Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 450-459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5406

Comparative Assessment of Low Job Competitiveness Among University Graduates Using Naïve Bayes and KNN Algorithms

Ricardo Hamonangan[1], Irma Palupi[2,*], Putu Harry Gunawan[2]

[1] School of Computing, Informatics, Telkom University, Bandung, Indonesia
[2]Center of Excellent Human Centric Engineering (HUMIC), Indonesia
Email: [1]ricardohamonangan@students.telkomuniversity.ac.id, [2,*]irmapalupi@telkomuniversity.ac.id,
[3]phgunawan@telkomuniversity.ac.id
Correspondence Author Email: irmapalupi@ telkomuniversity.ac.id
Submitted: **20/06/2024**; Accepted: **30/06/2024**; Published**: 30/06/2024**

Abstract−Tracer studies investigate the career outcomes of graduates, encompassing job search experiences, employment conditions, and the application of acquired skills post-graduation. These studies are pivotal for universities and colleges to assess graduate success and shape educational policies. This study aims to elucidate the factors contributing to low job competitiveness through the application of classification models like KNN and Naïve Bayes. It also evaluates how competencies developed during university studies impact this scenario. Key issues addressed include the identification of factors causing low job competitiveness and the assessment of competencies trained during university education. Utilizing a dataset comprising two classes and seven features, the KNN method achieved an accuracy of 71.00%, while Naïve Bayes achieved 70.00%. The data set size is 1853 (around 20% of the survey sample) of unemployed alumni. The results indicate that the lack of specific competencies, particularly those related to practical skills and real-world application, is a major factor contributing to low job competitiveness. The results highlight a specific competency as most crucial in the KNN model, whereas different competencies play significant roles in the Naïve Bayes model. Despite variations in competency importance across models, all features significantly contribute to predictions. This research enhances the classification of workforce competitiveness levels within tracer studies and underscores the potential of KNN and Naïve Bayes algorithms to identify factors influencing low job competitiveness. These findings support informed decision-making in academic and career development initiatives, emphasizing the critical influence of university-trained competencies on job market readiness.

Keywords: Tracer Study; University graduates, K-Nearst Neighbors (KNN); Naïve Bayes; Job Competitiveness.

1. INTRODUCTION

Universities play a critical role in equipping students with the tools they need to thrive in the dynamic job market. Traditionally, a university's success has been measured by its graduates' ability to secure employment after graduation. However, recent trends indicate a troubling rise in the number of graduates struggling to find jobs that align with their educational backgrounds. The mismatch between graduates' skills and job market demands has become increasingly pronounced. This discrepancy often leads to underemployment, where graduates take positions that do not require their level of education or expertise, resulting in wasted potential and dissatisfaction. Factors contributing to this issue include the rapid pace of technological change, shifting economic conditions, and evolving employer expectations that outpace the current curricula. Additionally, the saturation of certain job markets with too many qualified candidates exacerbates the problem, making it harder for graduates to stand out. The urgency of this research lies in addressing this concerning phenomenon, aiming to understand and mitigate the factors hindering graduate competitiveness. Without timely intervention, universities risk producing graduates who are ill-prepared for the job market, thereby undermining the value of higher education and contributing to broader economic inefficiencies.

To address this challenge, universities have increasingly relied on tracer studies. These studies systematically assess graduates' employment status, job search experiences, and the application of competencies acquired during their university education [1], [2]. Tracer studies provide invaluable insights that enable universities to adapt to societal changes, meet employer expectations, and continuously evaluate and revise curricula [2]. The job market is constantly evolving due to technological advancements, globalization, and economic shifts. Tracer studies help universities identify emerging skill gaps and adapt their programs accordingly. Employers are constantly seeking graduates with specific skillsets and experiences. Tracer studies provide valuable data on employer needs, allowing universities to tailor their programs to meet those expectations. By analyzing graduate outcomes, universities can identify areas where their programs may fall short and make informed decisions about curriculum revisions.

However, the effectiveness of traditional methods for analyzing tracer study data can be limited. This is where machine learning (ML) models come into play. This research employs ML algorithms, specifically Naïve Bayes and KNN, to analyze tracer study data and identify factors contributing to low job competitiveness among university graduates. ML algorithms can efficiently manage and analyze large datasets collected through tracer studies. This allows universities to identify patterns and trends within the data, revealing factors that might otherwise be overlooked. This deeper understanding can inform the development of targeted interventions to improve graduate competitiveness [3]. By analyzing historical data, ML models can be trained to predict future trends in graduate employment. This allows universities to proactively address potential challenges and prepare students for the job market accordingly. Universities can leverage ML insights to make informed decisions about resource allocation, program development, and career counseling services.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 450-459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5406

Several recent studies have demonstrated the effectiveness of machine learning algorithms in analyzing tracer study data. For instance, a study at STIKOM Bali used the Naïve Bayes algorithm to predict alumni employment waiting times, achieving an accuracy of 48.629% [4]. This highlights the potential of Naïve Bayes, although there is room for improvement. Other studies utilizing Naïve Bayes achieved accuracy rates of 75.33% [5], 87.50% [6], and another study confirmed its effectiveness in predicting alumni employment timelines [7], further solidifying its potential. K-Nearest Neighbors (KNN) is another algorithm that has proven effective. Research has shown its accuracy in predicting on-time graduation status (accuracy: 93.2%) [8]. While comparisons with Naïve Bayes suggest a slight edge for Naïve Bayes in some cases (accuracy: 83.83% vs. 82.34% for KNN) [9], both algorithms demonstrate significant potential.

