Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

Decision Tree Algorithm for Predicting Alumni Job Competitiveness Through Waiting Time Working

Bagus Panuluh, Irma Palupi*, Putu Harry Gunawan

School of Computing, Informatics, Telkom University, Bandung, Indonesia
Email: bagusp@student.telkomuniversity.ac.id, irmapalupi@telkomuniversity.ac.id, phgunawan@telkomuniversity.ac.id
Correspondence Author Email: irmapalupi@telkomuniversity.ac.id
Submitted: **12/07/2024**; Accepted: **09/09/2024**; Published**: 09/09/2024**

Abstract−The absorption of alumni from universities into the world of work is an essential indicator that universities must pay attention to. One-way universities can pay attention to their alums is through tracer studies, where they can evaluate their curriculum's relevance to what is needed in today's world of work. One aspect that can be seen from the tracer study to assess the competitiveness of alums is the waiting time for alums to get their first job. This is because the sooner alums get jobs, the better the curriculum the university provides to students. This research aims to apply machine learning to predict the waiting time for alums from Telkom University to get their first job and find out what factors influence the waiting time for work. The algorithm used in the research is the Decision Tree with hyperparameter tuning using Grid Search and feature selection application. There are 3 methods of feature selection used for comparison: Spearman's Rank Correlation, Chi-square, and Principal Component Analysis. This research produces the best prediction model in applying Chi-square and hyperparameter tuning with an accuracy of 0.79, recall of 0.79, precision of 0.80, and F1-Score 0.75. Several features, such as the number of companies registered, how to find and get work, internship and practicum experience, ethical competency, discussion, and IT skills, have the biggest effects on the model.

Keywords: Tracer Study; Waiting Time; Correlation; Decision Tree; Feature Selection

1. INTRODUCTION

The high level of alum absorption in the world of work contributes to a region. This significantly impacts economic development by reducing unemployment and stimulating economic activity through their jobs, taxes, and spending. At the social level, employment improves the quality of life of alums and their families, and they often contribute to community development in various ways. In addition, feedback from alums also has an impact on the world of education.

The absorption of university alumni in the world of work is an essential indicator in measuring the success of universities in preparing their graduates to enter the world of work. This process includes several stages, from academic education to developing non-academic skills needed in the world of work. Universities that are successful in alum placement usually have good curricula and skills development programs [1]. Additionally, support from university career centers plays an important role in helping graduates identify employment opportunities. Thus, good alumni absorption reflects the university's commitment to providing a holistic education that focuses not only on academics but also on comprehensive preparation for the world of work.

At Telkom University, alumni conditions are monitored through tracer studies. A tracer study is a method of tracking the career or life progress of educational institution graduates after completing a particular study or training program. [2]. The main focus of the tracer study is to understand the extent to which the education or training received by graduates has prepared them to enter the world of work. In the context of universities or other educational institutions, implementing Tracer studies can help institutions demonstrate the quality of the education they provide, identify the advantages and disadvantages of study programs, and improve the educational climate so that it is more in line with the needs of the world of work [3]. Data collected through tracking studies includes information such as status after graduating from college, assessment of aspects of ability, income level, and job placement level [3]. The results of the tracer study provide valuable input for educational institutions to improve the quality of education and prepare students more effectively to face the demands of the world of work [4].

The material in the tracer study comes from Direktorat Jenderal Pendidikan Tinggi (DIKTI) with the addition of several fields required by various faculties [1]. One of the materials in the Telkom University tracer study is about the waiting time for alums to get their first job. The length of time it takes for alums to enter the world of work can reflect the extent to which the educational program has succeeded in preparing them to face the challenges of the job market. By understanding the waiting time for alums to find work, universities can assess the effectiveness of their educational strategies and adapt them to changing job market dynamics to increase the competitiveness of prospective graduates. Understanding and evaluating the character and results of alums can be done using machine learning methods [5]. By utilizing machine learning algorithms, universities can analyze historical data and variables that influence alums's success in finding their first job.

The aim of this research is to predict the waiting time for Telkom University alumni in 2020-2021 to get their first job using a machine learning algorithm, namely Decision Tree. In previous research at Telkom University, research was conducted to analyze the correlation between student competence during lectures and the waiting time to get their first job. Using one of the machine learning algorithms, namely Logit Regression and correlation calculations using Spearman's Rank-Order Correlation and Chi-square, it was found that waiting time for work has a

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

high correlation with competencies developed in the university curriculum, namely with a correlation of more than 90% [6]. Other research at Telkom University was also carried out to predict alums' waiting time for work using a machine learning algorithm, namely an Artificial Neural Network (ANN). In this research, the accuracy of the ANN model was obtained at 0.77. Then, efforts were made to increase the accuracy by applying K-Fold, and the accuracy was increased to 0.87 [3].

One of the commonly used machine learning algorithms is Decision Tree (DT) [7]. This algorithm was chosen in this research because it has an algorithm that is easy to understand, easy to use, has a good prediction level, and high accuracy results [8]. This was proven in previous research, namely employability predictions from survey data given to IT graduates and employers in Egypt using five types of algorithms, namely DT, Random Forest, Gaussian Naive Bayes, Support Vector Machine, and Logistic Regression. Through several performance calculations, it was found that DT produces the highest accuracy, namely reaching 100% [1].

In other research with the aim of classifying work waiting times using DT and Support Vector Machine, the highest accuracy was found in the DT model. After setting the hyperparameters on the DT model with the parameter set being maximum depth, the highest accuracy was obtained at 65%. [9]. Apart from that, in research conducted in one of the Indian cities to predict student employability using the DT, Random Forest, Naive Bayes, and K-Nearest Neighbour algorithms, the results showed that DT had an accuracy of 87.79%. [10].

