Computer Engineering and Applications Vol. 10, No. 2, June 2021

# Application of the Relief-f Algorithm for Feature Selection in the Prediction of the Relevance Education Background with the Graduate Employment of the Universitas Sriwijaya

Sugandi Yahdin[1], Anita Desiani[2*], Nuni Gofar[3], Kerenila Agustin[4], Desty Rodiah[5]

[1,2,4]*Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Sriwijaya*
[4]*Department of Soil Science, Faculty of Agriculture, Universitas Sriwijaya*
[5]*Department of Computer Informatics, Faculty of Computer Science, Universitas Sriwijaya*
**anita_desiani@unsri.ac.id*

## ABSTRACT

Career Development Center (CDC) at Universitas Sriwijaya provided a tracer study dataset for graduates. The data contained feature questions about the relevance of background education and graduate employment, namely about lectures, research projects experience, internships experience, English skill, internet knowledge, computer skill and others. the data was filled in by graduates in 2014, 2015, and 2016. Applying the Relief-f algorithm was to select the pattern features that most influence the relevance of education background and graduate employment. This study used Naive Bayes and KNN methods to measure the success rate of the Relief-f algorithm. The results of the accuracy of the data before the feature selection process for the naïve Bayes method were 73.43% and the KNN method was 66.24%, after the feature selection process the accuracy obtained in both methods increased to 74.38% for the Naive Bayes method and 72.22% for the KNN method. The best pattern features selected were 8 features: department relationship with work, the competence of education background, English skill, research projects experience, extracurricular activities, the competence of education background, internships experience, and communication skills. Based on the accuracy obtained, it was concluded that the Relief-f algorithm worked well in the feature selection and improved the accuracy.

**Keywords**: Relief-f, CDC, Universitas Sriwijaya, graduate employment, tracer study

## 1. INTRODUCTION

Feature selection is one of the data pre-processing, the process of selecting a subset of an important attribute using certain criteria [1]. Feature selection is an effective way of data reduction and is usually used to reduce data with many dimensions by removing features that have no relevance to the dataset, so as to save memory and time used [2]. There are several types of algorithms in feature selection, one of which is Relief-f, the algorithm was developed from the Relief algorithm [3]. The Relief algorithm has disadvantages, namely that it cannot handle incomplete data (incomplete data) and is limited to 2 classes [4]. The Relief-f algorithm can handle multiclass datasets (more than 2 classes) and incomplete data (incomplete data) and can handle internal data. discrete or continuous form [5].

The Relief-f feature selection algorithm is widely used in various research fields. Wang, Sanin & Szczerbicki [6] used feature selection to evaluate the quality of features and is used for architectural recommendations to improve prediction based on Decisional DNA (DDNA). Xie et al [7] used the Relief-f feature selection algorithm to balance stroke data. Tahir & Loo [8] used the Relief-f algorithm Selection feature to reduce complexity by ranking the feature selection and in the result Relief-F can reduce the training time of classification for all datasets by 52.14%. The performance of the Relief-f feature selection algorithm has been used by Sharma & Dey [9] to analyze the accuracy of using feature selection using the Relief-F algorithm. Sun et al [10] used Relief-f in feature selection to reduce computational complexity for classification in multi-label cases. The use of Relief-f in Deepika & Sathyanarayana [11] was reduced high dimensionality and dealed with data uncertainty. Zaffar, Hashmani & Savita [12] conducted comparisons of several feature selection algorithms, one of which was Relief-F by applying it to a student dataset.

The CDC was formed to respond to the Low achievement of graduate tracking points against AIPT forms. The CDC provided an online questionnaire for tracer studies on its website (http//:cdc.unsri.ac.id). This questionnaire consisted of 17 questions that refer to the DIKTI standard. The questionnaire had questions about the assessment of the conditions and regulations of learning in Unsri by graduates. The results of this tracer study are useful in providing input to the campus to improve services and facilities as well as the quality of existing learning so that the quality of graduates can be improved [13]. The dataset available in The CDC had a large number of attributes or various supporting features. To find out what features could support to improve the quality of graduates was looked for what the relevance of the educational background and graduate employment at Universitas Sriwijaya. One of the algorithms used to determine the most influential features on a dataset was Relief-f algorithm. The Relief-f algorithm was used in this study to select the features that have the most influence on predicting the relevance between educational backgrounds with Universitas Sriwijaya graduates employment.