Building upon the success of previous studies, this research aims to utilize both Naïve Bayes and KNN algorithms to analyze tracer study data and identify the factors contributing to low job competitiveness among university graduates. By analyzing the specific competencies provided by universities, this study seeks to generate valuable insights that can be used to enhance educational and career development strategies

This research offers a two-pronged approach. First, by employing machine learning algorithms, this study delves deeper into the factors hindering graduate competitiveness. This knowledge is crucial for developing targeted interventions to address these specific challenges. Second, the insights gained from this research can be used to inform the development of more effective programs and services that better prepare students for the job market. This includes potentially revising curricula to ensure graduates possess the in-demand skills and experiences sought by employers.

As universities have a vital role to play in ensuring graduate success in the job market, tracer studies, coupled with machine learning analysis, provide a powerful tool for identifying the challenges faced by graduates and developing effective solutions. This research, by utilizing Naïve Bayes and KNN algorithms, aims to contribute to this critical endeavor by shedding light on the factors hindering graduate competitiveness and paving the way for improved educational and career development strategies. The objective of this study is to identify the factors contributing to low job competitiveness among university graduates by analyzing tracer study data using Naïve Bayes and KNN machine learning algorithms. The findings from this research will be used to inform the development of more effective educational and career development strategies for graduates.

2. RESEARCH METHODOLOGY

2.1 Overall Methodology

This research involves creating a procedures to identify factors contributing to low competitiveness. The dataset utilized originates from the Telkom University 2022 Tracer Study. The study compares the accuracy of the K-Nearest Neighbors and Naïve Bayes algorithms, incorporating feature importance permutation. The process flowchart is illustrated in Figure 1.

The flowchart in Figure 1, demonstrates the process of building a machine learning model. The process begins with data collection, which can come from various sources, such as tracer studies that track graduates' career progress. The collected data must undergo cleaning and preparation before being used to train the model. In this case, the tracer study data serves as a specific example. After preparation, the data is divided into two groups: training data and testing data. The training data is used to teach the model by identifying patterns that will help it make predictions. The testing data is then used to evaluate the model's performance on new, unseen data. To ensure the model's accuracy is reliable, validation techniques such as K-Fold Cross Validation are employed during the validation stage. Finally, the model's overall performance is assessed using metrics like accuracy. Additionally, an analysis of feature importance is conducted to understand which factors most significantly influence the model's predictions.

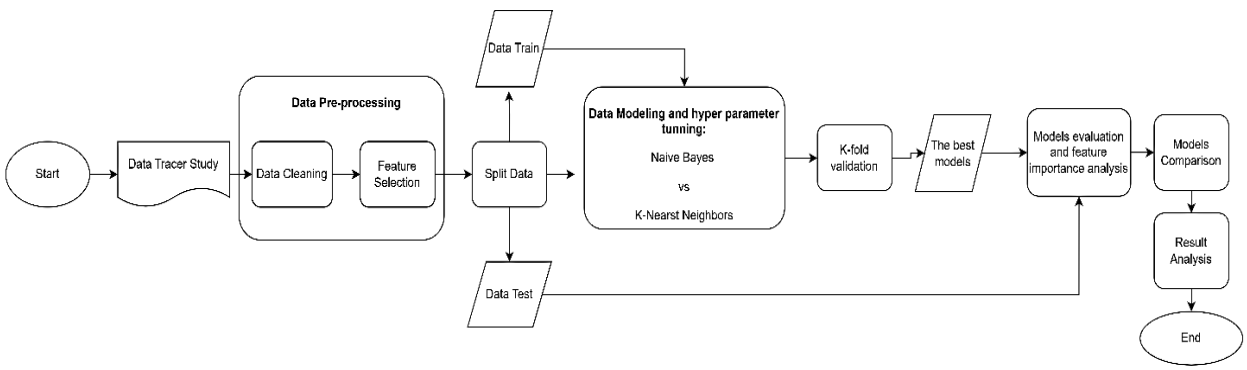

Figure 1. Research Stages.

2.2 Tracer Study Data

Tracer study is a research method carried out by universities to track and analyze the success of alumni in obtaining jobs after graduation [10]. This data or information can help universities improve their curriculum and learning

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 450-459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5406

systems. In other words, "Graduate Surveys" or "Alumni Research" is a university effort to study and evaluate the quality of graduates in terms of their employment outcomes [11]. Additionally, tracer studies also influence the accreditation of a university, which is why universities routinely conduct alumni tracing studies.

At the data collection stage of the tracer study, datasets are gathered. The dataset used in this study is the 2022 Telkom University alumni tracer dataset. This dataset includes 8 columns and 1852 rows of data representing alumni with non-working status. The 8 columns consist of: (1) Ethical , (2) Skill Based on Field of Science, (3) Time Management, (4) Information Technology, (5) Communication, (6) Teamwork, (7) Self Development, and (8) Job competitiveness.

