IT Journal Research and Development (ITJRD)
Vol.10, No.1, August 2025, E-ISSN : 2528-4053 | P-ISSN : 2528-4061
DOI : 10.25299/itjrd.2025.20594

Analysis of Skill Requirements in The Information Technology Job Market on Jobstreet Indonesia Using Machine Learning Algorithms

Muhammad Rifqi Majid[1], Hery Dian Septama[2], Mahendra Pratama[3]
Informatic Engineering, Department of Engineering, Lampung University
rifqimajid72@gmail.com[1], hery@eng.unila.ac.id[2], mahendra.pratama15@eng.unila.ac.id[3]

Article Info

Article history:

Received Des 28, 2024
Revised Mar 11, 2025
Accepted Jul 18, 2025

Keyword:

JobStreet
Data mining
CRISP-DM
Skills
Classification

ABSTRACT

With the rapid advancement of information technology, the demand for skills in this field is growing significantly. Jobstreet provides various qualifications, including jobs in information technology. Therefore, classification is necessary to identify skill trends. Job vacancy data from Jobstreet can be utilized as raw data to generate a comprehensive classification of information technology (IT) skills. This research focuses on exploring machine learning algorithms in the context of classification to analyze skill trends. It also compares model accuracy in data classification, visualizes data mining results, and identifies sub-categories and skill trends required by the industry. The study adopts the CRISP-DM framework and employs k-Nearest Neighbor (KNN), Naïve Bayes Classifier (NBC), and Support Vector Machine (SVM) algorithms. The research methodology includes data collection through scraping techniques, data processing using machine learning algorithms (tokenization, stopword removal, stemming, n-gram visualization, and word embeddings), and data visualization through Looker Studio. The results show that the SVM model excels with an accuracy of 86.75%, followed by KNN at 83.33%, and NBC at 79.49%. The most in-demand job sub-categories include Business/System Analyst (34.1%), Network & System Administration (22.6%), and Developer/Programmer (8%). This study demonstrates the superiority of the SVM algorithm over other algorithms, highlighting its strong performance in text classification tasks.

Corresponding Author:

Muhammad Rifqi Majid
Department of Engineering
Lampung University
Bandarlampung, Lampung, Indonesia
Email: rifqimajid72@gmail.com

1. INTRODUCTION

In the digital era, information technology (IT) plays a crucial role across various industries. This development has significantly increased the demand for skilled IT professionals. As a result, job vacancy data—such as that from Jobstreet—has become a valuable resource for identifying skill trends in the market. Jobstreet, the largest job portal in Southeast Asia, offers a wide range of job categories. It also streamlines the qualification-matching process between job seekers and employers,

particularly in the IT sector. Due to the continuous rise in IT-related job demand, Jobstreet's data provides rich insights into in-demand skills.

Current approaches often fail to capture the complexity and diversity of IT skills, indicating a need for more robust techniques. To address this challenge, innovative methods are required to enhance classification performance and uncover deeper insights from the data.

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM), a systematic methodology widely used in data mining research [1-2]. CRISP-DM consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [2-3]. Exploratory Data Analysis (EDA) techniques are integrated during the data understanding phase to uncover measurable insights [4]. A study by Caetano et al. [5], demonstrated the effectiveness of CRISP-DM in data mining applications. However, their research was limited to evaluating the methodology in the context of regression, not text classification.

Recent studies have compared machine learning algorithms such as *Naive Bayes Classifier* (NBC), *k-Nearest Neighbor* (KNN), and *Support Vector Machine* (SVM). For instance, Solekhah et al. [6], found that KNN achieved the highest accuracy of 82%, while Naive Bayes reached 79% [7-8]. Similarly, Hermawan in 2023 demonstrated that Support Vector Machine (SVM) effectively modeled text classification in the context of sentiment analysis, achieving an accuracy of 73% [9]. However, a common challenge in text classification lies in addressing data imbalance, where one class significantly outweighs others in size. This imbalance can reduce both the accuracy and reliability of the model [10].