Based on the performance results of the DT model in previous research, this research uses DT as the primary model for making predictions. This research focuses on developing a model to predict the waiting time for alums to get their first job using the DT algorithm. It was also identified which features had the most significant effect on working waiting time-based on the best model that had been developed. Details of the research methodology are explained in Section II. The results of the research are described in Section III. Meanwhile, conclusions from the research results are given in Section IV.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The research aims to predict the waiting time for alumni to get their first job as a representation of competitiveness using the DT algorithm. To implement the proposed methods, the Python programming language is used, combining various libraries such as Pandas and Numpy for preparing data, Scikit-learn for feature selection and creating machine learning model, Imblearn for handling imbalanced datasets, and Seaborn and Matplotlib for visualization. Figure 1 illustrates the stages of the research process.

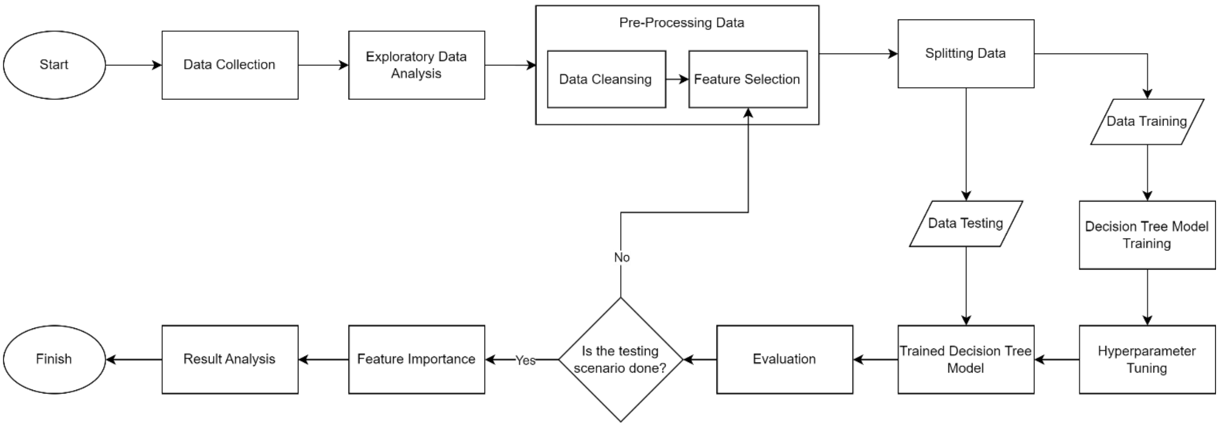

Figure 1. Research Flowchart Diagram

The flowchart outlines the key stages in preparing and processing data for a machine learning model. Initially, raw data undergoes pre-processing, which involves cleaning and formatting it for analysis. Data collection follows, gathering information from various sources. Exploratory Data Analysis is then performed to understand data characteristics through visualization and summary statistics. Feature selection identifies the most relevant data features to enhance model accuracy. The data is split into training data and testing data, followed by data cleansing to remove errors and inconsistencies. The model is then trained on the training data and evaluated on the testing data. Evaluation measures are computed, including recall, accuracy, precision, and F1-score. Hyperparameter tuning adjusts model settings for optimal performance. The process concludes with the selection of a finalized model for use.

2.2 Dataset Collection and Pre-Processing

Gathering the dataset was the initial step in the study process. The Telkom University Career Alumni Endowment provided the 2022 Tracer study dataset, which was used in this study. Table 1 show an overview of the dataset used.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

Table 1. Dataset Attributes

Attribute	Description	Attribute	Description
study_program	Alumni's study program	time_management*	Time management competency
faculty	Alumni's faculty	communication*	Communication competency
funds_source	Source of college funds	teamwork*	Teamwork competency
status	Employment status	self_development*	Self-development competency
province	Province of work	english*	English language competency
edu-work_relation	Education background and work relation	IT_skills*	IT skills competency
job_search	How the way to get a job	lecture*	Emphasis on lecture methods
qty_applied	Quantity of company applied	demonstration*	Emphasis on demonstration methods
qty_responded	Quantity of company response	research_project*	Emphasis on research projects methods
qty_interview	Quantity of company invites to interview	internship*	Emphasis on internship methods
company_sector	Sector of the company	practicum*	Emphasis on practicum methods
company_type	Types of company	field_work*	Emphasis on field work methods
ethic*	Self ethical competency	discussion*	Emphasis on discussion methods
knowledge*	Knowledge competency	waiting_time	Waiting time to get a first job

* Rated of alums ability and university contribution.

Table 2 shows the data description used for modelling alumnus competency profiling. The dataset is pre-processed before model training to ensure high accuracy by cleaning irrelevant data. Several procedures are applied during the pre-processing stage, including handling missing values, removing duplicate data, label encoding, addressing imbalanced data, feature selection, and splitting the dataset. Missing values are handled by replacing empty data using the median for numerical data and the mode for categorical data. Attributes with a missing value rate above 50% are neglected. For categorical data types, they are encoded and transformed into numeric data. The methods used are label encoding and dummy encoding. Label encoding is used for attributes with values in the form of a sequence, while dummy encoding is used for attributes with values that are not in the form of a sequence [11].

Table 2. Dataset Profiling

		Quantity
Number of features		28
Number of rows		4945
Feature types	Numerical	4
	Categorical	24
Missing Value	Company type	9
	Waiting_time	79
Duplicate rows		7

2.3 Features Selection

Feature Selection is used to select attributes with the highest correlation with the target variable. Several methods are applied in feature selection, including Spearman's Rank Correlation Coefficient, Chi-Square, and Principal Component Analysis (PCA). Some feature selection was applied to compare the differences in results from using the two methods and find out which method produces higher accuracy after carrying out the modelling process and measuring model performance.