## 2. MATERIAL AND METHODS

### 2.1. DATA COLLECTION

The data used in this study were secondary data in Career Development Center (CDC) of Universitas Sriwijaya. The data was collected in 2017 and 2018, the data was filled in by graduates in 2014, 2015 and 2016 with 1143 existing data and 17 existing features. This study used 14 features because the other 3 features had more than 50% missing data. The selection of features was based on what features were available on the CDC form as well as the features or factors that support the relevance between education background with graduate employment of Universitas Sriwijaya. There were 14 features used, of which 13 were supporting features and 1 feature was the target or label feature. The following in Table 1 explained the description of each feature used in this study.

TABLE 1.
Features Description in CDC Tracer Study Data Set of Universitas Sriwijaya

| No. | Feature | Annotatiom | Data Type | Missing Data |
|---|---|---|---|---|
| 1. | F1(Learning level in lectures) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 2. | F2(The research experience) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 3. | F3(The internship experience) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 4. | F4 (Learning level in practicum) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 5. | F5 (Level of learning in field work) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 6. | F6 (The communication skill) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 7. | F7 (Time to get work) | month | Continu | None |
| 8. | F8 (The competence of education background) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 9. | F9 (The extracurricular activities ) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 10. | F10 (English skill) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 11. | F11 (Internet knowledge ) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 12. | F12 (Computer skill) | 1: Excellent 2: Very Good 3: Good 4: Low 5: Poor | Category | None |
| 13. | F12 (Department relationship with work) | 1: Thightest 2: Very Tight 3: Tight 4: Somewhat Tight 5: Not Tight | Category | 124 |
| 14. | F14 (The relevance education background with graduate employment) | 1: High 2: Same 3: Lower 4: No Relation | Category (as a feature class) | 124 |

The features in table 1 were divided into two part. The F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11, F12 and F13 features were pattern features and the f14 feature is a class that has 5 labels. The Universitas Sriwijaya had 10 faculties, namely the Faculty of Computer Science with 74 graduates data, the Faculty of Public Health with 9 graduates data, the Faculty of Medicine with 39 graduates data, the Faculty of Economics with 131 data, the Faculty of Engineering with 196 graduates data, the Faculty of Mathematics and Natural Sciences as many as 104 graduates data, the Faculty of ISIP as many as 51 data, Faculty of Law with 64 graduates data, Faculty of Agriculture with 116 graduates data, Faculty of KIP with 359 graduates data. In Table 1, in F13 and F14 there were 124 blank data and there was no information about graduates employment, so that the data was removed and the total data used in this study was 1019 graduates data.

## 2.2. APPLICATION OF THE RELIEF-F FEATURE SELECTION ALGORITHM

The steps in applying the Relief-f algorithm were as follows:

1. Calculate the probability of each class.

The available data is calculated the probability of each class occurring in the data. Relief-f estimates W [A] from feature A by estimating the difference in the probability of:

$$W[A] = P(X \mid Y) - P(X \mid Z) \tag{1}$$

where $X$ is difference value on A, $Y$ is closest instance of different class, and $Z$ is closest instance of the same class. The probability $W[A]$ was be used in calculating the weight of the features in Relief-f algorithm.

2. Determine K-Near hit and K-Near miss values.

The diff function in the Relief-f algorithm was used to calculate the difference in the value of feature A between instances I1 and I2, where I1 = Ri and I2 are H or M, when performing weight updates. For discrete features (e.g. categorical or nominal), the diff function was defined as [14]:

$$diff\left(A, I_1, I_2\right) = \begin{cases} 0; if\left(A, I_1\right) = value\left(A, I_2\right) \\ 1; else \end{cases} \tag{2}$$

And for continuous features (e.gordninal or numeric), diff is defined as:

$$diff\left(A, I_1, I_2\right) = \frac{\left|value\left(A, I_1\right) - value\left(A, I_2\right)\right|}{\max(A) - \min(A)} \tag{3}$$

