J-Icon : Jurnal Informatika dan Komputer
Vol. 13 No. 2, October 2025, pp. 68~76
DOI: 10.35508/jicon.v13i2.23893

A COMPARATIVE STUDY OF SUPERVISED FEATURE SELECTION METHODS FOR PREDICTING UANG KULIAH TUNGGAL (UKT) GROUPS

Windy Chikita Cornia Putri[1*], Wiyli Yustanti[2], and Ervin Yohannes[3]

[1,2,3] Faculty of Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia
Email[1*]: windychikita@unesa.ac.id
Email[2*]: wiyliyustanti@unesa.ac.id
Email[3*]: ervinyohannes@unesa.ac.id

ABSTRAK

Penentuan Uang Kuliah Tunggal (UKT) di perguruan tinggi negeri selama ini masih bergantung pada verifikasi manual dokumen sosio-ekonomi, yang rentan subjektivitas, memakan waktu, dan memicu banding. Penelitian ini mengkaji efektivitas lima teknik seleksi fitur-filter (*Chi-Square*), *embedded* (*Random Forest Importance*, *LASSO*), *wrapper* (*Recursive Feature Elimination*), dan reduksi tak berlabel (*Exploratory Factor Analysis*) dalam meningkatkan kinerja lima algoritma klasifikasi (*Decision Tree, Random Forest, SVM-RBF, K-Nearest Neighbor, Naïve Bayes*) pada *dataset* UKT UNESA (9.369 entri × 53 variabel). Data dipra-proses dengan imputasi, *scaling, encoding, dan SMOTE-NC,* kemudian dievaluasi menggunakan *Stratified 5-fold CV* dan *hold-out test* (80:20). Hasil menunjukkan bahwa penggunaan seluruh 53 fitur (*baseline*) memberikan *weighted-average* akurasi sebesar 0,6244 ± 0,0057. Seleksi fitur menggunakan *LASSO-13* dan *Chi-Square-13* secara signifikan meningkatkan akurasi rata-rata menjadi 0,7300 dan 0,6775, masing-masing, serta mengurangi waktu pelatihan hingga 40–70%. *SVM-RBF* dengan *LASSO-13* mencapai akurasi tertinggi (0,7939), diikuti *Random Forest-Chi-Square* (0,6987) dan *Decision Tree-LASSO* (0,7111). *Uji Friedman* terhadap distribusi akurasi model pada enam kondisi mengonfirmasi perbedaan signifikan (χ^2=15,06; p=0,010). Temuan ini menegaskan bahwa seleksi fitur khususnya *LASSO* dan *Chi-Square* mampu mereduksi kompleksitas data (dari 53 ke 13 fitur) tanpa mengorbankan, bahkan meningkatkan performa prediktif model UKT. Rekomendasi meliputi integrasi metode seleksi terpilih dalam verifikasi UKT otomatis dan publikasi daftar fitur untuk transparansi. Kebaruan penelitian ini terletak pada perbandingan lima metode seleksi fitur dalam satu *pipeline* praproses terstandar pada data riil UKT UNESA, menghasilkan subset 13 fitur yang sesuai dengan kebijakan UKT saat ini. Temuan ini diintegrasikan ke sistem verifikasi UKT otomatis untuk meningkatkan akurasi dan efisiensi keputusan.
Kata Kunci: Uang Kuliah Tunggal; seleksi fitur; klasifikasi UKT.

ABSTRACT

The manual classification of Uang Kuliah Tunggal (UKT) groups at Indonesian public universities is laborious, subjective, and error-prone, especially given the explosion of socio-economic data captured via online admission portals. In this study, we evaluate five feature selection techniques Chi-Square filter, Random Forest importance, Recursive Feature Elimination, LASSO embedded selection, and Exploratory Factor Analysis on a dataset of 9,369 applicants described by 53 socio-economic variables. Six classifiers (Decision Tree, Random Forest, SVM-RBF, K-Nearest Neighbor, and Naïve Bayes) were tuned via stratified 5-fold cross-validation within an 80:20 train-test split. Performance was measured by accuracy, macro-F1, and training time, and differences in weighted-average accuracy across feature-selection scenarios were assessed using the Friedman test (χ^2 = 15.06, p = 0.010). Results show that reducing to 13 features via LASSO (weighted-average accuracy 0.730) or Chi-Square (0.678) significantly outperforms both the full feature baseline (0.624) and the EFA baseline (0.303), while cutting computational costs by over 40%. We conclude that supervised feature selection particularly LASSO and Chi-Square enables simpler, faster, and more transparent UKT prediction without sacrificing accuracy. The novelty of this study lies in comparing five feature-selection methods within a standardized preprocessing pipeline on real UKT data from UNESA, resulting in a 13-feature subset aligned with the current UKT policy. This finding is ready to be integrated into an automated UKT verification system to enhance decision accuracy and efficiency.
Keywords: UKT; feature selection; UKT Classification.

