Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 911−920
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5655

Predicting Employability of University Graduates Using Support Vector Machine Classification

Muhamad Fachri Haikal[1], Irma Palupi[2*]

[1,2] School of Computing, Informatics, Telkom University, Bandung, Indonesia
Email: [1]haikalfachri@student.telkomuniversity.ac.id, [2,*]irmapalupi@telkomuniversity.ac.id,
Correspondence Author Email: irmapalupi@telkomuniversity.ac.id
Submitted: **22/07/2024**; Accepted: **09/09/2024**; Published**: 09/09/2024**

Abstract− Ensuring graduates' smooth transition into the job market is crucial as competition rises with increasing graduate numbers. This research addresses predicting employability, focusing on Telkom University students' initial job income. Using a dataset of 6089 Telkom University 2022 alumni, split 80:20 for training and testing, the study utilized Support Vector Machine (SVM) for data analysis due to the limitations of traditional linear regression in handling potential non-linearity in the data. Feature manipulation techniques like Principal Component Analysis, Spearman's rank correlation, and the Chi-square test of independence were applied, followed by SMOTE-ENN to tackle data imbalance. The SVM model, with Randomized Search hyperparameter tuning and analyzed through Permutation Feature Importance, identified key employability factors. The enhanced SVM model, utilizing SMOTE-ENN, Spearman's rank correlation for feature selection, and randomized search, achieved precision, recall, f1-score, and accuracy of approximately 0.70, 0.73, 0.71, and 0.73, respectively. Competency features such as ethics, English skills, IT skills, and knowledge emerged as the most influential factors.

Keywords: feature manipulation; imbalanced dataset; tracer study analysis; student employability; SVM

1. INTRODUCTION

In the era of dynamic job markets and rapid technological advances, predicting graduates' absorption into the job market is crucial for higher education institutions. With increasing competition due to a rising number of graduates each year, the level of graduate absorption serves as a key indicator of the success and quality of these institutions [1], [2], [3]. Therefore, higher education institutions must focus on enhancing the employability of their graduates by preparing them effectively for the world of work. Various factors can affect the employability of graduates, whether because of their competency and educational curriculum, such as soft skills, hard skills, or emphasis on course knowledge [4], [5]. To address this problem, it is essential for these institutions to understand the factors influencing graduates employability and to leverage academic data along with machine learning techniques to make accurate predictions [6].

Academic data is usually utilized in tracer study analysis. Tracer study examines high-performing institutions to document work outcomes, workplace shifts, and satisfaction levels with university services, learning environments, and facilities. It assesses the effectiveness and relevance of undergraduate programs in the face of rapid developments [7]. Tracer studies aim to evaluate the impact of educational programs, improve educational content, facilitate graduates' transition to the job market, and ensure the relevance of curricula to industry needs. This study provides feedback for curriculum development, improves learning conditions, and increases the attractiveness of educational programs. Higher education institutions use data from these studies to assess educational relevance, support accreditation, and create complete profiles of graduates [8].

Tracer study analysis has traditionally relied on simple regression methods, such as linear regression, despite the lack of empirical evidence supporting their effectiveness. In student employability profiling, identifying the key competencies that directly influence employability is a crucial aspect of the analysis. Machine learning has potential to accurately predict graduates' competitiveness, identify hidden patterns, and process big data quickly [9]. Therefore, the use of machine learning is well-suited and necessary for this purpose. It can also be used to understand what factors influence graduate competitiveness. For example, machine learning can link academic performance and extracurricular skills with graduate employability so institutions can improve their programs and curricula [2].

Several studies have examined graduate employability through tracer studies. One study focused on Information Technology graduates from a state university in the Philippines, finding that 78.53% of the graduates were employed, with most considering their jobs related to their college program. The study emphasized the need for regular curriculum reviews to keep programs relevant to industry needs [10]. Another study analyzed data from over 100.000 graduates to map their soft skills and employment status, using a predictive model with 77% accuracy. It found that employability is influenced by factors like gender, family income, field of study, CGPA, internships, and key soft skills such as communication, analytical skills, and teamwork. These studies highlight the importance of both technical and soft skills in securing employment [11].

At Telkom University, several studies have used machine learning in tracer studies. For example, research using Logistic Regression shows that educational suitability and first-job waiting time are highly correlated by more than 90%, and it shows that college knowledge, technical skills, and communication are the three most important competencies in the professional careers of Telkom University alums. They also applied the SMOTE-ENN resampling technique to handle data imbalance [4]. Another study used an Artificial Neural Network (ANN) to predict graduate

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 911−920
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5655

performance based on waiting time for the first job with an accuracy of up to 87%, which increased after using the SMOTE technique [12].