3. Perform weight calculations for each feature.

$$W[A] = W[A] - \sum_{j=1}^{k} \frac{diff\left(A, R_i, H_j\right)}{m.k} + \sum_{C \neq class\left(R_i\right)} \frac{\left[\frac{P(C)}{1 - P(class(R_i))} \sum_{j=1}^{k} diff\left(A, R_i, M_j(C)\right)\right]}{m \cdot k} \tag{4}$$

Feature weight is $W[A]$= feature weight '$A$'. $W[A]$ had ranges from -1(worst) to +1(best) [15]. The Relief-f algorithm performed a weight calculation cycle through random training instances$(R_i)$, where m was a parameter determined by the user [3]. The $R_i$ instant was the 'target' instant and the weight $W$ was updated based on the difference in the observed feature value between the 'target' instant and all other instances calculated.

4. The results of the weight of each feature were sorted from largest to smallest (rank).
5. The feature that has the smallest weight was eliminated gradually and the result of the accuracy of each elimination was compared.
6. Determine which features have the most influence on predicting the relevance between education backgrounds with graduates employment, and draw conclusions from the comparison of the results obtained for accuracy.

### 2.3. EVALUATION OF THE FEATURE SELECTION METHOD

The evaluation process uses the K-NN and naïve Bayes methods for prediction the relevance between education background and graduate employment. The KNN method was chosen because Relief-f also uses the k-nearest neighbor technique to search for instant proximity to each other. The application both of methods was to see the difference in the prediction results before and after when the selection feature was carried out. Meanwhile, Naive Bayes was used as a comparison of the prediction results from the KNN. The result of the Naive Bayes was measured whether the performance also increased after the selection feature was carried out. The results of the two prediction methods could give conclusion how well the feature selection method performs when applied to the prediction problem. the performances measured and used in this study were [16]:

1. Accuracy (acc), the accuracy value was defined as the percentage of accuracy of data that is classified correctly after testing the classification results.

$$acc = \frac{(TP + TN)}{(TP + FN + FP + TN)} \times 100\% \qquad (5)$$

2. Precision can be interpreted as a match between requests for information and answers to requests.

$$precision = \frac{TP}{(FP + TP)} \times 100\% \qquad (6)$$

3. Recall is defined as the ratio of selected relevant items to the total number of relevant items available.

$$recall = \frac{TP}{(TP + FN)} \times 100\% \qquad (7)$$

## 3. RESULT AND DISCUSSION

### 3.1. APPLICATION OF SELECTION ALGORITHM SELECTION RELIEF-F

In the feature selection process, the Relief-f algorithm calculated the weight of each pattern feature in the dataset on the relevance between education background and graduate employment (F14). The resulting weight for each feature was sorted based on the value of the weight of the features from the largest to the best. The feature selection process reduced the dimensionson the relevance of education background and graduate employmentbased on the weight value for each feature. Table 2 showed the sequence of pattern features from the largest to the distance calculated by the weight of the Relief-f algorithm.

Table 2 showed the weight for each feature, where the feature that had the greatest weight was the optimal feature, namely F13, F8, then folLowed by other features, namely F10, F2, F9, F5, F3, F6, F12, F11, F4, F1, and F7.

TABLE 2.
Weight Value of Each Pattern Feature with the Relief-f Algorithm in the Prediction of the Relevance Education Background with the Graduate Employment

| No. | Weights | Annotation |
|---|---|---|
| 1. | 0,08483 | F13 (Department relationship with work) |
| 2. | 0,07768 | F8 (The competence of education background) |
| 3. | 0,07704 | F10 (English skill) |
| 4. | 0,06999 | F2 (The research experience) |
| 5. | 0,06952 | F9 (The extracurricular activities) |
| 6. | 0,06052 | F5 (Level of learning in field work) |
| 7. | 0,04967 | F3 (The internship experience) |
| 8. | 0,04844 | F6 (The communication skill) |
| 9. | 0,04341 | F12 (Computer skill) |
| 10. | 0,03967 | F11 (Internet knowledge) |
| 11. | 0,03613 | F4 (Learning level in practicum) |
| 12. | 0,03364 | F1 (Learning level in lectures) |
| 13. | 0,00291 | F7 (Time to get work) |

## 3.2. EVALUATION OF FEATURE SELECTION RESULTS

The values in Table 2 were used as the basis for selecting the best features used as training data and test data by looking at rank and high levels of accuracy by removing features that were less influential in predicting the relevance of education background and graduate employment. To measure how the Relief-F made an impact in predicting, this study used the Naive Bayes and KNN methods. The dataset was divided into two group training and testing data. Theaccuracy results of the training and testing data were shown in Table 3.