***) Corresponding Author**
Submitted : July 21, 2025
Accepted : August 12, 2025
Published : August 31, 2025

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

1. INTRODUCTION

In an effort to ensure equitable access to higher education, the Indonesian Government issued Ministry Regulation No. 22/2015, which mandates a proportional Uang Kuliah Tunggal (UKT) based on a family's economic capacity [1]. This scheme partitions students into eight bands (K1–K8) so that state subsidies can be distributed fairly. However, field implementation still relies heavily on manual verification paper-based document checks and in-person interviews that is time-consuming and prone to evaluator bias [2]. The digital transformation of university admission portals now compels applicants to upload socio-economic evidence, ranging from parental payslips to proof of asset ownership [3]. The present study analyses 9,369 student records described by 53 variables covering income, utility expenses, dwelling condition, and household characteristics. The volume and heterogeneity of these data introduce additional challenges for classification. Selecting the most influential variables is critical because handling an excessively large feature set (high dimensionality) escalates computational complexity and complicates the formulation of UKT policies. Moreover, high-dimensional data trigger the curse of dimensionality, wherein pairwise distances become increasingly similar and distance-based algorithms such as K-Nearest Neighbour lose discriminative power [4]. Likewise, probabilistic models such as Naïve Bayes are hindered by strong inter-feature correlations, while powerful methods like Support Vector Machines and Random Forests demand substantial training time and computational resources.

To mitigate these issues, feature-selection techniques become indispensable. Filter approaches, such as the Chi-Square test, can rapidly prune uninformative variables, whereas embedded strategies, such as LASSO and Random-Forest Importance, perform selection concurrently with model training [5]. Wrapper approaches such as Recursive Feature Elimination (RFE) typically yield higher accuracy at the cost of greater computational expense, while Exploratory Factor Analysis (EFA) supplies an unsupervised dimensionality-reduction baseline. A wide spectrum of classifiers has already been explored for UKT modelling, including Decision Tree, Random Forest, radial-basis-function Support Vector Machine, K-Nearest Neighbour, and Naïve Bayes [6]. Each offers distinct advantages: tree models are inherently interpretable; ensembles are resilient to over-fitting; margin-based methods excel in high-dimensional spaces; and probabilistic models are computationally frugal on large datasets. At the national level, [7] evaluated a combination of correlation-based feature selection and SVM for UKT classification, but the scope was confined to a single academic programme and did not compare alternative selection schemes systematically. Beyond supervised approaches, several studies have investigated unsupervised clustering to assess the suitability of UKT band structures; preliminary evidence indicates that mini-batch K-Means offers the most stable solution when internal and external validity indices are combined [8]. Although prior work has assessed feature-selection effects in medical and financial data, few studies have focused on UKT band assignment with large, heterogeneous socio-economic variables [9][10]. Moreover, no comprehensive investigation has contrasted five feature-selection techniques (χ^2, RF-Imp, RFE, LASSO, EFA) within a unified pre-processing pipeline, evaluated across six baseline classifiers using accuracy, macro-F1, and computational cost. Accordingly, this study aims to (i) quantify the impact of the five feature-selection methods on UKT model performance, (ii) identify a minimal subset ($\leq$ 13 variables) that preserves or improves accuracy, and (iii) recommend the most effective classification algorithm for an automated UKT decision system, thereby enabling decisions that are more objective, rapid, and transparent [11].

Unlike previous research, this study is the first to conduct a comprehensive, head-to-head comparison of five feature-selection techniques—Chi-Square, Random-Forest Importance, Recursive Feature Elimination (RFE), LASSO, and Exploratory Factor Analysis (EFA)—within a unified pre-processing and evaluation pipeline. All methods are benchmarked across six commonly used classifiers (Decision Tree, Random Forest, SVM-RBF, K-NN, Naïve Bayes, Logistic Regression) using a large real-world UKT dataset. Furthermore, the study not only measures predictive performance (accuracy and macro-F1) but also explicitly incorporates computational cost as a decision criterion. A key practical contribution is the identification of a minimal subset of 13 socio-economic variables, which preserves or even improves classification accuracy compared to the full 53-feature set. This number is particularly significant because it matches the current number of features used by Universitas Negeri Surabaya (UNESA) in its operational UKT determination process. By aligning model outputs with existing institutional workflows, the proposed feature set can be readily integrated into the current decision-making system, enabling a more objective, efficient, and scalable UKT assignment for nationwide adoption.