To compute a student's job competitiveness level, we use five key indicators: A_1 :=Why don't you look for a job? , A_2:= Looking for work in the last 4 weeks?, A_3 :=Submission Frequency, A_4:=Response Frequency, and A_5 :=Interview Frequency. Each of these indicators is first scaled to a range of $[0,1]$ to ensure uniformity in measurement. This scaling allows us to handle the data in a standardized format where 0 represents the minimum and 1 represents the maximum value observed for each indicator. For each indicator, we assign a weight w_i which signifies the portion of its contribution to the overall competitiveness level. The weights also lie within the $[0,1]$ range, ensuring that the combined weight of all indicators sums to 1. The weights reflect the relative importance of each indicator based on empirical analysis or expert judgment. The job competitiveness level (JCL) for a student is then computed using the formula (1).

$$JCL = \sum_{i=1}^{5} w_i A_i \quad (1)$$

Indicator A_1 captures the reasons behind a student's inactivity in job searching. The responses are normalized to a [0,1] scale where 1 might represent the most valid reason for not searching (e.g., pursuing further education) and 0 might represent less valid reasons, for instance lack of interest. A_3 measures the number of job applications submitted by the student within a given period. A_4 measures how frequently the student receives responses to their job applications. Higher frequencies are scaled closer to 1, while lower frequencies are closer to 0. A_5 measures the number of interviews the student has been invited to. Like the other indicators, it is scaled to [0,1], with 1 representing the highest observed interview frequency. By combining these scaled indicators with their respective weights, we obtain a composite score that represents the student's job competitiveness level. This method allows for a nuanced and balanced assessment that takes into account various aspects of job search behavior and outcomes.

2.3 Preprocessing

At the pre-processing stage involves several steps of the EDA. Exploratory Data Analysis (EDA) is a statistical approach developed by John Tukey in 1977 [12]. Explorational data analysis is an important process in conducting early investigations into data to find patterns, find anomalies such as outliers, and test early hypotheses and assumptions with some statistics and visual representations [13]. EDA is a tool used to better understand datasets and prepare them as well as possible for the implementation of machine learning algorithms.

The data exploration steps undertaken in this study include descriptive statistics, data visualization, handling of missing values, correlation analysis, exploration of categorical variables, outlier analysis, and data transformation. Descriptive stats are used to provide a statistical summary for each variable present in the dataset, such as mean and standard deviations. Data visualization involves creating graphs and plot to describe data distribution using barchart and boxplot to help understand data patterns. Further, the detection and handling of missing values aims to verify the presence of missed values in the datasets and design strategies for dealing with those lost values. The exploration of categorical variables is performed when the dataset has a category variable, with the aim of examining the distribution of the category and the relationship between the category variables and other variables. Outlier analysis aims to identify and deal with extreme or outlier values in a data set. Some lines that did not match the class were deleted to avoid machine confusion. Finally, data transformation is done with normalization or standardization for use in naive bayes and KNN classifications, thus facilitating data interpretation and producing optimal classification results.

2.4 Spearman's Rank Corelation

Spearman's Rank Corelation will be used in this study to find out that there is a relationship between the competitiveness column of work and the competence column. Calculations for the correlation of the competitive column and the compensation column are shown on the equation (2).

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

Where Rs is the value of Spearman's rank correlation coefficient. This coefficient is calculated based on the rating of values, not the actual value, thus making it a non-parametric correlation measure [14]. Then d is the difference between the rankings of variable 1 to other variables, and n is the sum of the data pairs.

2.5 Naïve Bayes

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 450-459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5406

Naïve Bayes is a simple probabilistic classification method that calculates the probability of a group of values by adding up the frequencies and combinations of values from a given dataset [15]. This method is based on the Bayes theorem which predicts future chances based on previous experience. This algorithm assumes that the object attribute is independent [16]. Mathematically, Bayes's theorem can be expressed in equation (3).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

Where $P(A|B)$ is the probability of event A occurring if event B has occurred. $P(A \cap B)$ is the probability of both events A and B occurring simultaneously, and P(B) is the probability of event B occurring. By dividing P(A∩B) by P(B), we get the probability of event A occurring if event B has already occurred.

Naive Bayes is used to predict job competitiveness based on features in the Tracer Study dataset. Once the data is divided into a training set and a testing set, the Naive Bayes model is trained using the training data. The training process involves calculating the posterior probability of each job competitiveness class, taking into account the observed feature values. Using Bayes' Theorem, the model can then estimate the probability that a sample falls into a certain class, based on its features.

2.6 K-Nearst Neighbors (KNN)

The K-Nearest Neighbors (KNN) method is used to classify data based on training datasets by considering k nearest neighbors [17]. The k value refers to the number of nearest relatives to be considered in the classification process [18]. Mathematically, classification using K-Nearest Neighbors (KNN) can be expressed in the equation (4).

$$\hat{y} = arg\,max\, y \sum_{i=1}^{k} I(yi) \quad (4)$$

Where $\hat{y}$ is the predicted class label for point X, I(yi) is an indicator function worth 1 if yi is class and 0 if not, and argmax returns the value that gives the largest amount.

2.7 K-Fold Cross Validation

K-Fold Cross Validation (k-FCV) is a validation technique used in machine learning to evaluate a model and measure its performance more accurately [19]. K-fold cross validation helps reduce bias or eliminate the possibility of a neutral or unfair tendency or view in an assessment that may occur when a data set is randomly divided into training and testing data. With k- FCV, all data will be used both as training and test data, so that the model can be evaluated more comprehensively.

In this research, K-Fold Cross Validation is used to evaluate the performance of the Naive Bayes and K-Nearest Neighbors (KNN) prediction models. This technique divides the dataset into 5 equal subsets, then trains the model 5 times, each time using 4 subsets as training data and 1 subset as testing data. This process ensures each subset is used once as testing data and four times as training data, providing a more accurate assessment and reducing the possibility of overfitting. The results of each iteration are calculated and averaged to obtain an overall evaluation value. The use of K-Fold Cross Validation helps in ensuring the resulting model has good performance and strong generalization on never-before-seen data.