This research uses the Support Vector Machine (SVM) model because previous research shows that SVM has the highest prediction accuracy compared to other methods such as Naïve Bayes and Decision Tree [9]. The other research showed SVM has better accuracy score than Logistic Regression [13].This research offers a different contribution by predicting the employability of graduates based on income, and it is hoped that it can provide additional understanding of the factors that influence the career success of Telkom University alums. The developed model involves feature manipulation and parameter-tuning techniques. This research also uses resampling techniques to handle data imbalance. In addition, the method used can also identify the factors that influence income levels and evaluate the performance of model predictions on graduate employability.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This section outlines the methodology employed in this research. Figure 1 shows the methodology used for the results of this research. The stages included in this research are data collecting, data preprocessing, feature manipulation, data splitting, data resampling, hyperparameter tuning, machine learning modelling, model evaluation, and determining feature importance.

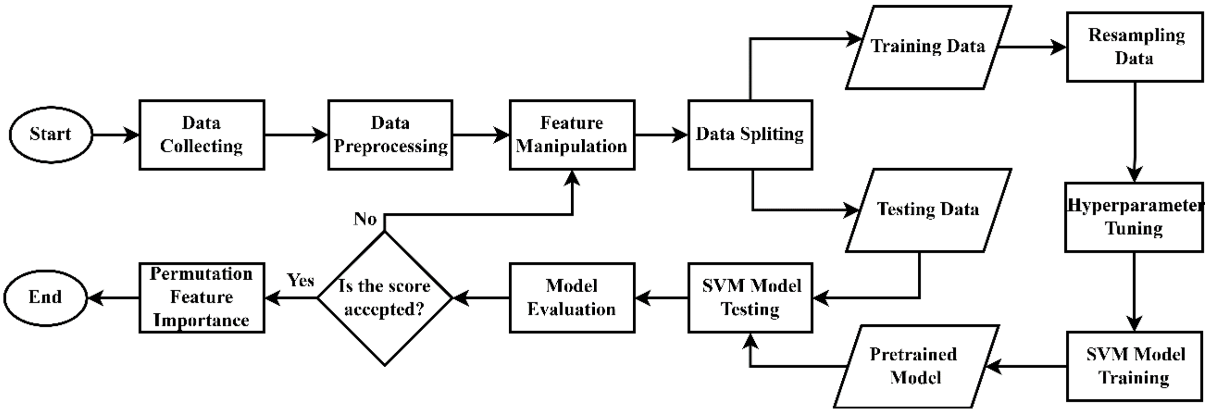

Figure 1. Research Methods Flow

The process begins with gathering relevant data, followed by preprocessing to clean and prepare the data for analysis. A feature manipulation schema is then performed to enhance the dataset, which is split into training and testing sets. The training data undergoes resampling to address imbalances and hyperparameter tuning is conducted to optimize the model's performance. After training, the machine learning model is tested and evaluated to ensure its performance scores. Finally, feature importance is determined to identify the most significant factors contributing to the model's predictions.

2.2 Data Collection

The dataset used in this research is the 2022 Telkom University alums tracer dataset. This dataset comprises 26 features and 6089 data points. Table 1 lists the features available in the dataset along with their descriptions.

Table 1. Dataset Features

Feature	Description	Feature	Description
Graduation Year	Student's graduation year	Communication*	Communication competency
Study Program	Student's study program	Colaboration*	Teamwork competency
Faculty	Student's faculty	Work-educ Rel	The relationship between work and education background
Fund Source	Sources of college funding	Practicum*	Emphasis on practical methods
Waiting Time	Waiting time to get a job	Field Work*	Emphasis on field work methods
Companies Applied	Qty. company applied	Discussion*	Emphasis on discussion methods
Companies Responded	Qty. the company responded	Lecture*	Emphasis on lecture methods
Companies Interviewed	Qty. the company invites an interview	Demonstration*	Emphasis on demonstration methods
Self-dev*	Self-development competency	Research Project*	Emphasis on research project methods

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 911−920
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5655

Ethic*	Self-ethical competency	Internship*	Emphasis on apprenticeship methods
Knowledge*	Knowledge field competency	City/Regency	City/district of work
English Skills*	English language competency	Province	Province of work
IT Skills*	Information technology competency	Salary	First job income

*Competency features and curricula contribution.

These features include graduates' competencies, curriculum contributions, demographic information, and employment features. By analyzing these features, we aim to gain insights into the employability and career outcomes of the graduates.

2.3 Data Preprocessing

This stage is carried out on the dataset before being entered into the model. Missing values are handled using mode imputation techniques, except for targets that are removed to reduce bias. Outliers such as incorrect data are also deleted to avoid model mistakes. The Salary targets in the dataset are grouped into two groups as shown in figure 2. Then, an encoding label is applied to the categorical data so that the model can understand and analyze it better.