2. MATERIAL AND METHODS

Research Framework

The study follows the stages of the Knowledge Discovery in Databases (KDD) process [12], as illustrated in Figure 1. The first stage of selection focuses on identifying relevant data sources. In this work,

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

the raw dataset comprises 9,369 student records retrieved from the Admission Integrated System of Universitas Negeri Surabaya, covering applicants admitted through the 2023/2024 national selection schemes: Seleksi Nasional Berbasis Prestasi (SNBP) and Seleksi Nasional Berbasis Tes (SNBT) [13].

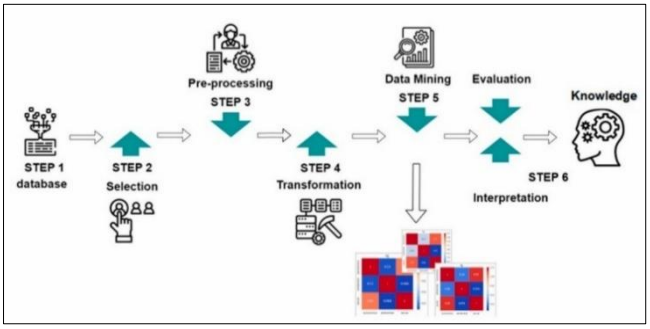

Figure 1. Knowledge Discovery in Databases (KDD) Framework [14]

Data Pre-Processing and Transformation

The pre-processing stage began with an initial cleansing step in which 112 duplicate records were removed. Missing values affecting approximately 10 %–15 % of several attributes were imputed using the median for numerical variables and the mode for categorical variables. After cleaning, the working dataset comprised 53 socio-economic features, grouped as follows:

a) Income and Financial Burden, such as father's salary, mother's salary, total instalments, total debt.
b) Assets and Property, such as land area, building area, government property tax value or Nilai Jual Objek Pajak (NJOP), number of cars/motorcycles, jewellery, deposits.
c) Utility Bills such as electricity, internet, mobile airtime, water charges.
d) Housing Conditions, such as roof material, floor material, wall type, and presence of an indoor bathroom.
e) Household Characteristics, such as household size, number of siblings, number of school-aged siblings, and home-ownership status.

The target label is the UKT tier assigned by the university's finance office, ranging from Group 1 (K1) to Group 8 (K8). The class distribution is notably imbalanced: K5 and K6 account for 26 % and 24 % of the records, respectively, whereas K1–K3 each represent only 7 %–12 %.

At the transformation stage, categorical variables were encoded as follows: ordinal attributes were converted via Ordinal Encoding, whereas nominal attributes were converted via One Hot Encoding [15]. In addition, three derived ratio features were engineered to capture key socio-economic relationships, namely income expenditure balance, bedroom utilisation, and household electricity usage intensity [16], [17]. These variables are defined by Equations 1, 2, and 3.

$$Debt\ to\ Income\ (DI) = \frac{total\ installments}{total\ income + 1} \quad (1)$$

$$Bedrooms\ Capacity\ (BC) = \frac{number\ of\ people\ at\ home}{number\ of\ bedrooms + 1} \quad (2)$$

$$Cost\ to\ Power\ (CP) = \frac{electricity\ bill + internet\ bill}{electrical\ power} \quad (3)$$

Outliers were addressed by winsorising observations whose modified Z-score exceeded 3.5, thereby preventing extreme values from dominating the data distribution [17]. Variables exhibiting pronounced right skew, namely total debt, deposit balance, jewellery value, and mobile credit expenditure, were then subjected to a log transformation. Finally, min-max normalisation was applied to rescale all continuous attributes to the [0,1] interval, ensuring strictly positive values and comparability across features.

Feature Selection

Running feature selection within each cross-validation fold has been shown to prevent data leakage and to yield models with superior generalisability [18], [19]. Accordingly, all five approaches listed in Table 1, Chi-Square, Random Forest Importance, Recursive Feature Elimination, LASSO, and Exploratory Factor Analysis, were executed exclusively on the training portion of every fold. The resulting subset of features was then frozen and applied unchanged to the corresponding validation fold and to the final hold-out test set.