2.8 Evaluation

On my evaluation I used the confusion matrix. The confusion matrix describes the number of true predictions (True Positive and True Negative) as well as the amount of false forecasts (False Positive, False Negative). This is very important because it provides detailed information about the performance of the classification model in classifying data [20].

Tabel 1. Confusion Matrix

Classsification	Actually Positive (+)	Actually Negative (-)
Predicted Positive (+)	True Positive (TP)	False Positive (FP)
Predicted Negative (-)	False Negative (FN)	True Negative (TN)

Based on confusion matrix in Table 1, the metrics for model performance can be represented by several scores such as Accuracy, Precision, Recall and F1 scores, with the formulas are given by (5-8).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (8)$$

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 450-459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5406

Accuracy is a metric that shows how many true predictions an algorithm produces. However, when the distribution of data is unbalanced, high accuracy achievements with this method do not necessarily reflect its ability to distinguish different categories effectively and efficiently [21].

2.9 Feature Importance

Feature importance for KNN and Naïve Bayes in classification provides insights into how each feature contributes to the model's decision making process. However, the methods for determining feature importance differ between these two algorithms due to their distinct operational principles.
KNN is a non-parametric, instance-based learning algorithm that classifies a data point based on the majority class of its k-nearest neighbors. To determine feature importance in KNN, permutation feature importance is commonly used. First, the initial accuracy of the KNN model is calculated on the test dataset. Then, for each feature, its values are randomly shuffled among the data points, breaking the relationship between the feature and the target variable while keeping the values of other features unchanged. The model's accuracy is recalculated on the test dataset with the permuted feature. The feature importance is computed as the decrease in accuracy caused by permuting the feature. The more significant the drop in accuracy, the more important the feature is considered to be. This can be formulated as:

$$Feature\ Importance(i) = Accuracy_{original} - Accuracy_{permutedfeature\ i}. \quad (9)$$

This method helps understand how crucial each feature is for the KNN model's predictive performance.

Naïve Bayes classifiers assume independence between features given the class label and use the Bayes theorem for classification. Feature importance in Naïve Bayes can be determined through various approaches, including mutual information and likelihood ratio tests. Here, we use the approach using mutual information. Mutual information measures the amount of information obtained about one random variable through another random variable. For feature importance, it quantifies how much knowing the value of a feature reduces uncertainty about the class label. Mathematically, the mutual information $I(X_i; Y)$ between a feature X_i and the target variable Y is given by formula (10).

$$I(X_i; Y) = \sum_{x_i,y} P(x_i, y)\ \frac{logP(x_i,y)}{P(x_i)P(y)} \quad (10)$$

where $P(x_i, y)$ is the joint probability distribution of X_i and Y, and $P(x_i)$ and $P(y)$ are the marginal distributions of X_i and Y, respectively. Features with higher mutual information values are considered more important, as they provide more information about the class label.

3. RESULT AND DISCUSSION

This section encompasses three distinct scenarios aimed at achieving optimal prediction accuracy for job competitiveness classifications using statistical exploratory analysis and implementing two classification methods. In Scenario I, a single-test assessment determines the best data classification from four job competitiveness classes or two classes. Scenario II compares the K-Nearest Neighbor (KNN) and Naïve Bayes methods across two job competitiveness classes, utilizing the three highest correlated features and all available features. Furthermore, Scenario III tests two job competitiveness classes using KNN and Naïve Bayes methods with k-fold cross-validation (kf=5). These scenarios not only aim to identify the best classes and features for achieving optimal prediction accuracy but also incorporate feature importance analysis to pinpoint factors influencing job competitiveness levels.

3.1 Dataset Infomation

This research utilizes a dataset comprising eight columns, of which seven are categorical and one is numeric. The categorical data is transformed into numeric for analysis.

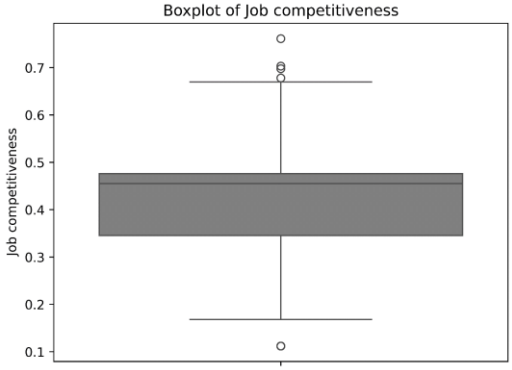

Figure 2. Boxplot of Job Competitiveness

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 450-459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5406

In Figure 2, the data distribution of computed score JCL is illustrated, allowing us to establish categorical intervals for both 4-class and 2-class classifications. For the 4-class classification, the intervals are defined as follows: 0 - 0.20 represents the "Very Poor" category, 0.21 - 0.40 corresponds to the "Poor" category, 0.41 - 0.60 signifies the "Moderate" category, and 0.61 - 0.80 denotes the "Strong" category. Meanwhile, the 2-class classification is divided into 0 - 0.40 for the "Poor" category and 0.41 - 0.80 for the "Strong" category. The frequency distribution of each class partition is given by Figure 4.