Spearman's rank correlation coefficient assesses the relationship between 2 variables. Its value range is from -1 to 1. A value of -1 indicates a negative perfect correlation, a value of 1 indicates a positive perfect correlation, and a value of 0 indicates no correlation [12]. The calculation of the Spearman rank correlation coefficient is defined as follows.

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

The variable ρ shows the correlation value from Spearman's rank correlation coefficient, where the closer the value of ρ is to 1 or -1, the stronger the relationship between the variables. The variable d shows the difference in ranking and n indicates the number of data pairs.

The Chi-square method is a statistical test that is applied to categorical data to determine whether there is a strong correlation between two variables [13]. The null hypothesis (H_0) in this test symbolizes that the two variables are independent or have no relationship. Meanwhile, the alternative hypothesis (H_1) means the two variables have a

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

relationship. The Chi-square test statistic is calculated based on the difference between the observed and expected frequencies in the absence of an association.

$$x^2 = \sum \frac{(O-E)^2}{E} \quad (2)$$

In the above calculation, O represents the observed frequency and E represents the expected frequency for each category combination. After that, the degrees of freedom are calculated using the following calculations. In this calculation, r represents the number of entries in the table and c is the number of columns.

$$df = (c-1)(r-1) \quad (3)$$

The Principal Component Analysis (PCA) method is used primarily to reduce the dimensions of certain data sets. This methods is one of the most efficient and accurate methods for reducing the dimensionality of data and provides the desired results. In this methods, a given data set's characteristics are reduced to a few desired qualities known as principal components [14].

In its application, PCA involves eigenvalue and eigenvector calculations [14]. The eigenvalue shows how much variation each main component can explain. The greater the eigenvalue, the more variation is presented by the main element in question. By understanding eigenvalues and eigenvectors, PCA allows us to select the most significant principal components and produce data with lower dimensions.

2.4 Decision Tree

Decision Tree (DT) is an algorithm commonly used for classification. Because this algorithm is a part of supervised learning, it requires labelled data to be applied. The way this algorithm works is by forming a tree-like structure consisting of several nodes. A DT algorithm often consists of multiple levels of nodes, with the root node at the top and the leaf node at the tip or bottom [15]. Every subset specifies the values a node can accept, and every node represents an attribute in a category that must be classified [7]. The evaluation of input variables or features is shown by all internal nodes that have at least one child node [16].

A data set's impurity is measured by the Gini index, which calculates the likelihood that an element chosen at random would have a false label if its labelling were determined by its subset [17]. The procedure for measuring the Gini Index is as follows.

$$Gini(L) = 1 - \sum_{i=1}^{c} p_i^{\,2} \quad (4)$$

L is the node in DT, c is the number of classes, and p_i is the proportion of samples in class i. The Gini index reaches a minimum value (0) when all samples at a node belong to one class and reaches a maximum value when the data is evenly divided between classes. In other words, a smaller Gini value indicates that data separation is better because it produces a more homogeneous group [9]. Recursively, the process of identifying decision rules that produce splits to minimize impurities is carried out until either all pure leaf nodes or a predetermined cut point is achieved [18].

2.5 Modelling

The modelling process using DT model algorithm requires hyperparameter tuning to optimize performance. Hyperparameter tuning will be done with Grid Search. This method is a brute-force approach that will try every possible combination of values with cross-validation [18]. Later, the combination of values that produces the highest performance will be selected as the best model.

Key parameters that are tuned include min_samples_leaf and min_samples_split. The min_samples_leaf parameter specifies the minimum number of samples a leaf node must have, while min_samples_split determines the minimum number of samples a node must have before it is split [15]. The values of these parameters, which constitute the sample space for hyperparameter tuning, are detailed in Table 3.

Table 3. Sample Space of DT Hyperparameter

Parameter	Sample
min_samples_split	2, 3, 4, 5, 6, 7, 8, 9, 10
criterion	gini
min_samples_leaf	1, 2, 3, 4, 5, 6, 7, 8, 9, 10

During this process, several modelling experiments are conducted to obtain a DT model with the highest accuracy. The experiments include:

a. Comparing the performance of models with and without hyperparameter tunning parameters.
b. Comparing model performance from different feature engineering, including Spearman's rank correlation coefficient, chi-square, and PCA.
c. Comparing model performance from number of target’s classes defined.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

2.6 Model Evaluation and Gini Importance

One way to evaluate the model created is to use a confusion matrix, which is a table used in evaluating model classification performance to measure the extent to which the model can predict classes correctly [4]. The actual classes of the data are represented by the rows in the table, while the classes that were successfully predicted by the model represented by the columns. Based on the confusion matrix, other performance calculations can be carried out to assess the effectiveness of the model that has been created [19]. Some of these performances are accuracy, precision, recall and f1-score.

Gini Importance is one of the most common mechanisms for calculating the importance of an attribute from data. This mechanism is widely used because it is easy to use and is directly available in most tree-based algorithms such as Random Forest and DT [15]. When creating a decision tree, the algorithm evaluates various features to determine the most effective classification split. Gini importance measures how important a particular feature is in influencing decision making or in other words how good it is at separating data into different classes [20]. A higher Gini importance score implies that a feature plays a more influential role in making accurate predictions [21].