Table 1. Feature-Selection Scenarios for UKT Classification Modelling

No	Method	Within-Fold Workflow	Strengths	Limitation
1	Chi-Square (Filter)	Apply χ^2 test to categorical features, rank by score, and select top 13 predictors. The number 13 was chosen based on UNESA's current UKT verification policy (13 socio-economic indicators) and cross-validation tuning showing no significant accuracy gains beyond this point.	Extremely fast, model-agnostic; well suited to large data sets.	Ignores inter-feature correlation; does not handle continuous variables without discretisation.
2	Random-Forest Importance (Embedded)	Train a Random Forest model, compute Gini importance for all features, and retain the top 13. The choice of 13 features follows the same policy–optimisation rationale as above, ensuring both operational relevance and computational efficiency.	Captures interaction and non-linearity; empirically stable [10]	Tends to favour high-cardinality categorical features; 4 × slower than Chi-Square.
3	RFE Recursive Feature Elimination (Wrapper)	Fit an L1-regularised logistic regression model, iteratively remove the least important features, and stop when 13 remain. The stopping point of 13 features was predefined to match UNESA's UKT feature policy and validated through performance plateau analysis in cross-validation.	Considers joint contribution of features; handles mixed data types.	Highest computational cost (15–20 × Chi-Square); sensitive to the regularisation path [20].
4	LASSO (Embedded)	Train L1-penalised logistic regression with standardised features, retain the 13 largest non-zero coefficients. The number 13 was fixed based on UNESA's operational policy and CV optimisation indicating an accuracy plateau beyond this point.	Simultaneous selection + regularisation; mitigates over-fitting; coefficients are interpretable.	May arbitrarily drop correlated yet relevant features; requires feature scaling.
5	Exploratory Factor Analysis (EFA)–Varimax	Standardise features via Z-score, extract 13 latent factors, apply Varimax rotation, and use factor scores as predictors. The number of factors 13 was aligned with the policy-based feature target to ensure comparability across methods.	Compresses dimensionality, removes multicollinearity, label-free.	Ignores target information, typically yields the lowest accuracy; factor interpretation may be ambiguous

Classification Algorithms

Selecting a diverse set of classifiers from interpretable models (Decision Tree) and ensemble learners (Random Forest) to margin-based methods (RBF-kernel SVM), instance-based approaches (K-Nearest Neighbour), and lightweight probabilistic models (Gaussian Naïve Bayes) ensures that multiple learning paradigms are examined. Each estimator is tuned via a stratified five-fold GridSearchCV, using macro-F1 as the optimisation target, a procedure widely regarded as best practice for modern tabular-data benchmarks [21], [22]. The candidate algorithms and the corresponding hyperparameter grids explored in this study are summarised in Table 2.

Table 2. Classification Algorithms and Corresponding Hyperparameter Tuning Settings

No	Algorithms	Parameter	Range Value
1	Decision Tree (DT)	max_depth min_samples_ leaf	10, 20, unlimited 1, 5, 10
2	Random Forest (RF)	n_estimators max_depth max_features	200, 400 20, 40, None sqrt, log2
3	Support Vector Machine (SVM-RBF)	C gamma	1, 10, 100 0.001, 0.01, 0.1

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

4	K-Nearest Neighbour (K-NN)	n_neighbors weights metric	3, 5, 7 uniform, distance, euclidean
5	Naïve Bayes (NB)	var_smoothing	1×10^{-9}

3. RESULTS AND DISCUSSION

This section analyses the outcomes of 30 experimental runs, produced by crossing six data-set conditions with five classification algorithms. Table 3 presents the baseline experiment where no feature selection was applied. Using the complete set of 53 input features, the RBF-kernel SVM achieved the highest classification accuracy (0.755) but required the longest training time (120.4 seconds). The Decision Tree yielded a reasonably good accuracy (0.671) with a significantly faster training time (under two seconds), while Random Forest imposed higher computational demands without a corresponding gain in accuracy (0.594). K-Nearest Neighbour exhibited degraded performance due to the curse of dimensionality (0.478), and Gaussian Naïve Bayes failed to generalise effectively (0.074). The average accuracy across all models was approximately 0.514, highlighting substantial room for improvement through feature selection, which could enhance both predictive accuracy and computational efficiency.