Text representation techniques such as TF-IDF [11-13] , along with advanced NLP methods including tokenization, stopword removal, stemming, and n-gram visualization, have been proven to enhance feature quality in text classification [14-15]. For example, Khotimah et al. implemented EDA processes such as tokenization, stopword removal, stemming, n-gram visualization, and word embedding to extract relevant features from text data [16]. However, the application of these techniques in the classification of IT skills based on job vacancy data remains underutilized.

This study explicitly aims to address this gap by developing an accurate and reliable text classification model for IT skills using job vacancy data. It focuses on tackling data imbalance using machine learning algorithms such as NBC, KNN, and SVM [8]. Naïve Bayes (NBC) is known for its efficiency and robustness in high-dimensional data such as text. KNN offers simplicity and performs well in capturing local structure in data, while SVM has shown strong performance in handling non-linear separability and is less sensitive to data imbalance when combined with appropriate kernel functions. These characteristics make the three algorithms suitable candidates for evaluating the classification of IT skills from job vacancy data. The study also explores the use of advanced NLP techniques, including EDA and n-gram analysis, to improve model performance. Furthermore, the research aims to identify specific subcategories of IT skills, offering more detailed insights than general classifications. Using the CRISP-DM methodology, this study systematically integrates data collection, modeling, and visualization to deliver actionable insights [10]. To achieve these goals, the research seeks to answer the following questions: How effective are various algorithms in classifying IT skills?, To what extent do NLP techniques enhance classification performance? and What IT skills are most dominant in the job vacancy data?

The remainder of this paper is structured as follows: Section 2 outlines the proposed research methodology based on CRISP-DM, covering business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Section 3 presents the results and discusses the performance of machine learning algorithms and their implications. Finally, Section 4 concludes the study and summarizes the main contributions of this research.

2. RESEARCH METHOD

Following the CRISP-DM methodology as proposed by Pete Chapman et al. [1], this research involved a six-phase process. These phases include business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

2.1. Business Understanding

The business understanding phase aims to determine the urgency of conducting research, assess the situation, define the data mining objectives, and create a project plan. This phase covers understanding the problem statement, achieving research goals and limitations, obtaining supporting data, potentially utilization of data and analysis to obtain the best model. In this phase, the research objectives are clearly defined based on business needs.

In the business understanding phase, this research aims to address three primary objectives. First, it is dedicated to developing robust classification models for IT skill texts, leveraging data mining methodologies to enhance classification accuracy while tackling data imbalance challenges. Machine learning algorithms such as Naive Bayes Classifier (NBC), k-Nearest Neighbor (KNN), and Support Vector Machine (SVM) are employed to achieve this goal [6],[16]. Second, the study seeks to identify and analyze specific subcategories of IT skills, moving beyond broad categorizations to uncover more granular insights that can support targeted decision-making processes. Lastly, the research integrates advanced Natural Language Processing (NLP) techniques [12], including Exploratory Data Analysis (EDA) [21] and n-grams, to refine the classification process, ensuring higher precision and reliability in the generated models. These objectives collectively align to derive actionable and detailed insights from IT skill-related data.

2.2. Data Understanding

During the data understanding phase, a series of actions were undertaken to gain a deeper understanding of the data used in this research. In the data understanding phase, a comprehensive dataset exploration is conducted to gain deeper insights into its structure and potential. This phase focuses on data exploration, sampling, description, and data collection to ensure high-quality data and provide essential initial insights for further analysis. The initial data collection process was conducted, followed by identifying of existing data quality, discovering knowledge from the data, and subsequent data analysis to formulate hypotheses from the hidden information within the data [22]. The findings from this stage will inform the subsequent data modeling process and guide the creation of effective models.