3. RESULT AND DISCUSSION

3.1 Dataset Information

In this section, we present and analyze the findings of our study on waiting times for alumni obtaining their first job. The investigation encompasses data distribution, correlation analysis, model performance, and the influence of various features. We begin by exploring the distribution of waiting times, addressing class imbalance, and identifying key features impacting the target variable. Subsequently, we evaluate the performance of different models and feature selection methods, emphasizing the significance of Chi-Square feature selection combined with hyperparameter tuning. Finally, we discuss the implications of feature importance, particularly focusing on alumni competencies and university contributions, and their role in shaping job waiting times.

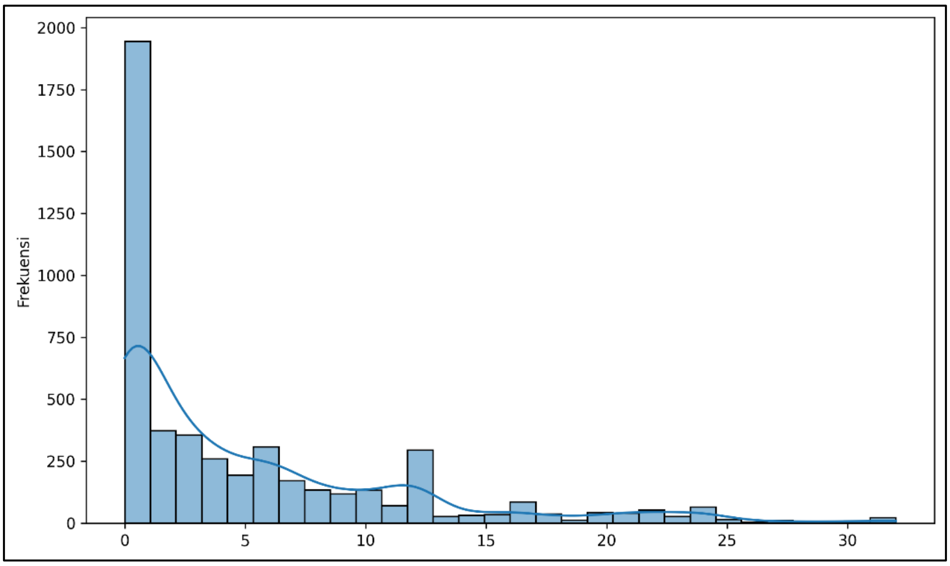

Figure 2. The distribution of waiting time for alumni getting the first job.

Based on Figure 2, which shows the distribution of waiting_time, which is the target class, the distribution of data from these features is not normal. The distribution of this data shows positive skewness, which indicates that the existing data is more biased towards small values. The peak of the distribution is at the value 0, which is the highest value in this dataset, indicating that the majority of data has the lowest waiting time. Conversely, as you move towards higher values, the data frequency decreases, creating a long tail to the right of the graph. This shows that only a small amount of data has a high waiting time. In accordance with the objectives of the research, the data distribution is grouped into several classes. This is also to overcome abnormal data distribution.

In the target class, the waiting_time attribute is grouped into three classes, namely "Before Graduate", "1-6 Months", and ">6 Months". A comparison of the numbers for each class can be seen in Figure 3. In "Before Graduate" there are 1272 data, "1-6 Months" there are 2113 data, and in ">6 Months" there are 1428 data. There is imbalanced data with the second class or "1-6 Months" being the majority class when compared to other classes. An imbalanced dataset is handled using SMOTE-ENN. As part of the experimental scenario, target features were also grouped into two classes.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

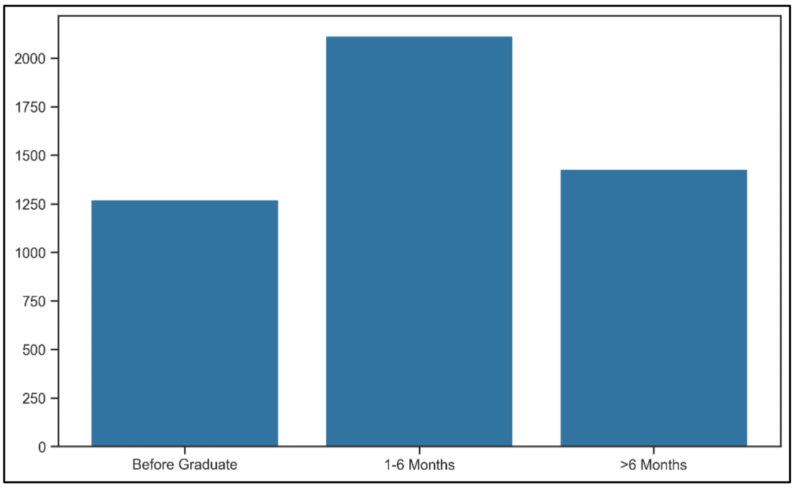

Figure 3. Distribution of Target class with 3 fractions.

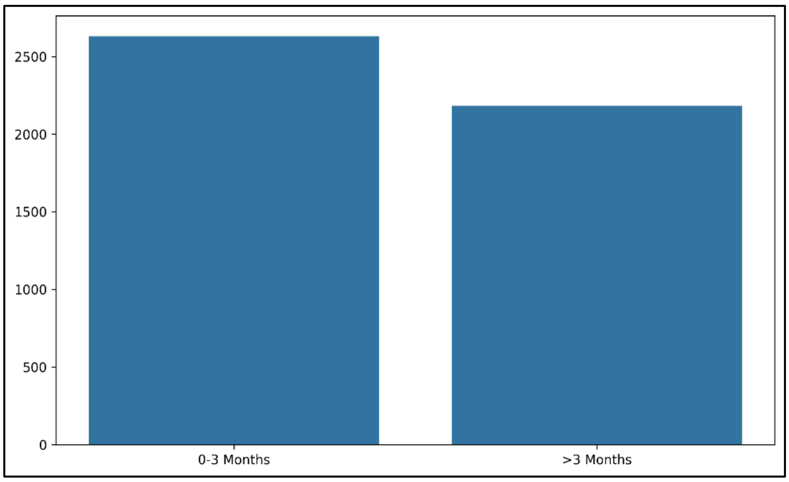

Figure 4. Distribution of Target class with 2 fractions.