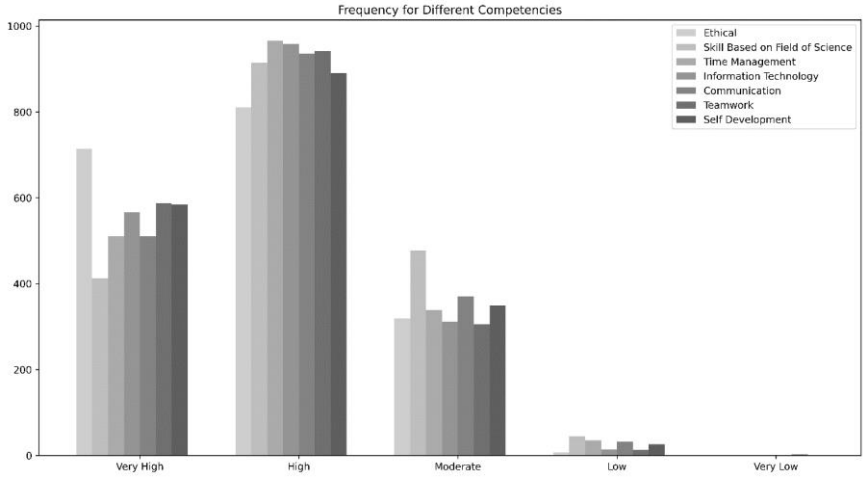

Figure 3. Barplot of Competencies

Figure 3. shows how many graduates scored in different levels for various skills. Skills like ethics, communication, and teamwork were rated highly by most graduates. "High" was the most common rating, with 800 to 1000 graduates scoring in that category. Ethics was rated even higher, with around 650 graduates scoring "Very High" in ethics. There were fewer graduates rated as "Moderate" or lower in any skill. Overall, most graduates seemed to have strong skills and high ethical standards.

Table 2. Dataset Score Statistics

Competency features	Correlation	Mean	Std Dev
Self Development	0.641	1.902	0.745
Communication	0.636	1.960	0.741
Skill Based on Field of Science	0.631	2.085	0.759
Time Management	0.615	1.947	0.732
Teamwork	0.602	1.866	0.712
Information Technology	0.595	1.878	0.703
Ethical	0.589	1.795	0.732
Job competitiveness		0.399	0.114

Table 2 analyzes the proficiency in specific skills and their correlation with alumni's high motivation to find employment. All skills demonstrated a positive association with alumni motivation, with correlations ranging from 0.589 to 0.641. "Self Development" emerged as the most crucial skill (0.641), closely followed by Communication and Skills in their Field of Science. The average score for each skill ranged from 1.795 to 2.085, with standard deviations between 0.703 and 0.759. This highlights that skills acquired during college, particularly in self-development, communication, and field-specific areas, significantly enhance alumni's motivation to secure employment.

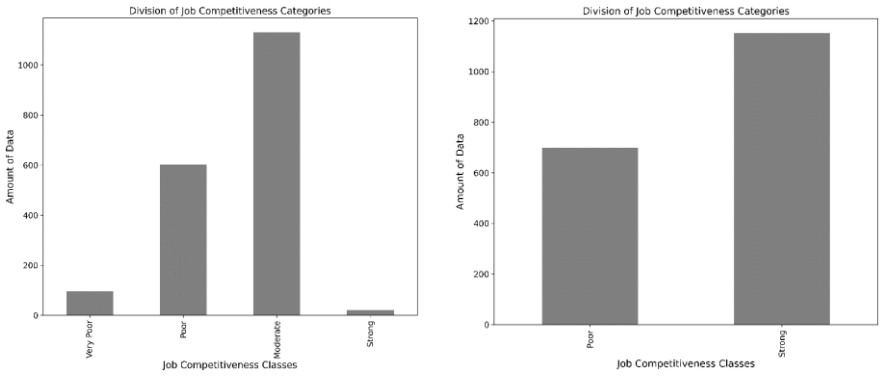

Figure 4. Four-class and two-class job competitiveness class

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 450-459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5406

Based on the information in Figure 4, the number of job competitiveness classes can be divided into four categories: strong, moderate, poor, and very poor. The data show that the strong category contains a total of 21 data points, while the moderate category has a larger number of data points, specifically 1131. The poor category contains 603 data points, and the very poor category contains 96 data points. Additionally, there is a separation into two broader categories: strong and weak. The strong category, in this case, comprises 1152 data points, while the weak category comprises 699 data points. This information provides a clear picture of the distribution and proportion of data within each job competitiveness category, thus providing a solid basis for further analysis.

3.2 Scenario I Testing Results

In this scenario, a single test is conducted to determine the best data classification from both four job competitiveness classes and two job competitiveness classes. The 4 classes include strong, moderate, poor, and very poor, while the 2 classes include strong and poor. The Naïve Bayes Method and the K-Nearest Neighbors Method are employed for this classification.

Table 3. KNN Evaluation Score with k = 15 Class 4 and Class 2

Class	4 Class				2 Class	
	Strong	Moderate	Poor	Very Poor	Strong	Poor
Accuracy			0.67		0.71	
Precision	0.00	0.70	0.60	0.00	0.72	0.68
Recall	0.00	0.88	0.44	0.00	0.87	0.45
F1-Score	0.00	0.78	0.50	0.00	0.79	0.54

Table 4. Naïve Bayes Evaluation Score 4 Class and 2 Class

Class	4 Class				2 Class	
	Strong	Moderate	Poor	Very Poor	Strong	Poor
Accuracy			0.58		0.70	
Precision	0.00	0.72	0.54	0.08	0.73	0.62
Recall	0.00	0.68	0.50	0.20	0.81	0.50
F1-Score	0.00	0.70	0.52	0.12	0.77	0.56