Job vacancy data was collected from the Jobstreet website using Python web scraping techniques [23]. This data encompassed information such as job titles, companies, required skills (descriptions), location, job sub-categories, job types, incentives (salaries), and data ingestion date. Exploratory Data Analysis (EDA) was employed to comprehend the characteristics of the data, including the distribution of job categories, text patterns, and relationships between variables. This process also identified potential anomalies or missing data that could impact the analysis. The data understanding phase is divided into several key steps: data collection, data comprehension, data quality verification, data visualization, and exploratory data analysis (EDA) [8].

a. Data Collection

Data sources include web scraping from the Jobstreet Indonesia API and other relevant sources. The dataset was collected in January, April, and June of 2024. Initial analysis revealed that the dataset consists of variables related to job vacancies in the IT sector.

b. Data Comprehension

Data description examines the dataset's structure, including the quantity of data, variable types, data formats, distribution comparisons, unique words, top rows, and frequencies. The analysis found that the dataset contains 2,340 rows. The data is generally in object type, with specific types such as strings, integers, and categorical data, offering a wide scope for further analysis.

Table 1. Data visualization (e.g. variable name, record value, quantity record, frequency)

	Job_title	Company	Descriptions	Location	Subcategory	Type	Salary	Date_ingestion
Count	2340	2304	2233	2340	2340	2340	550	2340
Unique	317	251	309	61	21	5	51	4
Top	Systems Analyst	OTO Group	What's your expected monthly basic salary? How...	South Jakarta	Business/System Analyst	full time	IDR 5,000,000 – IDR 6,000,000 per month	19/06/2024
Freq	305	127	79	725	805	1993	67	1440

Table 1 is the descriptive statistics for the job postings dataset. The table includes columns such as job title, company, descriptions, location, subcategory, type, salary, and date ingestion. For example, the most frequent job title is 'Systems Analyst' (305 occurrences), and the top company is 'OTO Group' (127 postings). The location with the highest frequency is South Jakarta (725 postings), while 'Business/System Analyst' dominates the subcategory (805 occurrences). Additionally, salary information is only available for 550 postings, with the most common range being IDR 5,000,000 – IDR 6,000,000 per month. Finally, data ingestion dates span 4 unique values, with 19/06/2024 being the most frequent (1440 entries).

The data consists of 8 variables, including job title, company, descriptions, location, subcategory, type, salary, and date_ingestion. A detailed description of each variable is presented in a table to understand its potential and inform subsequent analysis.

Table 2 defines the constituent variables in the dataset and provides their corresponding descriptions. The 'Job title' refers to the vacancy being offered, while 'Company' specifies the organization posting the job. The 'Descriptions' column details the roles, responsibilities, and qualifications required for prospective applicants. 'Location' indicates the job placement area, and 'Subcategory' classifies the specific IT skill expertise. The 'Type' variable identifies the job nature, such as full-time, contract, or casual positions. Additionally, 'Salary' highlights the incentives offered, and 'Date Ingestion' records when the data was acquired.

Table 2. Dataset Constituent Variables

Variable	Descriptions
Job title	Job vacancy needed
Company	Company name
Descriptions	The detailed responsibilities required by the company, typically accompanied by the qualifications that prospective applicants must meet.
Location	Job placement location
Subcategory	IT skill subcategory expertise
Type	Job type (Full-time, Contract/Temporary, or Casual/Vacation)
Salary	The incentives offered
Date Ingestion	Date of data acquisition

c. Data Quality Verification

Data quality verification focuses on identifying missing values, checking the distribution of text length, and detecting duplicates. Missing values are checked using the isnull() function, and the output is analyzed. The dataset contains several variables, and an analysis of missing values was conducted for each variable. The ‘job_title’ variable has no missing values, while ‘company’ has 36 missing entries. The ‘descriptions’ variable contains 107 missing values, and the ‘salary’ variable has 1,790 missing entries. The other variables—'location’, ‘subcategory’, ‘type’, and ‘date_ingestion’—have no missing data. The dataset’s data type is int64, indicating that most variables are numerical. This information is critical for ensuring the integrity and completeness of the dataset before further analysis and modeling.