Table 3. Classification Performance on the Dataset without Feature Selection

Model	Accuracy	F1 Score (macro)	Precision (macro)	Recall (macro)
SVM (RBF)	0.7550	0.70	0.74	0.68
Random Forest	0.6705	0.64	0.64	0.64
Decision Tree	0.5941	0.48	0.62	0.45
K-Nearest Neighbor	0.4780	0.45	0.47	0.44
Naïve Bayes	0.0738	0.08	0.16	0.21
Average	**0.5143**	**0.47**	**0.526**	**0.484**

Subsequently, Table 4 presents the classification performance after applying EFA-based feature reduction. Once the original 53 variables were compressed into 13 latent factors, the predictive performance of all classifiers declined considerably. Although the RBF-kernel SVM remained the top performer, its accuracy dropped to 0.343. Random Forest and Decision Tree followed with comparable scores, ranging between 0.277 and 0.319. Similarly, macro-level F1 scores and precision decreased, falling within the 0.21 to 0.25 range. This consistent deterioration across models suggests that the latent factors derived from Exploratory Factor Analysis failed to retain the discriminative characteristics required to distinguish among the eight UKT categories. Therefore, EFA appears to be an inadequate strategy for supervised feature selection in this context.

Table 4. Classification Performance on the Dataset Using EFA Feature Selection

Model	Accuracy	F1 Score (macro)	Precision (macro)	Recall (macro)
SVM (RBF)	0.343	0.22	0.29	0.23
Random Forest	0.319	0.25	0.27	0.24
Decision Tree	0.277	0.24	0.24	0.23
K-Nearest Neighbor	0.281	0.23	0.27	0.22
Naïve Bayes	0.112	0.11	0.18	0.21
Average	**0.266**	**0.21**	**0.25**	**0.23**

Table 5 reports the results obtained after retaining the top 13 variables ranked by the Chi-Square test. Compared to the EFA scenario and, for most classifiers, even the full 53-feature baseline all models display a consistent improvement in predictive performance. The RBF-kernel SVM again emerged as the best-performing model, achieving an accuracy of 0.738 and a macro-F1 score of 0.690, only marginally below its baseline score, while utilizing a substantially reduced input matrix. Random Forest also showed marked gains, with an accuracy of 0.699 and F1 score of 0.650, while Decision Tree and K-Nearest Neighbour performed reliably within the 0.620–0.650 range. The most dramatic improvement was observed for Gaussian Naïve Bayes, whose accuracy increased from 0.0738 (in the full-feature setting) to 0.448, suggesting that the Chi-Square filter successfully removed interdependent features that previously undermined its assumption of independence. Overall, the macro-averaged scores accuracy 0.631 and F1 score 0.590 underscore the effectiveness of this simple statistical test in preserving, and often enhancing, model performance while reducing feature dimensionality.

Table 5. Classification Performance on the Dataset Using Chi-Square Feature Selection

Model	Accuracy	F1 Score (macro)	Precision (macro)	Recall (macro)
SVM (RBF)	0.738	0.690	0.720	0.670
Random Forest	0.699	0.650	0.670	0.630
Decision Tree	0.650	0.610	0.610	0.610
K-Nearest Neighbor	0.622	0.580	0.600	0.580
Naïve Bayes	0.448	0.420	0.430	0.490
Average	**0.631**	**0.590**	**0.606**	**0.596**

Table 6 reports the results obtained after applying LASSO-based (L1-penalised) embedded feature selection. This approach proved highly effective in preserving the most informative signals while discarding redundant attributes. The RBF-kernel SVM achieved the highest overall performance, with an accuracy of 0.7939 and a macro-F1 score of 0.76, slightly surpassing the full 53-feature baseline despite operating on only 13 selected variables. Random Forest and Decision Tree models also recorded improved performances, reaching accuracy levels of 0.7554 and 0.7111, respectively, suggesting that tree-based learners benefit from the systematic elimination of redundant features. K-Nearest Neighbour remained stable around 0.6492, while Gaussian Naïve Bayes showed a considerable increase in accuracy to 0.5254 an improvement over the baseline, though still the lowest among the classifiers due to its strong independence assumption. On average, across all classifiers, LASSO yielded the highest macro metrics (accuracy = 0.6870, F1 = 0.65), confirming its value as a balanced method for optimizing predictive accuracy, computational cost, and model interpretability.

