastuti, "Implementasi Data Mining Untuk Memprediksi Penyakit Hipertensi Dalam Kehamilan Menggunakan Algoritma C4. 5 (Study Kasus: Puskesmas RImba Melintang, Rokan Hilir," 2021.

[16] M. Sitanggang, E. Simamora, and F. D. Mobo, "Increasing Accuracy of Classification in C4.5 Algorithm by Applying Principle Component Analysis for Diabetes Diagnosis," *Numer. J. Mat. Dan Pendidik. Mat.*, vol. 6, no. 2, pp. 175–186, 2022, doi: 10.25217/numerical.v6i2.2610.

[17] E. R. B. Sebayang, Y. H. Chrisnanto, and M. Melina, "Klasifikasi Data Kesehatan Mental di Industri Teknologi Menggunakan Algoritma Random Forest," *IJESPG (International J. Eng. Econ. Soc. Polit. Gov.*, vol. 1, no. 3, pp. 237–253, 2023.

[18] N. Nurdin, F. Fajriana, M. Maryana, and A. Zanati, "Information System for Predicting Fisheries Outcomes Using Regression Algorithm Multiple Linear," *J. Informatics Telecommun. Eng.*, vol. 5, no. 2, pp. 247–258.

[19] Y. Wang, P. Jia, L. Liu, C. Huang, and Z. Liu, "A systematic review of fuzzing based on machine learning techniques," *PLoS One*, vol. 18;15(8):e, 2020, doi: 10.1371/journal.pone.0237749.

[20] T. P. Quinn *et al.*, "Machine Learning in Psychiatry (MLPsych) Consortium. A primer on the use of machine learning to distil knowledge from data in biological psychiatry," *Mol Psychiatry*, vol. Feb;29(2):, 2024, doi: 10.1038/s41380-023-02334-2.