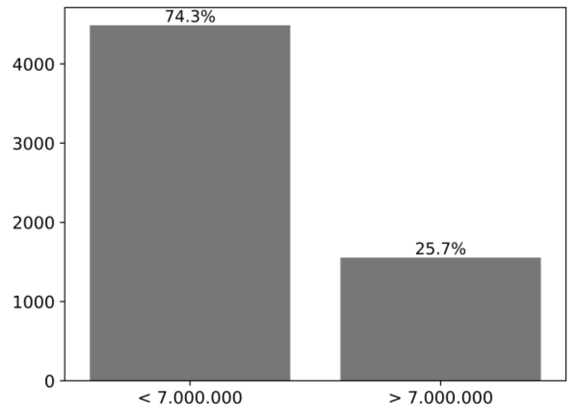

Figure 2. Salary Distribution on Dataset

Figure 2 shows the distribution of graduate's salary in the dataset, divided into two categories, those earning less than 7,000,000 and those earning more than 7,000,000. The distribution shows that 74.3% of the dataset falls into the category of earning less than 7,000,000, which has a frequency count 4487. On the other hand, 25.7% of the dataset earns more than 7,000,000, represented by the shorter bar on the right side of the chart with a frequency count 1554.

2.4 Feature Manipulation

This research used three feature manipulation schemes to prepare features before machine learning modelling: the Spearman's rank correlation coefficient, Principal Component Analysis (PCA), and the Chi-squared test of independence.

PCA as a technique for feature engineering. PCA addresses the issue of having too many features, which can lead to overfitting, by reducing the number of variables while preserving as much information as possible. PCA reduces the dimensionality of the data by projecting it onto a lower-dimensional space defined by the principal components, which capture the most significant patterns in the data while maintaining as much variance as possible [14].

Spearman's rank correlation coefficient measures the relationship between variables. This coefficient is between -1 and 1 for perfect correlation and 0 for neutrality [15]. The Spearman's rank correlation coefficient (ρ) calculation formula is defined as:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} \tag{1}$$

Where ρ is the Spearman's rank correlation value, d_i is the difference between two ranks of data pairs, dan n is the number of data pairs. Using this score in feature selection requires a threshold to define the features that are significant to the given target.

The Chi-squared test of independence is used to determine significant relationships between two categorical features in a single sample. The null hypothesis (H_0) states that the features are independent, while the alternative hypothesis (H_1) indicates that there is a significant correlation [16]. The statistical formula for the chi-square test of independence (x^2) is defined as:

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 911−920
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5655

$$x^2 = \sum[\frac{(O-E)^2}{E}] \tag{2}$$

where O is the observed frequency, E is the expected frequency if the features are independent, and the sum Σ is taken from all variable levels. The calculated chi-square statistics (x^2) and degrees of freedom (DF) are then compared with the critical values of the chi-square distribution to determine whether the null hypothesis can be rejected, indicating a significant relationship between the features.

2.5 Data Splitting

Data splitting involves splitting a dataset into a training subset and a testing subset. Data splitting was proportionate in this research, with 80% of the data for training and 20% for testing.

2.6 Data Resampling

According to the target frequency in our dataset, as shown in Figure 2, one class comprises 74.28% of the samples, and the other only 25.72%. Therefore, the SMOTE-ENN resampling technique is applied to handle data imbalance, which can cause high bias and prediction errors in the model [4].

2.7 Hyperparameter Tuning

We used Randomized Search as hyperparameter tuning in our model. Randomized Search is hyperparameter tuning that samples a subset of the search space, making it computationally lighter than Grid Search. Additionally, Random Search provides a comparable accuracy improvement to Grid Search [17]. For the SVM hyperparameter sample space used is shown in Table 2.

Table 2. Sample Space of SVM Hyperparameter

Parameter	Sample
C	0.01, 0.1, 1, 10
kernel	rbf
gamma	0.01, 0.1, 1, 10, 'scale'

The SVM model's hyperparameters include C, kernel, and gamma. The C parameter balances training error and testing error. We fixed the kernel to 'rbf' (Radial Basis Function) because it works well in high-dimensional spaces. The gamma parameter determines the reach of a single training example's influence. Using Randomized Search, we efficiently explored these hyperparameters, finding the possible best settings for improved model accuracy and generalization.