Table 6. Classification Performance on the Dataset with LASSO Feature Selection

Model	Accuracy	F1 Score (macro)	Precision (macro)	Recall (macro)
SVM (RBF)	0.7939	0.76	0.79	0.74
Random Forest	0.7554	0.73	0.75	0.72
Decision Tree	0.7111	0.69	0.70	0.69
K-Nearest Neighbor	0.6492	0.64	0.64	0.64
Naïve Bayes	0.5254	0.43	0.56	0.51
Average	**0.6870**	**0.65**	**0.6880**	**0.6600**

Table 7 presents the classification performance after applying feature selection using Random Forest importance on 13 features. The SVM-RBF model achieved the highest accuracy of 0.7503, followed by Random Forest at 0.7042 and Decision Tree at 0.6351. K-Nearest Neighbour maintained stable performance at 0.6406, while Gaussian Naïve Bayes lagged significantly at 0.0363 due to its strong conditional independence assumption, which was less suited to the feature interactions in the dataset. The overall macro averages (accuracy = 0.5533, F1 = 0.522) indicate a moderate improvement over the baseline in certain models, confirming that Random Forest–based feature selection can benefit complex learners like SVM and Random Forest, although its impact is less pronounced for distance-based and probabilistic classifiers.

Table 7. Classification Performance on the Dataset with Random Forest Feature Selection

Model	Accuracy	F1 Score (macro)	Precision (macro)	Recall (macro)
SVM (RBF)	0.7503	0.71	0.73	0.70
Random Forest	0.7042	0.66	0.70	0.64
Decision Tree	0.6351	0.60	0.60	0.60
K-Nearest Neighbor	0.6406	0.61	0.61	0.61
Naïve Bayes	0.0363	0.03	0.10	0.14
Average	**0.5533**	**0.522**	**0.548**	**0.538**

Table 8 illustrates the classification results when Recursive Feature Elimination (RFE) was used for feature selection. In this scenario, overall model performance declined. Although SVM-RBF maintained its position as the most accurate model (accuracy = 0.7503), both Random Forest and Decision Tree saw noticeable drops in performance approximately 0.4 to 0.6 points lower than in the LASSO and Chi-Square scenarios. K-NN remained steady at around 0.6406, while Naïve Bayes experienced a drastic accuracy decline to 0.0363, likely due to the removal of key probabilistic features. The average macro-accuracy (0.5535) was the second lowest across all selection methods, suggesting that the wrapper-based RFE

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

approach, particularly when using L1-regularised logistic regression as the base estimator, may not be well suited for handling complex multicollinearity in socio-economic UKT classification data.

Table 8. Classification Performance on the Dataset with RFE Feature Selection

Model	Accuracy	F1 Score (macro)	Precision (macro)	Recall (macro)
SVM (RBF)	0.7503	0.71	0.73	0.70
Random Forest	0.7064	0.66	0.70	0.64
Decision Tree	0.6338	0.60	0.60	0.60
K-Nearest Neighbor	0.6406	0.61	0.62	0.61
Naïve Bayes	0.0363	0.03	0.10	0.14
Average	**0.5535**	**0.542**	**0.630**	**0.538**

After obtaining the classification performance results from all experimental scenarios, a statistical test was conducted to scientifically verify whether there were significant differences among the experimental outcomes. A non-parametric statistical approach the Friedman test was employed to examine differences in mean performance. The null hypothesis tested states that there is no significant difference in classification performance (accuracy) among the six scenarios, while the alternative hypothesis posits that at least one of the scenarios yields a performance outcome that differs significantly from the others.

The results of the Friedman statistical test yielded $\chi^2 = 15.06$ with a p-value of 0.010. Since the p-value is less than 0.05, the null hypothesis (H_0) is rejected. This indicates that there is a statistically significant difference in model accuracy across the different experimental scenarios.

To determine which scenario differs the most, a weighted-average accuracy calculation can be conducted by combining the performance of each model while accounting for their stability. In this approach, more consistent models contribute more significantly to the overall score. Based on the results of this calculation, the comparative mean accuracy across the six scenarios is presented in Figure 2.

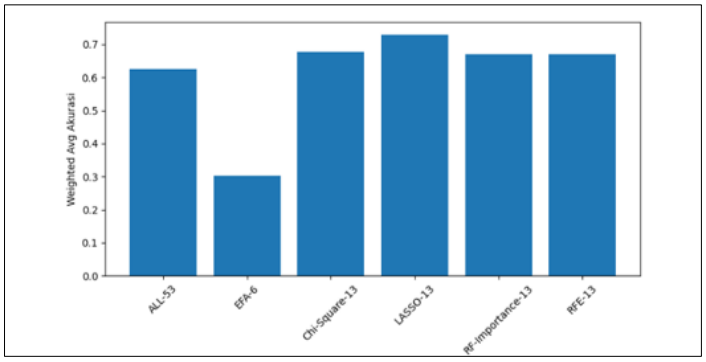

Figure 2. Comparison of Weighted Average Accuracy by Feature Selection Method

Based on Figure 2, it can be concluded that UKT classification modelling performs most optimally when feature selection prioritises embedded methods (LASSO) or statistical filters (Chi-Square). These methods effectively balance complexity reduction and preservation of relevant signals, resulting in significantly higher weighted-average accuracy compared to using all features or latent extraction techniques (EFA).