Based on Table 3 and Table 4, the evaluation scores for improving accuracy in this study using the KNN Method and the Naïve Bayes Method with a data split of 20:80 for test and training data, the best accuracy was obtained for the 2-class scenario. The KNN method achieved an accuracy of 0.71, with a precision of 0.72 for the strong class, a recall of 0.87 for the strong class, an F1 score of 0.79 for the strong class, a precision of 0.68 for the poor class, a recall of 0.45 for the poor class, and an F1 score of 0.54 for the poor class. The Naïve Bayes method achieved an accuracy of 0.70, with a precision of 0.73 for the strong class, a recall of 0.81 for the strong class, an F1 score of 0.77 for the strong class, a precision of 0.62 for the poor class, a recall of 0.50 for the poor class, and an F1 score of 0.56 for the poor class. Subsequently, Scenario II testing was conducted with 2 classes using the KNN and Naïve Bayes methods by comparing the three highest correlation features with all available features. Scenario III testing was conducted with 2 classes using the KNN and Naïve Bayes methods with k-fold cross-validation (kf = 5).

3.3 Scenario II Testing Results

In Scenario II testing, the best model for job competitiveness, which includes 2 classes (strong and poor), is used. The KNN and Naïve Bayes methods are employed with a data split of 80:20 for training and testing data, comparing the three highest correlation features with the full set of seven features.

Table 5. Evaluation Score of 2 Classes using K-Nearest Neighbors Algorithm with 3 and 7 Features

Class	Feature	Precision	Recall	F1-Score	Accuracy
Strong	3	0.65	0.91	0.76	0.65
Poor		0.64	0.25	0.35	
Strong	7	0.72	0.87	0.79	**0.71**
Poor		0.68	0.45	0.54	

Table 6. Evaluation Score of 2 Classes using Naïve Bayes Algorithm with 3 and 7 Features

Class	Feature	Precision	Recall	F1-Score	Accuracy
Strong	3	0.66	0.91	0.76	0.66
Poor		0.66	0.28	0.39	
Strong	7	0.73	0.81	0.77	**0.70**
Poor		0.62	0.50	0.56	

Table 5 and Table 6 present the evaluation scores for the 2-class scenario (strong and poor) using the K-Nearest Neighbors (KNN) algorithm and the Naïve Bayes algorithm, respectively. The performance of each algorithm is

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 450-459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5406

assessed with two sets of features: the three highest correlation features and the full set of seven features. Table 5 shows the evaluation scores using the K-Nearest Neighbors algorithm. With 3 features, the strong class achieved a precision of 0.65, a recall of 0.91, an F1-score of 0.76, and an overall accuracy of 0.65. The poor class, using the same 3 features, achieved a precision of 0.64, a recall of 0.25, and an F1-score of 0.35. When using all 7 features, the strong class achieved a precision of 0.72, a recall of 0.87, an F1-score of 0.79, and an overall accuracy of 0.71. The poor class, with 7 features, achieved a precision of 0.68, a recall of 0.45, and an F1-score of 0.54. Table 6 displays the evaluation scores using the Naïve Bayes algorithm. With 3 features, the strong class achieved a precision of 0.66, a recall of 0.91, an F1-score of 0.76, and an overall accuracy of 0.66. The poor class, using the same 3 features, achieved a precision of 0.66, a recall of 0.28, and an F1-score of 0.39. When using all 7 features, the strong class achieved a precision of 0.73, a recall of 0.81, an F1-score of 0.77, and an overall accuracy of 0.70. The poor class, with 7 features, achieved a precision of 0.62, a recall of 0.50, and an F1-score of 0.56.

The results show that using all seven features generally improves the performance of both algorithms, particularly for the strong class. The KNN algorithm demonstrates slightly better performance compared to Naïve Bayes.

3.4 Scenario III Testing Results

Scenario III focuses on evaluating job competitiveness using the KNN and Naïve Bayes methods with k-fold cross-validation (kf=5). This methodological approach differs from previous scenarios in that it systematically partitions the data into multiple subsets (folds) for training and validation. Each fold serves as both training and validation data across different iterations, ensuring robust evaluation of the model's performance and generalizability. This approach provides a more comprehensive assessment compared to earlier scenarios, which used simple train-test splits.

Table 7. Evaluation Score of 2 Classes using K-Nearest Neighbors Algorithm with kf = 5

Class	kf	Precision	Recall	F1-Score	Accuracy
Strong	1	0.71	0.91	0.80	0.71
Poor		0.72	0.37	0.49	
Strong	2	0.70	0.92	0.79	0.71
Poor		0.72	0.36	0.48	
Strong	3	0.66	0.86	0.74	0.64
Poor		0.54	0.27	0.36	
Strong	4	0.69	0.88	0.77	0.68
Poor		0.63	0.34	0.44	
Strong	5	0.71	0.89	0.80	0.72
Poor		0.72	0.43	0.54	
Avarange					0.69

Table 8. Evaluation Score of 2 Classes using Naïve Bayes Algorithm with kf = 5

Class	kf	Precision	Recall	F1-Score	Accuracy
Strong	1	0.74	0.82	0.78	0.71
Poor		0.64	0.51	0.57	
Strong	2	0.73	0.83	0.77	0.70
Poor		0.63	0.49	0.55	
Strong	3	0.66	0.76	0.71	0.61
Poor		0.48	0.36	0.41	
Strong	4	0.67	0.70	0.73	0.63
Poor		0.41	0.36	0.42	
Strong	5	0.65	0.79	0.71	0.60
Poor		0.44	0.28	0.34	
Avarange					0.65

Table 7 and Table 8 present a comparison of the performance of two classification methods, K-Nearest Neighbor (KNN) and Naive Bayes, in predicting the level of graduation strength of graduates of a study program. Both methods were evaluated using the five-fold cross-validation method. The results show that KNN has slightly better accuracy than Naive Bayes in predicting gradient strength levels. KNN accuracy ranges from 0.64 to 0.72, while Naive Bayes ranges from 0.60 to 0.71.