2.8 Support Vector Machine

SVM is a popular machine-learning model used for classification. We used SVM as our machine learning model. This model produces the best-dividing line (hyperplane) to separate classes in the dataset [3]. The objective of this model is to find a hyperplane that divides the data into two distinct classes. Out of all potential hyperplanes capable of achieving this division, SVM seeks the one that maximizes the margin, meaning it has the greatest distance from the classes to the separating hyperplane while minimizing misclassification errors. [18]. If data is n-dimensional, n represents number of features on the dataset, and the hyperplane will be an (n-1) vector function, which formula is represented by [14]:

$$f(x) = w^T.x_i + b = 0 \tag{3}$$

where w^T is weight vector for class T, x_i is input vector, and b is bias. In this work, the SVM method is used to explore how students' competencies acquired during their university studies relate to their performance in the workplace, or employability. We demonstrate that this model can depict this relationship more effectively than the traditional linear model typically employed.

2.9 Model Evaluation

One way to measure the performance of a classification model is to use a confusion matrix. In the SVM model, to handle binary classification problems, a confusion matrix generally represents model predictions with actual values [19]. The standard form of the confusion matrix can be seen in Table 3.

Table 3. Confusion Matrix

Classfication	Predicted	
	Positive(+)	Negative (-)
Actual	True Positive (TP)	False Negative (FP)
	False Positive (FP)	False Negative (FN)

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 911−920
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5655

In this matrix, True Positive (TP) represents positive instances correctly classified, False Negative (FN) represents positive instances incorrectly classified as negative, False Positive (FP) represents negative instances incorrectly classified as positive, and True Negative (TN) represents negative instances correctly classified. Using these values, we can calculate several performance metrics. Accuracy is the ratio of correctly predicted instances to the total instances. Recall measures the ability of the model to identify positive instances. Precision measures the accuracy of positive predictions. The f1-score is the harmonic means of precision and recall, balancing both concerns.

2.10 Permutation Feature Importance

Breiman introduced the permutation feature importance measure for Random Forest, but the procedure is model-independent and can be used for other machine-learning models. Feature Importance can be measured via the permutation method, which essentially measures the effect on model accuracy of randomly reshuffling each feature. In its approach, this method directly measures feature importance by observing how random reshuffling of each feature affects model performance model [20]. The highlighted features are competency features, and curriculum contribution features that can provide insight into developing existing education.

3. RESULT AND DISCUSSION

In this section, we present the findings from our study aimed at predicting student employability based on their initial income post-graduation. Table 4 presents a dataset description that undergoes data preprocessing.

Table 4. Dataset Profiling before Data Preprocessing

		Quantity
Number of features		26
Number of rows		6089
Feature types	Numerical	9
	Categorical	17
Missing value	Graduation Year	1283
	Salary	48
Outliers	Companies Applied	15
	Companies Interviewed	5
	Companies Responded	5
	Waiting Time	2

As per the previous method, the mode imputation technique handled missing values in the dataset while missing values were removed in target labels. Outliers were also removed from the dataset. Label encoding is applied to categorical features. The result of data preprocessing produces a dataset with 26 features and 6041 data ready for further analysis. Salary targets in the dataset are grouped into two groups, the one with a salary greater than and the other below “7.000.000”, as shown in its distribution in Figure 2. Table 5 shows the correlation of features with the target label.

Table 5. Correlation of Features to Target

Feature	Spearman	Pearson	Feature	Spearman	Pearson
Study Program	0.150	0.152	Knowledge*	0.038	0.038
Self-dev*	0.115	0.113	Companies Interviewed	0.035	0.031
Province	0.113	0.073	Internship*	0.034	0.028
Ethic*	0.111	0.108	Fund Source	0.034	0.000
IT Skills*	0.100	0.101	Demonstration*	0.031	0.028
Waiting Time	0.095	0.086	Colaboration*	0.027	0.030
City/Regency	0.069	0.027	Work-educ Rel	0.027	0.024
English Skills*	0.067	0.073	Graduation Year	0.019	0.019
Practicum*	0.055	0.049	Companies Applied	0.012	0.013
Communication*	0.055	0.060	Companies Responded	0.011	0.021
Discussion*	0.045	0.041	Faculty	0.002	0.001
Field Work*	0.044	0.039	Research Project*	0.000	0.001
Lecture*	0.041	0.041			

*Competency features and curricula contribution.

Overall, most correlations show very weak relationships between the analyzed features and the target labels. That shows that no single factor significantly influences the target label. Most results between Spearman and Pearson correlations are relatively consistent, although slight variations exist. For example, Province feature has quite different values between the two methods (0.113 vs 0.073). Spearman measures monotonic relationships, while Pearson

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 911−920
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5655

measures linear relationships. These small variations may indicate the existence of a non-linear relationship that is better captured by Spearman. After preprocessing data, we split data for next process. Table 6 shows the result of data splitting for model training and testing process with the proportion of 80% data for training and 20% data for testing.