TABLE 3.
Comparison Accuracy Results of Naïve Bayes and KKN in Feature Selection in the Prediction of the Relevance Education Background with the Graduate Employment

| No. | N Feature | Used Feature | Ommited Feature | Accuracy | |
|---|---|---|---|---|---|
| | | | | Naïve Bayes | KNN |
| 1. | 13 | F13,F8,F10,F2,F9,F5,F3, F6,F12,F11,F4,F1,F7 | - | 73,4053% | 66,24% |
| 2. | 12 | F13,F8,F10,F2,F9,F5,F3, F6,F12, F11,F4,F1 | F7 | 73,896% | 70,16% |
| 3. | 11 | F13,F8,F10,F2,F9,F5,F3, F6,F12,F11,F4 | F7,F1 | 73,7978% | 71,14% |
| 4. | 10 | F13,F8,F10,F2,F9,F5,F3, F6,F12,F11 | F7,F1,F4 | 73,7978% | 71,44% |
| 5. | 9 | F13,F8,F10,F2,F9,F5,F3, F6, F12 | F7,F1,F4,F11 | 74,1904% | 72,03% |
| 6. | 8 | F13,F8,F10,F2,F9,F5,F3, F6 | F7,F1,F4,F11,F12 | 74,3867% | 72,22% |
| 7. | 7 | F13,F8,F10,F2,F9,F5,F3 | F7,F1,F4,F11,F12,F6 | 74,2885% | 71,44% |

Table 3 showed the accuracy comparison obtained from the prediction results using naive bayes and KNN. It could be seen that the accuracy result of the experiments number 1 to 6 has increased but the experiment number 7, the accuracy has decreased. The 6th experiment showed the best accuracy results using 8 features. There were 8 selected features, namely F13, F8, F10, F2, F9, F5,  F3. The results of applying the relief-f algorithm to naïve Bayes and KNN could also be measured based on the precision and recall result obtained. The precision and recall were calculated based on the labels in the feature class. The feature class had 4 classes namely higher (1), same (2), lower (3), and no relation (4). Percentage of performance results from the Naïve Bayes method that uses the Relief-f algorithm and without using the Relief-f algorithm were shown in Figure 1.

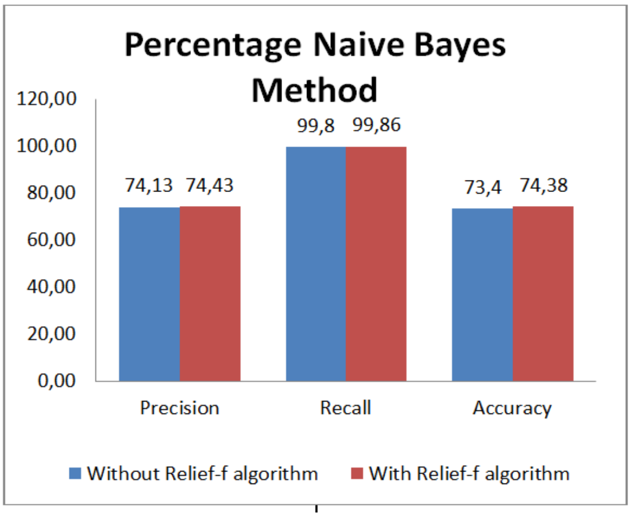

FIGURE 1. Percentage of Naïve Bayes Method for The Relevance Education Background and Graduate Employment

Percentage of performance results from the KNN method that uses the Relief-f algorithm and without using the Relief-f algorithm were shown in Figure 2.