Specifically, the KNN evaluation results for folds 1 to 5 show accuracies of 0.71, 0.71, 0.64, 0.68, and 0.72 respectively. Meanwhile, for Naive Bayes, the accuracies for folds 1 to 5 are 0.71, 0.70, 0.61, 0.63, and 0.60. From these results, it can be concluded that KNN tends to provide more accurate predictions in classifying the graduation strength level.

However, it is important to note that accuracy is not the only metric to consider when choosing a suitable classification method. Other metrics such as precision, recall, and f1-score are also important for a comprehensive evaluation. In this case, although KNN has higher accuracy, Naive Bayes shows better performance in some cases

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 450-459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5406

with more balanced precision and recall values. Therefore, the choice of classification method should consider various factors and can be tailored to the specific needs of the case under study.

3.5 Feature Importance Analysis

In analyzing feature importance for both KNN and Naïve Bayes models, we gain insights into which competencies significantly influence job competitiveness levels among university graduates.

Table 9. Feature Importance KNN

Competency features	Importance
Self Development	0.0135
Communication	0.0472
Skill Based on Field of Science	0.0194
Time Management	0.0482
Teamwork	0.0719
Information Technology	0.0541
Ethical	0.0186

In Table 9, Teamwork emerges as the most influential competency with an importance value of 0.0719 in the KNN model, suggesting its critical role in determining job competitiveness. Information Technology (0.0541) and Time Management (0.0482) also demonstrate substantial importance, emphasizing the importance of technological skills and effective time management practices in enhancing competitiveness. Communication follows closely with an importance value of 0.0472, highlighting its significant role despite not being the most dominant feature. Skill Based on Field of Science and Ethical have lower importance values of 0.0194 and 0.0186, respectively, indicating their smaller yet still meaningful contributions to the model. Self Development shows the lowest importance value at 0.0135, suggesting its lesser impact compared to other competencies. Overall, understanding these importance values helps identify which competencies are crucial for enhancing job competitiveness among graduates.

Table 10. Feature Importance Naïve Bayes

Competency features	Importance
Self Development	0.218
Communication	0.116
Skill Based on Field of Science	0.165
Time Management	0.126
Teamwork	0.068
Information Technology	0.106
Ethical	0.199

In Table 10 for the Naïve Bayes model, Self Development stands out with the highest importance value of 0.218, indicating its pivotal role in predicting job competitiveness levels. Ethical follows closely with an importance value of 0.199, underscoring its significant contribution. Skill Based on Field of Science (0.165) and Time Management (0.126) also exhibit moderate importance, suggesting their essential roles in determining competitiveness. Communication (0.116) and Information Technology (0.106) contribute moderately to the model, while Teamwork (0.068) shows a smaller but notable influence. These findings provide actionable insights into which competencies are most critical for improving job competitiveness among university graduates.

Understanding these feature importance values enables stakeholders to prioritize interventions and strategies aimed at enhancing specific competencies that drive higher job competitiveness. For instance, focusing on improving skills related to Teamwork, Information Technology, and Time Management could lead to significant improvements in graduates' employability and career success. Similarly, emphasizing the development of Self Development and Ethical competencies can foster a professional outlook that resonates positively in the job market. By leveraging these insights, educational institutions and policymakers can tailor curriculum enhancements and career development programs to better prepare graduates for competitive job opportunities.

4. CONCLUSION

This study examined factors affecting university graduates' job competitiveness using machine learning. The analysis showed promising results, with both tested algorithms effectively identifying key influences. An important finding is that considering a wider range of factors appears to be more accurate than focusing on just a few. Additionally, the study suggests that one algorithm, KNN, might be slightly better suited for this type of analysis. This research looked at Telkom University graduates in 2022 to see what factors influenced their job prospects. They tried two different methods (KNN and Naive Bayes) to analyze the data. They found that using a simpler rating system (2 categories) for job competitiveness and considering more factors (7 instead of 3) led to more accurate results (around 70% accuracy). Overall, the study suggests that various skills, including teamwork, communication, and self-development, all play a

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 450-459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5406

role in a graduate's job success. This research has significant implications for universities. By using machine learning on graduate data, universities can gain valuable insights to improve their programs and services. This could lead to better preparation for the job market and increased graduate employability. Future research can explore ways to incorporate even more data and utilize more advanced techniques to gain an even deeper understanding of this complex issue. Ultimately, this line of research can be a powerful tool for universities to ensure their graduates are successful in the competitive workforce.

REFERENCES

[1] Divisi Karir dan Hubungan Alumni Lembaga Pengembangan Kemahasiswaan dan Alumni, "Laporan Tracer Study 2021/2022," pp. 1-564. Diakses pada: 1 April 2024.

[2] A. C. Albina and L. P. Sumagaysay, "Employability tracer study of Information Technology Education graduates from a state university in the Philippines," *Soc. Sci. Humanit. Open*, vol. 2, no. 1, p. 100055, 2020, doi: 10.1016/j.ssaho.2020.100055.

[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," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 8, pp. 169–174, 2022, doi: 10.14569/IJACSA.2022.0130820.