Data Visualization

A statistical analysis of the value distribution and statistical modeling was performed to understand the distribution of values within the subcategory variable [24]. This process aims to evaluate data quality and identify potential issues. The results of the value distribution for the subcategory variable are provided in the appendix. According to the table, the most frequent distribution of values is held by the Business/System Analyst with 805 entries, followed by Network & System Administration with 538 entries, and Developer/Programmer with 198 entries.

Based on the analysis, the distribution of stopwords in the descriptions variable that appear most frequently are [('and', 17075), ('to', 9100), ('in', 6952), ('you', 6472), ('the', 6131), ('of', 5953),

Figure 1. Wordcloud visualization

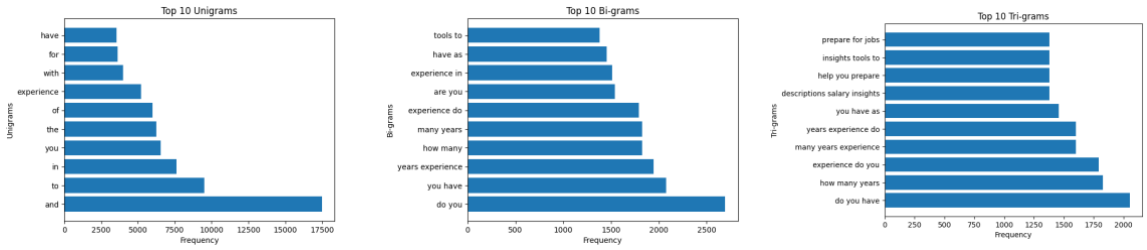
('a', 4520), ('experience', 3744), ('with', 3715), ('for', 3606)]. Based on these findings, the implementation of a word cloud is used to identify key terms and important phrases frequently found in the descriptions text. The word cloud visually represents the text patterns in the ‘descriptions’ variable, with the aim of recommending the removal of irrelevant words that do not support the research objectives, thus improving the accuracy of the developed model, as shown in Figure 1.

d. N-gram

EDA provides a comprehensive data analysis using n-gram methods (unigram, bigram, trigram). These techniques are fundamental in natural language processing, helping to uncover recurring language patterns and key phrases. Analysis of unigram, bigram, and trigram data reveals important linguistic structures within the dataset, which will be valuable for refining the classification models. The insights from EDA guide the refinement of models to enhance their accuracy and relevance.

Figure 2 (a) presents a bar chart uni-grams, which highlights the frequency of the ten most common single words (unigrams) in a given dataset. The x-axis represents the unigrams, arranged by frequency, while the y-axis indicates their count or occurrence. The height of each bar reflects the relative frequency of each unigram, with taller bars representing words that appear more frequently. The top 10 unigrams, listed in descending order, are ‘have’, ‘for’, ‘with’, ‘experience’, ‘of’, ‘the’, ‘you’, ‘in’, ‘to’, and ‘and’.

Similarly, Figure 2 (b) displays a bar chart bi-grams, which focuses on the most frequent pairs of consecutive words in the dataset. The top 10 bi-grams, listed from highest to lowest frequency, are ‘tools to’, ‘have as’, ‘experience in’, ‘are you’, ‘experience do’, ‘many years’, ‘how many’, ‘years experience’, ‘you have’, and ‘do you’. Compared to the results of the unigram analysis, the bi-gram analysis provides a deeper understanding of the discussed topics. The phrases that emerge in the bi-gram analysis reveal more specific word relationships and clearer contextual meaning.



(a) (b) (c)

Figure 2. Visualization of the distribution of word pairs n-grams, (a) unigram, (b) bigram & (c) trigram that appear sequentially in the research corpus, illustrating word-to-word relationships

On Figure 2 (c), compared to the results of the unigram analysis, the tri-gram analysis provides a deeper understanding of the topics discussed. The phrases that appear in the tri-gram analysis show more specific contexts and relationships between words. Additionally, phrases such as ‘prepare for jobs’ and ‘insights tools’ indicate supporting terms related to career preparation. The dominance of phrases related to work experience and job searching suggests that this dataset can be a valuable resource for further research on labor market trends and job seeker expectations.