Figure 4 shows a bar plot of the target feature, namely waiting_time when divided into two classes. The division is based on the median value of waiting_time which shows a value of 3. Based on this, the class is divided into 0-3 months and more than three months. Classes with 0-3 months totalled 2631 and classes with >3 months totalled 2182. In this class division, there is still imbalanced data, with the 0-3-month class being the majority. Even though the data imbalance in dividing into two classes is not too big when compared to dividing into three classes, it is still handled using SMOTE-ENN.

Table 4. Correlation of features with targets feature

Attribute	Spearman	Kendall	Attribute	Spearman	Kendall
study_program	0.023	0.019	communication*	0.070	0.064
faculty	0.039	0.032	teamwork*	0.066	0.060
funds_source	0.040	0.038	self_development*	0.033	0.030
province	0.025	0.021	english*	0.004	0.002
edu-work_relation	0.083	0.072	IT_skills*	0.066	0.060
job_search	0.103	0.088	lecture*	0.070	0.061
qty_applied	0.263	0.209	demonstration*	0.069	0.061
qty_responded	0.233	0.187	research_project*	0.069	0.061
qty_interview	0.216	0.175	internship*	0.087	0.076
company_sector	0.023	0.019	practicum*	0.093	0.081
company_type	0.035	0.032	field_work*	0.107	0.094
ethic*	0.049	0.045	discussion*	0.153	0.135
knowledge*	0.043	0.038	time_management*	0.013	0.011

Table 4 shows the correlation of each feature with waiting time for work using Spearman and Kendall as a comparison. Spearman correlation measures the monotonic relationship between 2 variables based on ranking. Meanwhile, Kendall correlation is an ordinal relationship between two variables by calculating concordant and

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

discordant pairs [22]. Based on these correlation calculations, there is no significant difference between Spearman and Kendall. Both show a correlation that tends to be low for the majority of features towards waiting_time in the dataset, with a range between 0.2 to 0.002. Both Spearman and Kendall produced the highest correlation with waiting_time on the features qty_applied, qty_responded, and qty_interview.

Table 5. P-Value features with target feature

Attribute	P-Value	Attribute	P-Value
study_program	0.64862	communication*	0.45831
faculty	0.00105	teamwork*	0.38699
funds_source	0.52456	self_development*	0.70510
province	0.00201	english*	0.21428
edu-work_relation	0.00111	IT_skills*	0.19054
job_search	0.00000	lecture*	0.01420
qty_applied	0.00000	demonstration*	0.00836
qty_responded	0.00000	research_project*	0.01051
qty_interview	0.00000	internship*	0.00126
company_sector	0.70510	practicum*	0.00143
company_type	0.10858	field_work*	0.00004
ethic*	0.33951	discussion*	0.00000
knowledge*	0.95971	time_management*	0.68734

In contrast using correlation test as the result in Table 5, through p-value of each feature on waiting time for work, there are several features that have a strong influence on the target class. Nearly half of the features in the dataset have a p-value below 5%. For example, job_search, qty_applied, qty_responded, and qty_interview with a p-value reaching 0.0000, which means that several of these features have a significant influence on waiting time for work. Even though in order the features qty_applied, qty_responded, and qty_interview are both the top features from both Spearman correlation and P-Value, calculations using P-Value show that these three features have a more significant influence on the target class.

3.2 ModellingApplication

The features in the dataset will not be used entirely, so three types of feature selection methods will be applied, namely Spearman Rank Correlation, Chi-Square, and PCA, as explained in section II. Each application of feature selection will be used to select the ten best features. Spearman feature selection is obtained from ten features with the highest correlation with the target class, namely waiting_time. Feature selection with Chi-Square is obtained from ten features with the highest Chi-Square Score or the lowest P-value. While PCA will reduce the original data into ten new features as the main components. Then, the ten features are used during the modelling process.

The application of modelling begins with handling imbalanced data using SMOTE-ENN. This method combines oversampling and undersampling techniques to improve the distribution of classes in the dataset. Furthermore, the stratified k-fold cross-validation process is carried out with k of 5. This will make each iteration, one fold is used as a testing set, while the other four folds are used as training sets. In other words, at each iteration, the dataset is divided into 20% as testing data and 80% as training data.

Table 6. Best Parameter Value in Each Schemes

Modelling Scheme	min_sample_split	min_sample_leaf
Spearman Rank Correlation (2 Class Target)	2	1
Spearman Rank Correlation (3 Class Target)	10	2
Chi-Square (2 Class Target)	3	8
Chi-Square (3 Class Target)	7	10
PCA (2 Class Target)	2	1
PCA (3 Class Target)	4	1

One of the experiments is to apply hyperparameter tuning to the DT model. Table 6 shows the value of each best parameter when hyperparameter tuning is performed. Each model and feature selection method tested was searched for its best parameters using Grid Search. The use of Grid Search was chosen to test every possibility of each parameter combination. On the other hand, experiments with default parameters using min_sample_split values of 2 and min_sample_leaf values of 1.

3.3 Modelling Results

To find the best model, several tests are carried out on hyperparameter tuning and selecting the feature selection method. Three types of feature selection: Spearman Rank Correlation, Chi-Square, and PCA are used to select the ten

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

best features for modelling. In addition, experiments are also conducted to compare the model's performance with different numbers of classes, specifically 2 and 3 classes.