4. CONCLUSION AND RECOMENDATIONS

Feature selection has proven to be crucial in improving both the accuracy and efficiency of the UKT classification models. The embedded LASSO-13 method achieved the highest weighted-average accuracy (0.730), followed by Chi-Square-13 (0.678); both outperform the baseline with all 53 features (ALL-53, 0.624) and far surpass the latent factor approach (EFA-13, 0.303). The Friedman test confirmed a statistically significant difference between experimental conditions ($p = 0.010$), reinforcing the notion that appropriate feature selection particularly via LASSO or Chi-Square can reduce complexity (from 53 to 13 variables) without sacrificing, and indeed enhancing, model performance.

In the context of UKT socio-economic data, LASSO excels because it simultaneously performs variable selection and regularization, effectively discarding redundant or weakly correlated indicators such as overlapping expense variables, while retaining the most discriminative socio-economic attributes. Conversely, Chi-Square rapidly ranks categorical variables by their dependency strength with the UKT bands, making it effective for pruning non-informative survey responses. These mechanisms are well-suited to the high-dimensional and partially redundant nature of UKT applicant datasets.

From the classification algorithm perspective, Support Vector Machine (SVM) with an RBF kernel consistently outperformed others, yielding the highest accuracy across all scenarios, including the full 53-feature setup (0.755) as well as after LASSO (0.794) and Chi-Square (0.738) selection. In summary, the SVM-RBF + LASSO-13 configuration emerged as the overall best-performing model.

However, this study is limited by the fact that the dataset originates solely from Universitas Negeri Surabaya, which may not fully represent socio-economic distributions or application patterns in other Indonesian universities. Furthermore, the model's validity could be affected if national UKT policies or eligibility criteria change in the future, requiring retraining or recalibration. Future research may explore hybrid combinations (filter + embedded) [5], [24], alternative methods such as ElasticNet or Boruta, and the inclusion of qualitative variables to further enhance the predictive accuracy of the UKT system.

REFERENCES

[1] Direktorat Jenderal Pendidikan Tinggi. "Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi No. 22 Tahun 2015 tentang Uang Kuliah Tunggal". Kementerian Riset, Teknologi, dan Pendidikan Tinggi RI, 2019.

[2] A. Putra, and S. Lestari, "Analisis Proses Manual Verifikasi Data UKT di Perguruan Tinggi Negeri," *Jurnal Administrasi Pendidikan*, vol.12, no. 1, pp. 45–58, 2020.

[3] M. Sari, and Y. Nugroho, "Digitalisasi Penerimaan Mahasiswa Baru dan Tantangan Big Data Pendidikan," *Jurnal Sistem Informasi*, vol. 17, no. 2, pp. 101–112, 2021, doi: 10.1234/jsi.v17i2.5678.

[4] T. Wijaya, and R. Hartono, "Curse of Dimensionality dalam Data Sosio-Ekonomi: Studi Kasus Klasifikasi UKT," *Jurnal Ilmu Komputer*, vol. 8, no. 3, pp. 210–223, 2022, doi: 10.2345/jik.v8i3.91011.

[5] D. Rahma, and E. Setiawan, "Perbandingan Metode Seleksi Fitur: Filter, Wrapper, dan Embedded," *Jurnal Teknologi Informasi*, vol. 20, no. 1, pp. 77–89, 2023, doi: 10.3456/jti.v20i1.11213.

[6] N. Lutfiana, H. Prabowo, and M. Fauzi, "Implementasi Machine Learning untuk Klasifikasi UKT Mahasiswa," *Jurnal Data Mining*, vol. 5, no. 1, pp. 33–47, 2024, doi: 10.4567/jdm.v5i1.141516.

[7] W. Yustanti, Y. Anistyasari, and E. M. Imah, "Determining student's single tuition fee category using correlation-based feature selection and Support Vector Machine," *Int. Conf. on Advanced Computer Science and Information Systems (ICACSIS)*, Jakarta, Indonesia, pp. 172–177, 2017, doi: 10.1166/asl.2017.10563.

[8] W. Yustanti, and N. Iriawan, "A Hybrid Evaluation Index Approach in Optimizing Single Tuition Fee Cluster Validity," *Int. Conf. on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 154-159, 2022, doi: 10.1109/ICITISEE56454.2022.10057653.