2.3. Data Preparation

Data preparation aims to select relevant data, integrate chosen data, clean unused data, and format data for processing using data mining techniques. By leveraging machine learning techniques such as regex and spaCy libraries [21], data collected undergoes cleaning and transformation processes using Natural Language Processing (NLP) [25]. This stage involves, removing redundant data and empty entries, handling outliers, converting data types, text normalization such as removing punctuation and numbers, removing stop words, and feature selection. Data preparation aims to select relevant data, integrate chosen data, clean unused data, and format data (e.g. case folding, cleaning, stopword removal, tokenizing, and joining) for processing using data mining techniques.

By leveraging machine learning techniques such as regex and spaCy libraries, data collected undergoes cleaning and transformation processes using Natural Language Processing (NLP) [26]. This stage involves:

a. Removing redundant data and empty entries

The isnull() function is used to identify missing values, revealing that 107 rows contain missing data. Subsequently, the missing rows are removed using the dropna() function, ensuring that the missing data does not compromise the accuracy of the model implementation.

b. Handling outliers

Outlier identification for the unique word ratio is conducted through several systematic steps. First, the unique word ratio for each description is calculated by dividing the number of unique words by the total word count in the text. Next, the mean and standard deviation of the ratios are computed to establish the normal range boundaries.

c. Converting data types

This study converted specific data types from object data to string using the astype() function to facilitate subsequent analysis processes.

d. Text normalization such as removing punctuation and numbers

Table 3 presents a comparison of the data before and after cleaning. The purpose of data cleaning is to improve data quality and ensure more accurate analysis results. The changes made include capital letter normalization, removal of irrelevant punctuation (e.g., ellipses), and correction of typographical errors such as truncated or misspelled words. These data cleaning steps aim to enhance data quality and ensure that the analysis results are more accurate and reliable.

Table 3. Data Cleaning

Before	After
Bachelor's degree in computer science, enginee...	bachelors degree in computer science engineeri...
Develop new user interface features that meet ...	develop new user interface features that meet ...
Deeply engaged in the full development lifecyc...	deeply engaged in the full development lifecyc...
Minimum working experience 2 years; Passion fo...	minimum working experience years passion for p...
Male/Female, max age 28 years; Bachelor's degr...	malefemale max age years bachelors degree in i...
...	...

e. Stopwords removal

In this study, stopword removal was performed using default stopword lists available in text processing libraries, such as NLTK, and external text sources. This step aims to enhance the efficiency and accuracy of the text classification model by retaining only words with significant informational value for further learning processes.

Table 3 presents the impact of stopword removal on the dataset, highlighting changes in the total word count and the proportion of meaningful words retained. The results of the stopword removal can be seen in Table 4. The text after cleaning is more focused on relevant key terms, such as ‘development’, ‘design’, and ‘programming’.

Table 4. Stopword Removal

Before	After
bachelors degree in computer science engineeri...	science engineering app development develop mi...
develop new user interface features that meet ...	develop uiux design web design design backend ...
deeply engaged in the full development lifecyc...	engage development lifecycle design develop te...
minimum working experience years passion for p...	programming programming js php mysql nosql res...
malefemale max age years bachelors degree in i...	informatics engineer python programming progra...
...	...

f. Feature selection

The subcategory variable is the label since it is categorical data, while the descriptions variable serves as a supporting feature. Table 4 illustrates the feature selection process carried out to identify characteristics or attributes relevant to this study. The selected features will be used for further analysis to ensure that only the most significant and meaningful attributes contribute to the research findings.

On the Table 5, it can be observed that this study focuses on analyzing data related to professions in the field of software development. The selected features reflect the skills and knowledge required in this domain.