The performance is calculated using the accuracy, recall, precision, and f1-score values. Considering the many scenarios for the experiment as explained in Section II, the calculation of all performances is obtained with the help of the Scikit-learns.metrics library to make it easier. Because the performance of recall, precision, and f1-score are calculated for each class, calculating the three performances applies a weighted average. This average calculates the average of metrics by giving weight to each class based on the proportion of the number of samples in that class. After that, the average value of each performance is calculated for each fold to get the final performance value.

Table 7. Model Performance with 3 Class Targets

Feature Selection	Accuracy	Recall	Precision	F1-Score
Spearman Rank Correlation (Default Parameter)	0.47	0.47	0.48	0.47
Spearman Rank Correlation (Hyperparameter Tuning)	0.48	0.48	0.9	0.48
Chi-Square (Default Parameter)	0.48	0.48	0.50	0.43
Chi-Square (Hyperparameter Tuning)	**0.50**	**0.50**	**0.55**	**0.44**
PCA (Default Parameter)	0.39	0.39	0.43	0.38
PCA (Hyperparameter Tuning)	0.40	0.40	0.44	0.38

Table 7 shows the performance of the model using three classes of target features. After carrying out several experiments, the results were obtained that the application of Chi-square with hyperparameter tuning using Grid Search produced the highest model performance compared to other testing results. By setting the parameters in min_sample_split and min_sample_leaf, the best combination of hyperparameters was obtained with a min_sample_split value of 7 and min_sample_leaf of 10. The best performance was obtained with an accuracy of 0.50, precision of 0.55, recall of 0.50, and F-1 Score of 0.44.

Table 8. Model Performance with 2 Class Targets

Feature Selection	Accuracy	Recall	Precision	F1-Score
Spearman Rank Correlation (Default Parameter)	0.67	0.67	0.74	0.69
Spearman Rank Correlation (Hyperparameter Tuning)	0.67	0.67	0.72	0.68
Chi-Square (Default Parameter)	0.78	0.78	0.77	0.75
Chi-Square (Hyperparameter Tuning)	**0.79**	**0.79**	**0.80**	**0.75**
PCA (Default Parameter)	0.63	0.63	0.72	0.65
PCA (Hyperparameter Tuning)	0.64	0.64	0.72	0.66

To compare model performance, an experiment was conducted with two target classes. As shown in Table 8, the scenario with two target classes demonstrated better performance than the scenario with three target classes. However, in both scenarios, the best model was achieved using chi-square feature selection combined with hyperparameter tuning. The optimal parameters identified were min_sample_split with a value of 3 and min_sample_leaf with a value of 8. For the two-class target scenario, the highest performance metrics were achieved, with an accuracy of 0.79, precision of 0.80, recall of 0.79, and F-1 Score of 0.75.

Additionally, a hypothesis test was conducted using two-way Analysis of Variance (ANOVA) to determine the influence of feature selection methods, hyperparameter tuning, and their interaction on model performance. The ANOVA was using samples from the highest performing model, namely from modelling using 2 classes and the accuracy results from each testing scenario. Samples for accuracy, obtained with replication from each fold in 5-fold cross-validation, were used in the testing.

Table 9. ANOVA Hypothesis Information

Factor	Hypothesis	Description
Feature Selection	H0	Feature selection has no significant effect on the performance of the model
	H1	Feature selection has a significant effect on the performance of the model
Hyperparameter Tuning	H0	Hyperparameter tuning has no significant effect on the performance of the model
	H1	Hyperparameter tuning has a significant effect on the performance of the model
Interaction	H0	Both feature selection and Hyperparameter have no significant effect on the performance of the model
	H1	Both feature selection and Hyperparameter have significant effect on the performance of the model

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

Table 9 provides a summary of the hypotheses tested. The null hypothesis (H0) indicates that changes in the feature selection process, hyperparameter adjustments, or a combination of both do not affect model performance or measured evaluation results. In contrast, the alternative hypothesis (H1) states that these factors significantly influence model performance results, so variations in feature selection, hyperparameter adjustments, or a combination of both will impact them.

Table 10. ANOVA results

Source of Variations	F	P-Value	F critical	Result
Feature Selection	108.8101	9.22E-13	3.402826	**H0 rejected**
Hyperparameter Tuning	0.316456	0.578963	4.259677	H0 accepted
Interaction	0.050633	0.950729	3.402826	H0 accepted

The results of the ANOVA calculation are displayed in Table 10, with the alpha value set at 0.05. H0 accepted occurs when the P-value value is more than the alpha value, while H0 rejected occurs when the P-value has a value less than the alpha value. Based on the experimental results, it was found that the feature selection setting had an F critical that was lower than the F value and a P-value below the alpha value, resulting in H0 being rejected, which means it influenced model performance. Meanwhile, hyperparameter settings have no influence on model performance based on the results of ANOVA calculations. Next, an analysis is carried out for the feature importance of DT in the best model scenario according to the one in Table 7, namely feature selection with chi-square combined with hyperparameter tuning.

3.4 Feature Importance

Based on the feature importance analysis in Figure 5, the number of companies applied for is the most influential factor in the model, indicating that applying to more companies increases the likelihood of a shorter waiting time for work. The job_search feature follows, suggesting that the method of job searching impacts waiting time. The emphasis on internships and discussions is also one of the most influential factors, which shows that companies are pretty focused on internship experience and discussion skills for alums when accepting jobs. The province feature also plays a significant role, showing that the choice of province for job applications can affect waiting time, with some provinces having more job vacancies than others. Additionally, competencies such as field work and practicum, while not at the top of the feature importance list, still influence waiting time. Interestingly, the qty_interview and qty_respond features, which had significant p-values in Table 5, were found to be the least important in feature importance analysis.