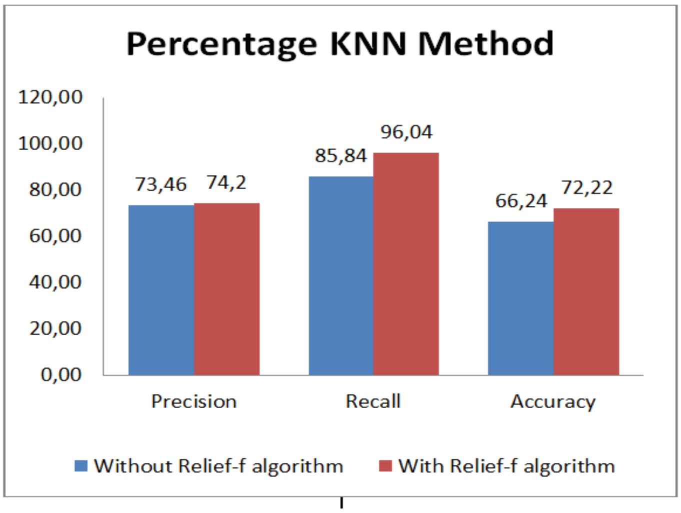

FIGURE 2. Percentage KNN Method The Relevance Education Background and Graduate Employment

Figure 1 and Figure 2 showed that the process of feature selection with the Relief-f algorithm by reducing features that were less influential on predictions could increase the accuracy compare if it did not use Relief-f algorithm. The prediction that using the Relief-f feature selection algorithm were selected as many as 8 best features and obtained predictions with an accuracy value of 74.38% by the Naive Bayes method and 72.22% by KNN. KNN gave the highest accuracy with Relief-f algorithm.

In Figures 1 and 2, the highest precision and recall for Naive Bayes were obtained by label 2 (same level). It meant that the relationship between educational background and graduate work was appropriate. The precision and recall obtained after the selection feature did not show a significant increase. It showed that the removal of some features did not really matter for predictions using Naive Bayes. In the KNN method, for precision and recall results before the application of feature selection, the highest precision and recall were also obtained by label 2. After the feature selection process, the precision on predicting increased by 5.74%, while the precision increased by 10.2 %. It can be concluded that most of the Sriwijaya University graduates work in accordance with their educational background.

Based on the result, the Relief-falgorithm can improve accuracy by selecting features and removing features that are less influential based on rank or feature weight sequence. The Relief-f algorithm has provided the selected 8 best featurse for predicting the relevance of education background and graduate employment. The featureswere F13 (Department relationship with work), F8 (The competence of education background), F10 (English skill), F2 (The research experience), F9 (The extracurricular activities) , F5 (Level of learning in field work), F3 (The intership experience) and F6 (The communication skill). To analyze the results of this study further, the results obtained were compared with several previous studies which can be seen in Table 4.

TABLE 4.
Comparison of Research Results The proposed method with previous research

| Author | Dataset | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|
| Tahir & Loo [8] | Pakistani Food dataset | 76,04 | 76,05 | 76,05 |
| Sharma & Dey [9] | Internet Movie Reviews | 68% | - | - |
| Deepika & Sathyanarayana [11] | Academic Student dataset | 64,65% | 77 | 65 |
| Zaffar et.al [12] | Academic Student Dataset | 70,83 | - | - |
| Proposed method with Naïve Bayes | Graduate Employment Universitas Sriwijaya | 74,38 | 74.43 | 99.86 |
| Proposes Method with KNN | | 72.22 | 74.20 | 96.04 |

Table 4 was consisted of several comparisons of research. Tahir & Loo [8] had fairly good accuracy, precision and recall values however, the study was unable to show differences in results before and after using Relief-f algorithm. It also obtained the highest accuracy value. The highest precision was obtained by Deepika & Sathyanarayana [11]. Even though the accuracy of the proposed method was not more than 75%, the two methods used in the proposed method had better results than the research by [9], [12]. The highest recall value was given by the two methods used in the proposed method. The recall result of both naïve Bayes and KNN was above 90%. The accuracy results of Sharma & Dey [9] was still not good enough and in this study it only displayed accuracy values but could not display the others. As well as research by Zaffar, Hashmani & Savita [12], it only displayed the accuracy result. Deepika & Sathyanarayana [11] had a fairly good accuracy and recall even though the precision was low. These results indicated that the use of the Relief-f algorithm to select features on the relevance of educational background and graduates employment provided significant results compared to some previous studies.