Table 5. Feature Selection

Subcategory	Descriptions
Developer/Programmer	science engineering app development develop mi...
Engineering - Software	develop uiux design web design design backend ...
Developer/Programmer	engage development lifecycle design develop te...
Engineering - Software	programming programming js php mysql nosql res...
Engineering - Software	informatics engineer python programming progra...
...	...

2.4. Modeling

The modeling phase involved implementing text classification models to measure the accuracy of the model in predicting data. Model implementation was carried out using Python notebooks in the VS Code.

The modeling phase utilizes only two variables: the subcategory variable as the label and the result_descriptions variable as the text feature. The variable X represents result_descriptions, while the variable y represents subcategory. Data classification involved dividing the dataset into two parts: a training set and a testing set. The dataset is randomly split into a training set and a testing set with a proportion of 80:20. The training set, comprising 80% of the data, consists of 1,872 rows, with X_train containing 1,497 rows and y_train containing 1,497 rows. Meanwhile, the testing set, comprising 20% of the data, consists of 468 rows, with X_test containing 375 rows and y_test containing 375 rows.

Data vectorization is performed using the TF-IDF Vectorizer from the scikit-learn library. The TF-IDF Vectorizer calculates the importance of a word in a document relative to other documents in the collection, enabling effective feature representation for the text data, as shown in Table 6. This table presents the TF-IDF scores for the most significant terms in the dataset, highlighting their relevance and contribution to the analysis.

In this study, we applied three different algorithms—K-Nearest Neighbors (KNN), Naive Bayes (NBC), and Support Vector Machine (SVM)—to test and compare their performance in addressing the problem at hand. The selection of these three algorithms aimed to compare the accuracy levels of each model. These algorithms were selected to develop the best-performing model, improving the overall modeling outcomes. The choice of algorithms is based on their ability to handle text data, particularly in text classification tasks. The models will be evaluated during the evaluation phase.

Table 6. Vectorization evaluate the importance of a word relative to a collection of documents

Vectorization X train		Vectorization X test	
(1, 398)	0.2474602900263872	(0, 350)	0.2498692520920671
(1, 6)	0.2174957106087361	(0, 319)	0.21973383332961915
(1, 271)	0.23118892830088858	(0, 200)	0.43340569383231703
(1, 200)	0.17223399763523217	(0, 117)	0.26041884298058404
…		...	
(1871, 270)	0.1827659159596059	(466, 22)	0.14008020044230768

2.5 Evaluation

A comprehensive model evaluation is conducted, reviewing the steps taken to build the model and ensuring that it achieves the research objectives. The evaluation stage is another step to thoroughly evaluate the data mining objectives in achieving the desired modeling. It is crucial for measuring the performance of the developed model.

The classification algorithm is evaluated using a confusion matrix based on evaluation measures such as accuracy, precision, recall, and F1-score [27]. The evaluation results indicate the effectiveness of each algorithm in handling text data. Furthermore, a deeper analysis is conducted to evaluate the distribution of predictions and identify significant classification errors, as shown in Figure 3. This figure visualizes the distribution of the model's predictions and highlights the areas where the algorithm made notable errors, providing insights for further model improvement.