Table 6. Data Splitting Result

	Training Data	**Testing Data**
Quantity	4832	1209
	6041	

The table provides a clear overview of how the data is divided, highlighting the proportions allocated for training and testing, and ensuring that the model's performance can be accurately evaluated and validated. Before testing the feature manipulation and hyperparameter tuning schemes, resampling testing was carried out to overcome data imbalance as shown in Figure 2. Table 7 shows the difference in SVM model performance with and without resampling using SMOTE-ENN.

Table 7. Model Performance Comparison With SMOTE-ENN

	Without Resampling		**Resampling**	
Class	0	1	0	1
Precision	0.00	0.72	**0.38**	0.80
Recall	0.00	1.00	**0.58**	0.64
F1-Score	0.00	0.84	**0.46**	0.71
Accuracy	0.72		0.62	

It can be noticed that the performance of the SVM model based on the accuracy score is worse when data resampling is performed than without resampling. However, the model with resampling could have predicted the other classes, indicated by an increase in precision, recall, and f1-score scores from 0.00 to 0.38, 0.58, and 0.46, respectively, in other classes. Therefore, in the following scheme, data resampling is still applied so that the model can better predict both classes.

The next step is to test the SVM model scheme involving feature manipulation and hyperparameter tuning using Randomized Search. Table 8 below presents the selected parameters for each scheme, detailing the kernel type, C value, and gamma value used.

Table 8. Selected Parameter on Each Schemes

Scheme	**Parameter**		
	kernel	**C**	**gamma**
Default Parameter without Feature Manipulation	rbf	1	‘scale’
Tuning without Feature Manipulation	rbf	10	0.01
Default Parameter with Spearman	rbf	1	‘scale’
Tuning with Spearman	rbf	10	0.1
Default Parameter with PCA	rbf	1	‘scale’
Tuning with PCA	rbf	10	0.01

These parameters were used for the respective experiments and play a crucial role in the model's performance. Next step is SVM models' training process which used stratified k-fold cross-validation with k=5 to make it the same amount of data division in each k. Table 9 below summarizes the model performance on the training data for each scheme, reporting accuracy, recall, precision, and f1-Score.

Table 9. Model Performance on Training Data

Scheme	**Training Score**			
	Accuracy	**Recall**	**Precision**	**F1-Score**
Default Parameter without Feature Manipulation	0.43	0.43	0.73	0.42
Tuning without Feature Manipulation	0.63	0.63	0.71	0.66
Default Parameter with Spearman	0.64	0.64	0.69	0.66
Tuning with Spearman	**0.71**	**0.71**	**0.68**	**0.69**
Default Parameter with PCA	0.48	0.48	0.68	0.50
Tuning with PCA	0.67	0.67	0.68	0.67

These results show the SVM models' performance metrics on each scheme's training data. The results indicate that Randomized search hyperparameter tuning with Spearman's rank correlation coefficient yielded the highest accuracy, recall, and f1-score suggesting improved model performance compared to other schemes on training data. After the training process, we tested SVM models on testing data. Table 10 below presents the performance of the SVM models on the testing data, highlighting accuracy, recall, precision, and f1-Score for each scheme.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 911−920
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5655

Table 10. Model Performance on Testing Data

Scheme	Testing Score			
	Accuracy	**Recall**	**Precision**	**F1-Score**
Default Parameter without Feature Manipulation	0.43	0.43	0.71	0.42
Tuning without Feature Manipulation	0.62	0.62	0.68	0.64
Default Parameter with Spearman	0.65	0.65	0.67	0.66
Tuning with Spearman	**0.73**	**0.73**	**0.70**	**0.71**
Default Parameter with PCA	0.46	0.46	0.67	0.47
Tuning with PCA	0.68	0.68	0.66	0.67

The scheme using Spearman's rank correlation coefficient for feature selection combined with hyperparameter tuning achieved the highest accuracy, recall, and f1-score, confirming its effectiveness in improving model performance. Following the best scheme, we presented Table 11, which shows the confusion matrix of this best scheme on testing data.

Table 11. Confusion Matrix of Testing Data for The Best SVM Scheme

Classfication	Predicted	
	Positive(+)	**Negative (-)**
Actual	107	226
	104	772

Following the execution of multiple experiments utilizing various testing schemes, the testing scheme that used Spearman's rank correlation coefficient for feature selection alongside Randomized Search for hyperparameter tuning yielded the best results. This conclusion was drawn based on the evaluation scores using confusion matrix, which indicated a precision score of 0.70, an f1-score of 0.71, as well as accuracy and recall scores of 0.73 each. The optimal parameter combination, derived from the hyperparameter tuning process, was identified to be a rbf kernel with a C value of 10 and a gamma value of 1.