[4] I. M. B. Adnyana, "Implementasi Naïve Bayes Untuk Memprediksi Waktu Tunggu Alumni Dalam Memperoleh Pekerjaan," *Semin. Nas. Teknol. Komput. Sains*, pp. 131–134, 2020.

[5] K. C. Piad, M. Dumlao, M. A. Ballera, and S. C. Ambat, "Predicting IT employability using data mining techniques," *2016 3rd Int. Conf. Digit. Inf. Process. Data Mining, Wirel. Commun. DIPDMWC 2016*, no. January 2014, pp. 26–30, 2016, doi: 10.1109/DIPDMWC.2016.7529358.

[6] M. Brilliant, I. A. Nurhasanah, and ..., "PERBANDINGAN ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR UNTUK KLASIFIKASI WAKTU TUNGGU ALUMNI DALAM MEMPEROLEH PEKERJAAN (Study Kasus SMKS PGRI 2 PRINGSEWU)," *SEAT J. ...*, pp. 15–20, 2023.

[7] T. D. Aulia, Y. Siagian, and P. Putri, "PENERAPAN ALGORITMA NAIVE BAYES UNTUK PREDIKSI WAKTU TUNGGU ALUMNI MENDAPATKAN PEKERJAAN PADA LEMBAGA Keyword : data mining ; naive bayes ; waiting time Kata Kunci : data mining ; naive bayes ; waktu tunggu PENDAHULUAN Perkembangan ilmu pengetahuan di bid," vol. 3, no. 2, pp. 85–92, 2023.

[8] T. Asril, "Prediction of Students Study Period using K-Nearest Neighbor Algorithm," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 6, pp. 2585–2593, 2020, doi: 10.30534/ijeter/2020/60862020.

[9] M. A. Maricar and Dian Pramana, "Perbandingan Akurasi Naïve Bayes dan K-Nearest Neighbor pada Klasifikasi untuk Meramalkan Status Pekerjaan Alumni ITB STIKOM Bali," *J. Sist. dan Inform.*, vol. 14, no. 1, pp. 16–22, 2019, doi: 10.30864/jsi.v14i1.233.

[10] D. I. Purnama, R. L. Islami, L. Sari, and P. R. Sihombing, "Analisis Klasifikasi Data Tracer Study Dengan Support Vector Machine Dan Neural Network," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 2, pp. 46–52, 2021, doi: 10.47970/siskom-kb.v4i2.191.

[11] R. Efiyanna, S. P. Hastono, and A. Syafiq, "TRACER STUDY : THE ALIGNMENT OF WORK TYPES WITH THE ORIGIN OF ALUMNI SPECIALIZATION OF FKM UI Departemen Biostatistik dan Kependudukan Fakultas Kesehatan Masyarakat Universitas Indonesia Departemen Gizi Masyarakat Fakultas Kesehatan Masyarakat Universitas," vol. 10, no. 1, pp. 26–34, 2019.

[12] E. Camizuli and E. J. Carranza, "Exploratory Data Analysis," no. 3, 2018, doi: 10.1002/9781119188230.saseas0271.

[13] V. Da Poian *et al.*, "Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry," *Front. Astron. Sp. Sci.*, vol. 10, no. May, pp. 1–17, 2023, doi: 10.3389/fspas.2023.1134141.

[14] K. Ali and A. Al-Hameed, "Spearman's correlation coefficient in statistical analysis," *Int. J. Nonlinear Anal. Appl*, vol. 13, no. May 2021, pp. 2008–6822, 2022, [Online]. Available: http://dx.doi.org/10.22075/ijnaa.2022.6079.

[15] I. Loelianto, M. S. S. Thayf, and H. Angriani, "Implementasi Teori Naive Bayes Dalam Klasifikasi Calon Mahasiswa Baru Stmik Kharisma Makassar," *SINTECH (Science Inf. Technol. J.*, vol. 3, no. 2, pp. 110–117, 2020, doi: 10.31598/sintechjournal.v3i2.651.

[16] M. R. Qisthiano, T. B. Kurniawan, E. S. Negara, and M. Akbar, "Pengembangan Model Untuk Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 3, p. 987, 2021, doi: 10.30865/mib.v5i3.3030.

[17] N. K. K. Ardana *et al.*, "Perbandingan Metode KNN, Naive Bayes, dan Regresi Logistik Binomial dalam Pengklasifikasian Status Ekonomi Negara," *Jambura J. Math.*, vol. 5, no. 2, pp. 404–418, 2023, doi: 10.34312/jjom.v5i2.21103.

[18] F. Rizka Yudana, M. Suyanto, and A. Nasiri, "Model Klasifikasi Untuk Menentukan Kesiapan Kerja Mahasiswa Dan Kelulusan Tepat Waktu Dengan Metode Machine Learning," vol. 1, no. 1, pp. 1–12, 2023, doi: 10.37680/ijitech.v1i1.xx.

[19] M. R. Wayahdi, D. Syahputra, S. Hafiz, and N. Ginting, "Evaluation of the K-Nearest Neighbor Model With K-Fold Cross Validation on Image Classification," *Infokum*, vol. 9, no. 1, pp. 1–6, 2020, [Online]. Available: http://infor.seaninstitute.org/index.php/infokum/index.

[20] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[21] S. Sharma and K. Guleria, "A Deep Learning based model for the Detection of Pneumonia from Chest X-Ray Images using VGG-16 and Neural Networks," *Procedia Comput. Sci.*, vol. 218, pp. 357–366, 2022, doi: 10.1016/j.procs.2023.01.018.