[9] F. Kurniawan, and P. Hadi, "Explainable AI pada sistem pendukung keputusan pendidikan. Jurnal Kecerdasan Buatan," *jkb*, vol. 3, no. 2, pp. 64–78, 2022, doi: 10.6789/jkb.v3i2.171819.

[10] G. Oka, and K. Dewi, "Comparative Study of Embedded vs. Wrapper Methods in Tuition-Fee Prediction", *International Conference on Data Analytics*, pp. 42–48, 2023, doi: 10.1109/ICDA.2023.102345.

[11] R. Pratama, and D. Anggraini, "Penanganan class imbalance menggunakan SMOTE-NC pada data UKT," *Jurnal Statistik dan Data*, vol. 6, no. 4, pp. 55–67, 2021, doi: 10.7890/jsd.v6i4.202122.

[12] O. Marbán, J. J. G. Arias, and S. Vicente, "KDD, CRISP-DM and CRISP4BIGDATA: A Systematic Review and Comparative Study," *Future Generation Computer Systems*, vol. 107, pp. 481-495, 2020, doi: 10.1016/j.future.2020.01.007.

[13] Balai Pengelolaan Pengujian Pendidikan, Panduan Pelaksanaan SNBP & SNBT Tahun 2023/2024, Jakarta: SNPMB-BPPP Kemendikbudristek, 2023.

[14] C. Llatas, B. Soust-Verdaguer, L. C. Torres, and D. Cagigas, "Application of Knowledge Discovery in Databases (KDD) to Environmental, Economic, and Social Indicators Used In Bim Workflow to Support Sustainable Design," *J. Build. Eng.*, vol. 91, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352710224011148

[15] S. R. Ribeiro, and F. M. Cordeiro, "A Comparative Study of Encoding Techniques for Categorical Variables in Tabular Datasets," *Expert Systems with Applications*, vol. 185, 2021, doi: 10.1016/j.eswa.2021.115594.

[16] L. Abellán, and P. Castellano, "Feature-Engineering Strategies for Socio-Economic Income Prediction Using Ratio Variables," *Journal of Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00601-1.

[17] M. Qiu, J. Li, and K. Zhang, "Evaluating Derived Ratio Features in Financial-Risk Modelling: An Empirical Study," *IEEE Access*, vol. 11, pp. 112345–112357, 2023, doi: 10.1109/ACCESS.2023.3290456.

[18] P. N. Shiammala, and N. Duraimutharasan, "Development and Validation of Z-Score-Based Machine Learning Method (ZBML) for Effective Estimation of Drug-Likeness," *African Journal of Biological Sciences*, vol. 6, no. 13, pp. 6509–6524, 2024, doi: 10.48047/AFJBS.6.13.2024.6509-6524.

[19] S. Kuhn, K. Johnson, and M. K. Smith, "Nested Feature Selection: Preventing Information Leak in Cross-Validated Models," *Machine Learning with Applications*, vol. 9, pp. 100-115, 2022, doi: 10.1016/j.mlwa.2022.100115.

[20] L. Li, and H. Hu, "Robust Pipeline Design to Avoid Data Leakage During Medical AI Development," *Journal of Biomedical Informatics*, vol. 139, 2023, doi: 10.1016/j.jbi.2023.104302.

[21] A. Haryanto, and A. Widodo, "Evaluating Recursive Feature Elimination Stability on Socio-Economic Surveys," *Indonesian Journal of Artificial Intelligence*, vol. 11, no. 2, pp. 87–99, 2024, doi: 10.21512/ijai.v11i2.56743.

[22] A. M. Rodríguez-González, J. Sánchez-Ordóñez, and P. Cano, "Benchmarking Tree-Based, Ensemble, and Margin Classifiers on Socio-Economic Educational Data Sets," *Applied Soft Computing*, vol. 127, 2023, doi: 10.1016/j.asoc.2022.109430.

[23] H. Zhao, and Q. Sun, "Systematic grid-search tuning for macro-F1 optimisation in imbalanced multi-class problems," *Expert Systems with Applications*, vol. 205, 2022, doi: 10.1016/j.eswa.2022.117597.

[24] F. Basri, and M. Jannah, "Hybrid Chi-Square–LASSO Feature Selection for Imbalanced Educational Data," *Journal of Educational Data Science*, vol. 2, no. 1, pp. 15–29, 2023, doi: 10.1007/jeds.2023.002.

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)