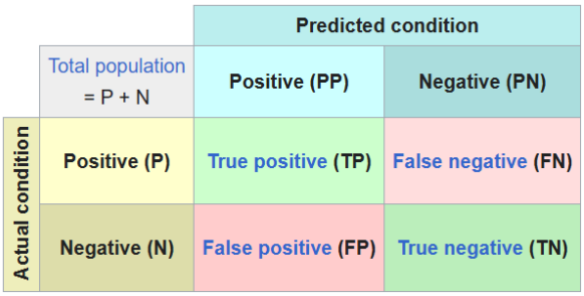


Figure 3. Evaluation matrix using confusion matrix

The confusion matrix in Figure 3 provides a detailed view of the model's prediction performance, allowing us to assess the distribution of true positives, false positives, true negatives, and false negatives. Building on this analysis, we can calculate several key evaluation metrics, including accuracy, precision, F1-score, and support, to quantify the model's performance more precisely [21]. The formulas represent the main evaluation metrics derived from the confusion matrix, namely accuracy, precision, recall, and F1-score. These metrics are essential in this research because each provides a different perspective on the performance of the classification model. For instance, in cases of data imbalance, accuracy alone can be misleading, as a model may achieve high accuracy simply by predicting the majority class. Therefore, precision and recall become crucial for evaluating how well the model specifically and comprehensively identifies the positive class. The F1-score then provides a balance between precision and recall, which is particularly useful when both are equally important. In the context of this study, these metrics offer a comprehensive view of the model's performance, especially in assessing its effectiveness in detecting patterns or anomalies that are the main focus of the research. The formulas for these metrics are as follows:

$$accuracy = \frac{TP + TN}{P + N} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} = \frac{TP}{P} \quad (3)$$

$$F1 - score = \frac{2\ precision\ x\ recall}{precision + recall} \quad (4)$$

2.6 Deployment

The developed and evaluated text classification model is applied in the deployment phase. This process encompasses several key stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The final research results are visualized using Google Looker Studio. Looker Studio enables the creation of interactive data visualizations and dashboards that display the results of the classification model in an informative and easy-to-understand manner [28]. Visualizations include the distribution of in-demand skills based on job categories, skill trends, and algorithm performance comparisons.

Furthermore, a label visualization with percentages to highlight the classes with the highest number of entries in the dataset. This visualization illustrates the proportion of skill occurrences within each label, emphasizing the dominance of specific skills and showing how this may affect model bias. The deployment phase ensures that the results of the research are easily accessible and comprehensuble, enabling in-depth analysis and aiding decision-making in real-world applications.

3. RESULTS AND ANALYSIS

The evaluation process involves measuring the performance of the models using evaluation metrics such as accuracy, precision, recall, F1 score, and the confusion matrix. This evaluation process is conducted to compare the accuracy levels of the three models: KNN, NBC, and SVM.

3.1. Performance of Machine Learning Algorithms.

This study implemented three machine learning algorithms—KNN, NBC, and SVM—to compare their classification performance. These algorithms were selected for their suitability in handling text-based data, particularly in text classification tasks. The developed models used text features extracted from the description variable as input and the subcategory variable as the output label. The dataset was split into 80% training and 20% testing data, and the text data was transformed using the TF-IDF Vectorizer to produce effective feature representations. Below is a summary of evaluation results based on accuracy, precision, recall, and F1-score, as presented in Table 7. This table provides a detailed comparison of the model's performance across different metrics, allowing for a comprehensive assessment of its effectiveness in the given task.

Table 7. Evaluation Matrix Of Machine Learning Model

	Accurcy	Precision	Recall	F1-score	Support
KNN	0.83	0.85	0.83	0.82	468
NBC	0.79	0.80	0.79	0.77	468
SVM	0.87	0.87	0.87	0.86	468

a. K-Nearest Neighbors (KNN)

KNN demonstrated good performance for balanced label distributions but struggled with minority classes. It excelled in categorizing frequently occurring labels but faced challenges with smaller data distributions, leading to reduced recall for less-represented labels.

b. Naive Bayes Classifier (NBC)

NBC was effective in handling short-text data and showed robustness against imbalanced datasets. Despite this, its overall accuracy was slightly lower compared to other algorithms, as it struggled to adapt to more complex text distributions.

c. Support Vector Machine (SVM):

SVM outperformed other algorithms by achieving the highest accuracy. Its strength lies in its ability to capture complex decision boundaries, making it particularly effective for text with significant variations. This capability allowed SVM to handle high-dimensional data and deliver superior classification performance.

The confusion matrix provided insights into the distribution of predictions and classification errors. SVM consistently exhibited higher precision, recall, and F1 scores, reflecting its robustness in managing text classification tasks. These results align with previous studies that highlighted SVM's effectiveness in separating data with clear margins and handling textual complexity.