## 4. CONCLUSION

The results of the application of the Relief-f feature selection algorithm obtained the most influential features on the prediction of the relevance education background with graduates employment at Universitas Sriwijaya are 8 features, namely F13 (Department relationship with work), F8 (The competence of education background), F10 (English skill), F2 (The research experience), F9 (The extracurricular activities), F5 (Level of learning in field work), F3 (The internship experience), and F6 (The communication skill). This is indicated by the accuracy value obtained from the prediction using the Naive Bayes method and the KNN method has increased before using Relief-f and after using Relief-f. The increase in the accuracy value obtained shows that the Relief-f algorithm can choose which features are the most influential and good enough to increase the performance in predicting the the relevance of educational background and graduates employment at Universitas Sriwijaya.

## REFERENCES

[1] T. Zar Phyu and N. N. Oo, “Performance comparison of feature selection methods,” *MATEC Web of Conferences*, vol. 42, pp. 2–5, 2016, doi: 10.1051/matecconf20164206002.

[2] R. P. L. Durgabai, “Feature Selection using ReliefF Algorithm,” *Ijarcce*, vol. 3, no. 10, pp. 8215–8218, 2014, doi: 10.17148/ijarcce.2014.31031.

[3] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and H. Moore, “Relief-Based Feature Selection : Introduction and Review,” *Journal of Biomedical Informatics*, 2018.

[4] S. Gore and V. Govindaraju, “Feature Selection Using Cooperative Game Theory and Relief Algorithm,” in *Knowledge, Information and Creativity Support Systems: Recent Trends, Advances and Solutions*, 2016, pp. 401–412.

[5] Y. He, J. Zhou, Y. Lin, and T. Zhu, “A class imbalance-aware Relief algorithm for the classification of tumors using microarray gene expression data,” *Computational Biology and Chemistry*, vol. 80, no. March, pp. 121–127, 2019, doi: 10.1016/j.compbiolchem.2019.03.017.

[6] P. Wang, C. Sanín, and E. Szczerbicki, "Prediction based on integration of decisional dna and a feature selection algorithm relief-F," *Cybernetics and Systems*, vol. 44, no. 2–3, pp. 173–183, 2013, doi: 10.1080/01969722.2013.762246.

[7] Y. Xie, D. Li, D. Zhang, and H. Shuang, "An improved multi-label relief feature selection algorithm for unbalanced datasets," *Advances in Intelligent Systems and Computing*, vol. 686, pp. 141–151, 2018, doi: 10.1007/978-3-319-69096-4_21.

[8] G. A. Tahir and C. K. Loo, "An open-ended continual learning for food recognition using class incremental extreme learning machines," *IEEE Access*, vol. 8, pp. 82328–82346, 2020, doi: 10.1109/ACCESS.2020.2991810.

[9] A. Sharma and S. Dey, "A Comparative Study of Feature Selection and Machine Learning Techniques for Sentiment Analysis," *RACS*, pp. 1–7, 2012.

[10] L. Sun, T. Yin, W. Ding, Y. Qian, and J. Xu, "Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems," *Information Sciences*, vol. 537, pp. 401–424, 2020, doi: 10.1016/j.ins.2020.05.102.

[11] K. Deepika and N. Sathyanarayana, "Relief-F and budget tree random forest based feature selection for student academic performance prediction," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, pp. 30–39, 2019, doi: 10.22266/IJIES2019.0228.04.

[12] M. Zaffar, M. A. Hashmani, and K. S. Savita, *Comparing the performance of FCBF, Chi-Square and relief-F filter feature selection algorithms in educational data mining*, vol. 843. Springer International Publishing, 2019.

[13] N. Gofar and P. Susmanto, *Tracer Study Universitas Sriwijaya Tahun 2018 (Lulusan Tahun 2016)*. Palembang: Noer Fikri, 2018.

[14] K. Kira and L. A. Rendell, "A Practical Approach to Feature Selection," in *9th International Workshop on Machine Intelligence, Aberdeen*, 1992, pp. 249–256.

[15] I. Kononenko, "Estimating Attributes : Analysis and Extensions of Relief," in *European Conference on Machine Learning*, 1994, pp. 171–182.

[16] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," *Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*, vol. 2018-Janua, pp. 294–298, 2018, doi: 10.1109/ICITISEE.2017.8285514.