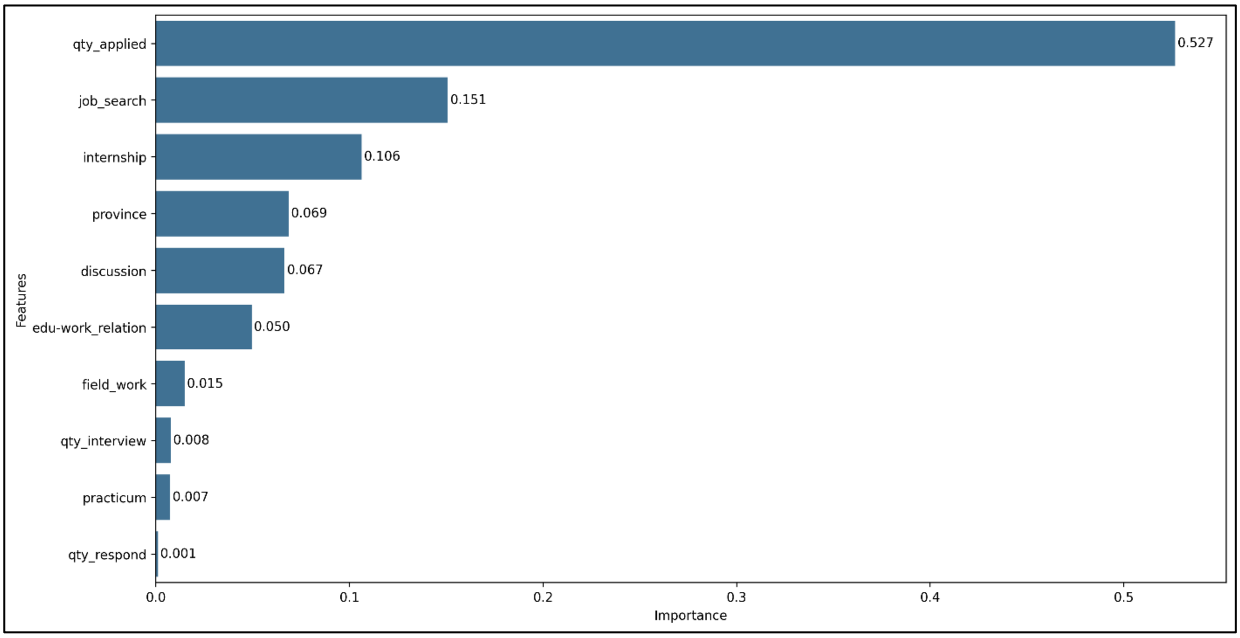

Figure 5. Feature Importance

Further analysis was conducted to evaluate the importance of features related to student competencies and university contributions, as shown in Figure 6. It shows feature importance if the features involved in the model are only data originating from features with values containing the level of alum competency and university contribution, which in the future will be referred to as competency feature importance. The model used is the same as that used for feature importance in Figure 5, namely DT with feature selection using chi-square and a combination of hyperparameter tuning. As a result, of course, some of the features in Figure 6 are still included in the competency importance features such as discussion, internship, field_work, and practicum. However, what is interesting is that several features were not previously in Figure 5 but are one of the top features in Figure 6. One of the most prominent

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

features is english and ethics, which are the top features in feature importance. Some other features available are IT_skill, and research_project. There is also a practicum feature, which was previously the lowest feature in Figure 5 but is actually the second most important feature in competency feature importance.