Furthermore, we provide hypothesis testing to determine whether factors involving feature manipulation and hyperparameter tuning affected model performance. Hypothesis testing is carried out based on accuracy scores in cross-validation during the model training process. The details of the hypotheses being tested are outlined in Table 12.

Table 12. Hypothesis Information

Factor	Hypothesis	Description
Feature Manipulation	H_0	There is no significant effect of feature manipulation on model performance.
	H_1	There is significant effect of feature manipulation on model performance.
Hyperparameter Tuning	H_0	There is no significant effect hyperparameter tuning on model performance.
	H_1	There is significant effect of hyperparameter tuning on model performance.
Interaction	H_0	There is no significant effect of hyperparameter tuning and feature manipulation on model performance.
	H_1	There is significant effect of hyperparameter tuning and feature manipulation on model performance.

Table 12 specifies the null (H_0) and alternative (H_1) hypotheses for each factor: feature manipulation, hyperparameter tuning, and their interaction. The goal is to determine whether these factors individually or combined have a significant impact on the model's accuracy. The following Table 13 presents the results of the hypothesis testing, including the F-statistic, P-value, and F-critical, to assess whether the factors and their interaction significantly impact model performance. This testing used a significance level of 0.05.

Table 13. Hypothesis Testing Results with Significance Level 0.05

Source of Variations	F	P-Value	F-critical	Result
Feature Manipulation	66.835	7.1422E-14	2.901	H_0 Reject
Hyperparameter Tuning	443.01	1.6022E-20	4.149	H_0 Reject

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 911−920
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5655

Interaction	16.694	1.0455E-06	2.901	H_0 Reject

Based on a comparison of either the P-value with significance level or the F-critical with F-statistic, the P-value shows a lower value than the significance level, and the F-critical also shows a lower value than the F-statistic, therefore each H_0 is rejected. These results found that the use of hyperparameter tuning, and feature manipulation affected model performance.

The final stage is to analyze competency and curricula contribution features important to the model. Decreased model performance based on accuracy, recall, precision, and f1-score scores measure feature importance. Figure 3 display boxplots for the competency features that influence the model most based on the previously mentioned metric score.