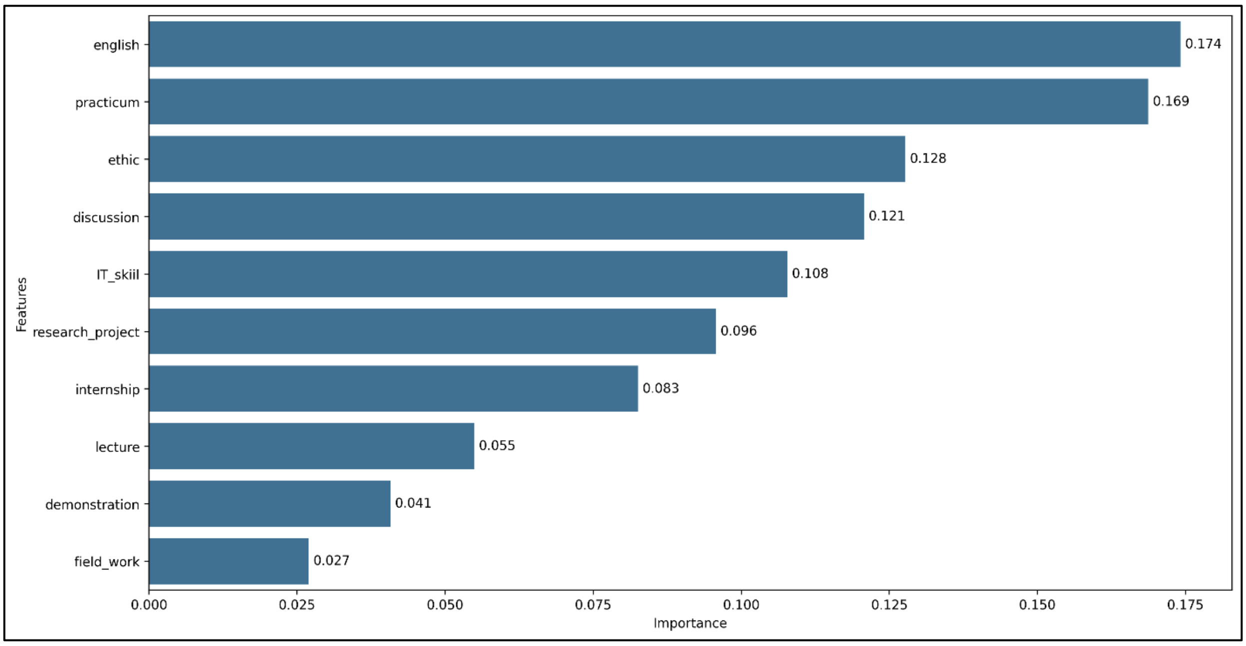

Figure 6. Competency Feature Importance

4. CONCLUSION

In this study, we applied machine learning techniques to a tracer study dataset to predict the waiting time for alums to get their first job. The findings show that the best performance is obtained from modelling the Decision Tree algorithm using target features with 2 classes and applying a combination of chi-square and hyperparameter tuning. The highest performance metrics achieved include accuracy, recall, precision, and F1 score, all exceeding 0.75. Hypothesis testing using ANOVA further confirms that the choice of feature selection method has a significant impact on model's performance. Interestingly, hyperparameter tuning did not significantly enhance performance, indicating that the default parameters are sufficient to produce an effective model. Feature importance analysis revealed that factors such as the number of companies applied to, job search methods, and the choice of province for job applications have the greatest influence on the model. Additionally, competencies such as internship experience, practicum, english, ethics, discussions, and IT skills substantially impact the waiting time for alumni to secure their first job. These insights suggest that universities can leverage machine learning especially Decision Tree algorithm and tracer study data to refine their curricula, aligning education more closely with job market demands, and potentially reducing the waiting time for prospective graduates to get their first job.

REFERENCES

[1] 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, 2022, doi: 10.14569/IJACSA.2022.0131043.

[2] N. Anggraini, M. J. D. Putra, and N. Hakiem, "Development of an Islamic Higher Education Institution Tracer Study Information System and It's Performance Analysis using ISO/IEC 25010," in *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, IEEE, Nov. 2019, pp. 1–6. doi: 10.1109/CITSM47753.2019.8965356.

[3] 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.

[4] M. A. Febriantono, S. H. Pramono, Rahmadwati, and G. Naghdy, "Classification of multiclass imbalanced data using cost-sensitive decision tree c5.0," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 1, pp. 65–72, 2020, doi: 10.11591/ijai.v9.i1.pp65-72.

[5] 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.

[6] P. H. Gunawan, I. Palupi, Indwiarti, A. A. Rohmawati, and A. T. Hanuranto, "Correlation Analysis of Student's Competencies and Employ-Ability Tracer Study of Telkom University Graduates," 2024, pp. 49–60. doi: 10.1007/978-3-031-53552-9_5.

[7] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 856−866
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5647

[8] T. T. Huynh-Cam, L. S. Chen, and H. Le, "Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance," *Algorithms*, vol. 14, no. 11, Nov. 2021, doi: 10.3390/a14110318.

[9] 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.

[10] B. Saini, G. Mahajan, H. Sharma, and Ziniya, "An Analytical Approach to Predict Employability Status of Students," *IOP Conf Ser Mater Sci Eng*, vol. 1099, no. 1, p. 012007, Mar. 2021, doi: 10.1088/1757-899X/1099/1/012007.

[11] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021, doi: 10.1109/ACCESS.2021.3104357.

[12] Q. R. S. Fitni and K. Ramli, "Implementation of Ensemble Learning and Feature Selection for Performance Improvements in Anomaly-Based Intrusion Detection Systems," in *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, IEEE, Jul. 2020, pp. 118–124. doi: 10.1109/IAICT50021.2020.9172014.

[13] 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.

[14] S. Waskle, L. Parashar, and U. Singh, "Intrusion Detection System Using PCA with Random Forest Approach," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, Jul. 2020, pp. 803–808. doi: 10.1109/ICESC48915.2020.9155656.

[15] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. O'Reilly Media, Inc., 2019.

[16] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput Biol Med*, vol. 136, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

[17] M. Shaheen, T. Zafar, and S. Ali Khan, "Decision tree classification: Ranking journals using IGIDI," *J Inf Sci*, vol. 46, no. 3, pp. 325–339, Jun. 2020, doi: 10.1177/0165551519837176.

[18] K. Gallatin and C. Albon, *Machine Learning with Python Cookbook, 2nd Edition*. O'Reilly Media, Inc., 2023.

[19] I. M. De Diego, A. R. Redondo, R. R. Fernández, J. Navarro, and J. M. Moguerza, "General Performance Score for classification problems," *Applied Intelligence*, vol. 52, no. 10, pp. 12049–12063, Aug. 2022, doi: 10.1007/s10489-021-03041-7.

[20] A. B. Soman, "Gini Impurity vs Gini Importance vs Mean Decrease Impurity | by Aneesha B Soman | Medium." Accessed: Jul. 12, 2024. [Online]. Available: https://medium.com/@aneesha161994/gini-impurity-vs-gini-importance-vs-mean-decrease-impurity-51408bdd0cf1

[21] S. M. Ferdous, S. B. S. Mugdha, and I. Dehzangi, "New Multi-View Feature Learning Method for Accurate Antifungal Peptide Detection," *Algorithms*, vol. 17, no. 6, p. 247, Jun. 2024, doi: 10.3390/a17060247.

[22] E. F. El-Hashash and R. H. A. Shiekh, "A Comparison of the Pearson, Spearman Rank and Kendall Tau Correlation Coefficients Using Quantitative Variables," *Asian Journal of Probability and Statistics*, pp. 36–48, Oct. 2022, doi: 10.9734/ajpas/2022/v20i3425.