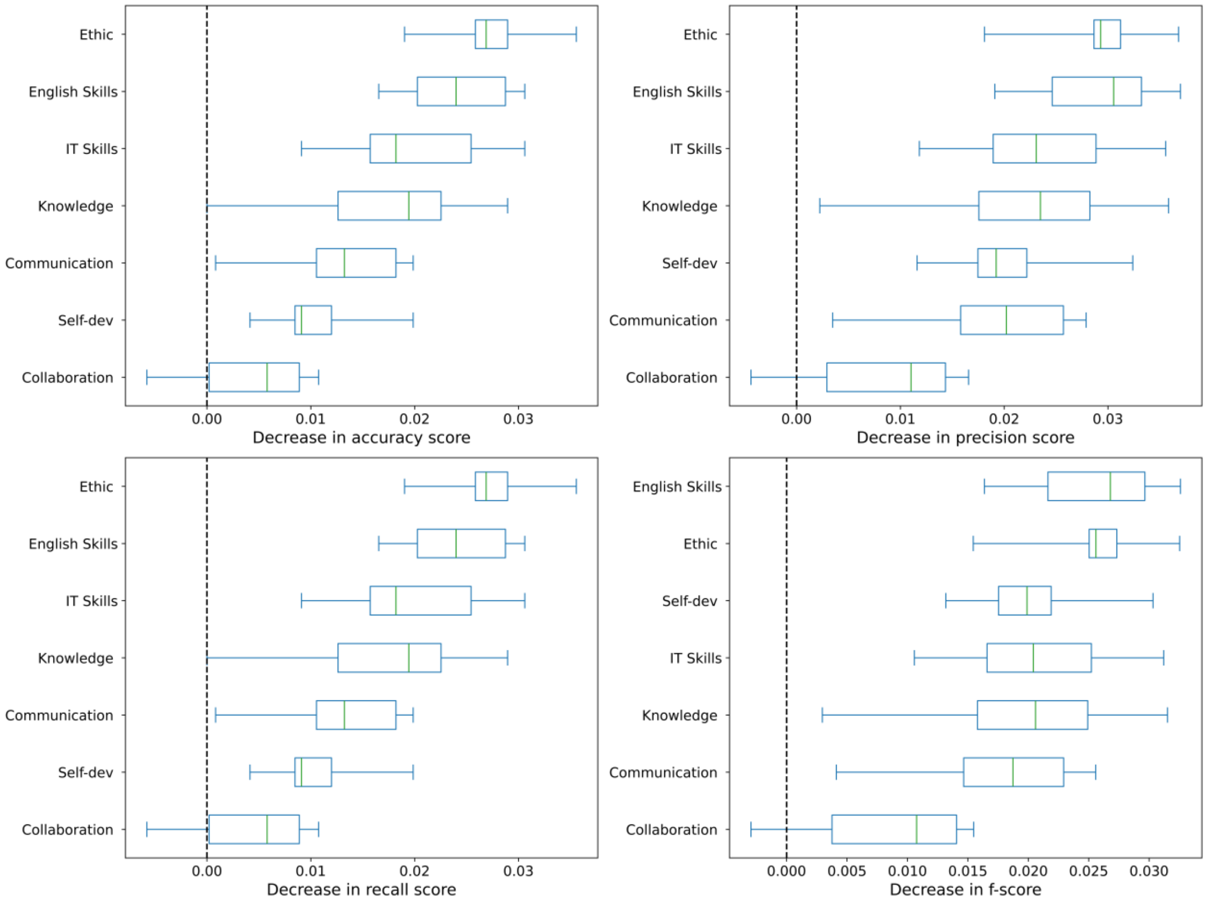

Figure 3. Boxplots Result of Permutation Feature Importance

The test results show that each competency feature affects the model because each metric score decreases, with an average decrease value above 0.004 on accuracy and recall score and above 0.008 for precision and f1-score for collaboration feature as it is the least important feature of all boxplots. Then, it is found that the competency features ethics, english Skills, information technology skills, and knowledge are the most frequent features that appear at the top of the four boxplots.

Similar research analyzed the Telkom University 2022 tracer study dataset but employed Logistic Regression to investigate the relationship between students' competencies and their waiting time to secure the first job after graduation. They found competencies like hard skills, communication, and course knowledge were most influential, providing insights through odds ratios quantifying predictors' effects. In contrast, our SVM based approach predicted employability via initial job income, utilizing feature manipulation and hyperparameter tuning techniques. While the target variable differed, both studies consistently highlighted the importance of competencies acquired during university studies, like ethics, English skills, IT skills, and knowledge, in shaping graduate employment outcomes. The Logistic Regression allowed direct modeling of categorical outcomes, while SVM offered an alternative perspective by targeting income, a direct career success measure. The methodologies provided complementary insights into the complex education and employment relationship.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 911−920
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5655

4. CONCLUSION

Several points from this research can be concluded. Initial correlation analysis using Spearman and Pearson methods indicated weak relationships between features and target labels, with minor variations suggesting potential non-linear relationships better captured by Spearman. Resampling using SMOTE-ENN improved the model's ability to predict minority classes, as evidenced by increased precision, recall, and f1-score, despite a lower accuracy. The SVM classification results showed that Randomized Search hyperparameter tuning combined with Spearman's rank correlation for feature selection led to the best model performance, achieving testing accuracy, recall, precision, and f1-score above 70%. Training and testing result have a consistent score that indicates that the model effectively balances complexity and generalization, avoiding overfitting and underfitting. Hypothesis testing confirmed that hyperparameter tuning and feature manipulation significantly impacted model performance. Lastly, permutation feature importance analysis identified ethics, English Skills, IT Skills, and knowledge as the most influential features of the model. In future research, several ways could be explored to build upon these findings. For instance, experimenting with alternative machine learning models such as Gradient Boosting, XGBoost, or deep learning techniques could provide valuable comparisons and potentially improved performance. Additionally, exploring other hyperparameter tuning techniques, such as Bayesian Optimization or Grid Search, might yield further enhancements in model performance. Investigating different feature selection methods, including Recursive Feature Elimination (RFE) could also be beneficial in identifying the most critical features.

REFERENCES

[1] Muhammad Hadiza Baffa, Muhammad Abubakar Miyim, and Abdullahi Sani Dauda, “Machine Learning for Predicting Students’ Employability,” *UMYU Scientifica*, vol. 2, no. 1, pp. 001–009, Feb. 2023, doi: 10.56919/usci.2123_001.

[2] G. ElSharkawy, Y. Helmy, and E. Yehia, “Employability Prediction of Information Technology Graduates using Machine Learning Algorithms,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, pp. 359–367, 2022, doi: 10.14569/IJACSA.2022.0131043.

[3] F. F. Abdulloh, M. Rahardi, A. Aminuddin, S. D. Anggita, and A. Y. A. Nugraha, “Observation of Imbalance Tracer Study Data for Graduates Employability Prediction in Indonesia,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, pp. 169–174, 2022, doi: 10.14569/IJACSA.2022.0130820.

[4] P. H. Gunawan, I. Palupi, A. A. Rohmawati, and A. T. Hanuranto, “Correlation Analysis of Student’s Competencies and Employ-Ability Tracer Study of Telkom University Graduates,” Proceedings of the Computational Methods in Systems and Software, Feb. 2024.

[5] A. Miranda and K. M. Lhaksamana, “Classification Analysis of Waiting Period for Telkom University Alumni to Get Jobs Using Decision Tree and Support Vector Machine,” *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 2, Sep. 2022, doi: 10.47065/bits.v4i2.1963.

[6] S. D. Abdul Bujang, A. Selamat, and O. Krejcar, “A Predictive Analytics Model for Students Grade Prediction by Supervised Machine Learning,” *IOP Conf Ser Mater Sci Eng*, vol. 1051, no. 1, p. 012005, Feb. 2021, doi: 10.1088/1757-899x/1051/1/012005.

[7] P. W. Yunanto, A. Idrus, V. M. Santi, and A. S. Hanif, “Tracer study information system for higher education,” *IOP Conf Ser Mater Sci Eng*, vol. 1098, no. 5, p. 052107, Mar. 2021, doi: 10.1088/1757-899x/1098/5/052107.

[8] R. C. Mae Narciso, “Tracer Study of BSBA Major in Human Resource Management Graduates (2017-2020) of UM Tagum College,” *International Journal of Entrepreneurship and Business Development*, vol. 4, 2021, doi: https://doi.org/10.29138/ijebd.v4i5.1462.

[9] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, “Classification and prediction of student performance data using various machine learning algorithms,” *Mater Today Proc*, vol. 80, pp. 3782–3785, Jan. 2023, doi: 10.1016/j.matpr.2021.07.382.

[10] A. C. Albina and L. P. Sumagaysay, “Employability tracer study of Information Technology Education graduates from a state university in the Philippines,” *Social Sciences and Humanities Open*, vol. 2, no. 1, Jan. 2020, doi: 10.1016/j.ssaho.2020.100055.

[11] N. M. Basir, Y. Z. Zubairi, R. Jani, and D. A. Wahab, “Soft Skills and Graduate Employability: Evidence from Malaysian Tracer Study,” *Pertanika Journal of Social Sciences and Humanities*, vol. 30, no. 4, pp. 1975–1989, Dec. 2022, doi: 10.47836/pjssh.30.4.26.

[12] Z. A. Adriani and I. Palupi, “Prediction of University Student Performance Based on Tracer Study Dataset Using Artificial Neural Network,” *Jurnal Komtika (Komputasi dan Informatika)*, vol. 5, no. 2, pp. 72–82, Nov. 2021, doi: 10.31603/komtika.v5i2.5901.

[13] L. Jovidia, P. H. Gunawan, and Indwiarti, “Telkom Univ Waiting Time Analysis: SVM vs. Logistic Regression with Tracer Study Data of 2022,” in *2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)*, 2024. doi: 10.1109/ICoSEIT60086.2024.10497466.

[14] M. Mustaqeem and M. Saqib, “Principal component based support vector machine (PC-SVM): a hybrid technique for software defect detection,” *Cluster Comput*, vol. 24, no. 3, pp. 2581–2595, Sep. 2021, doi: 10.1007/s10586-021-03282-8.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 911−920
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5655

[15] Q. R. S. Fitni and K. Ramli, "Implementation of Ensemble Learning and Feature Selection for Performance Improvements in Anomaly-Based Intrusion Detection Systems," in *The 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 2020.

[16] S. T. Nihan, "Karl Pearsons chi-square tests," *Educational Research and Reviews*, vol. 15, no. 9, pp. 575–580, Sep. 2020, doi: 10.5897/err2019.3817.

[17] L. Villalobos-Arias, C. Quesada-López, J. Guevara-Coto, A. Martínez, and M. Jenkins, "Evaluating hyper-parameter tuning using random search in support vector machines for software effort estimation," in *PROMISE 2020 - Proceedings of the 16th ACM International Conference on Predictive Models and Data Analytics in Software Engineering, Co-located with ESEC/FSE 2020*, Association for Computing Machinery, Inc, Nov. 2020, pp. 31–40. doi: 10.1145/3416508.3417121.

[18] V. Blanco, A. Japón, and J. Puerto, "A mathematical programming approach to SVM-based classification with label noise," *Comput Ind Eng*, Aug. 2022, doi: https://doi.org/10.1016/j.cie.2022.108611.

[19] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput Oper Res*, vol. 152, Apr. 2023, doi: 10.1016/j.cor.2022.106131.

[20] J. Gómez-Ramírez, M. Ávila-Villanueva, and M. Á. Fernández-Blázquez, "Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-77296-4.