The results indicate that SVM's superior performance is due to its capability to optimize decision boundaries in high-dimensional space, making it ideal for imbalanced and text-heavy datasets. Conversely, while NBC handled imbalanced data reasonably well, its simplicity limited its ability to model intricate text relationships. KNN's dependency on label frequency hindered its adaptability to skewed class distributions.

3.2. Natural Language Processing (NLP) Techniques

NLP techniques were employed to process textual data extracted from job descriptions, enhancing data quality and supporting efficient classification [25]. The process began with text normalization, where punctuation, numbers, and irrelevant words (stopwords) were effectively removed using libraries like NLTK. Next, text exploration involved using n-grams (unigram, bigram, trigram) to identify frequent linguistic patterns, revealing dominant keywords such as management, server, and security. EDA complemented this by uncovering data distribution patterns, anomalies, and relevant textual characteristics, while n-gram analysis highlighted significant linguistic structures within the job descriptions. The descriptions variable was utilized as text features, while the subcategory variable served as the classification label. These NLP techniques streamlined the data, reducing redundancy and preserving critical information, which significantly improved the classification models' efficiency and performance.

3.3. Visualization and Analysis with Looker Studio

Classification results were visualized using Google Looker Studio to provide deeper insights into the data. Key visualizations included distributing IT skill categories, skill trends based on word frequency, and algorithm performance comparisons.

The dataset consists of 2,340 records, and sampling techniques were applied to balance the distribution of labels across categories and reduce bias in the classification model being developed. During the data understanding phase, it was found that there are 21 distinct labels, representing subcategories within the field of Information Technology. In the distribution analysis, with notable imbalances: the Business/System Analyst category dominated with 805 entries, followed by Network & System Administration (538 entries) and Developer/Programmer (198 entries).

Based on the calculations, the keyword ‘management’ appears most frequently, with a total frequency of 1,686 occurrences, accounting for 3.4% of the overall keyword distribution in the dataset. It is followed by keywords such as ‘server’, ‘security’, and ‘application’, each with a total frequency of 1,598 occurrences and a 3% appearance rate. This analysis provides a detailed overview of the thematic focus within the dataset, offering targeted insights that can be utilized in subsequent modeling stages. This visualization focuses on the skills identified in job descriptions, providing an overview of the main skills frequently mentioned within each category and their occurrence percentages. Grouping and visually presenting these keywords helps users understand skill trends and demands in the modeled data.

The performance of classification algorithms was also evaluated, with SVM excelled across all metrics, demonstrating its ability to manage text complexity effectively. Additionally, Looker

Studio visualizations emphasized labels with the highest occurrence frequencies, shedding light on how label distribution can influence model bias.

3.4. Analysis

The results from this study align with findings in previous research while addressing key gaps. For instance, Solekhah et al. reported that KNN achieved higher accuracy than NBC, which is consistent with the results here [6]. However, this study's inclusion of SVM revealed its superior performance compared to both KNN and NBC, a finding corroborated by Hermawan (2023) [9], who noted SVM's effectiveness in sentiment analysis tasks. Additionally, the use of TF-IDF and advanced NLP techniques in this study enhanced text representation, which was less explored in prior works. This research also highlights the impact of addressing data imbalance using robust algorithms like SVM. Compared to earlier studies that struggled with imbalanced datasets, this study demonstrated how preprocessing and feature engineering could mitigate such challenges [30]. The findings underscore the importance of combining machine learning algorithms with advanced NLP methods to achieve reliable classification performance, especially in complex domains like IT skills categorization.

By integrating these insights, this study contributes to developing more accurate models, providing a foundation for future research in text classification and IT skill analysis.

4. CONCLUSION

This study identified key in-demand skills in the information technology sector by analyzing 2,340 job vacancy texts using machine learning and NLP techniques, thereby addressing the research question: “What are the most required IT skills in the current job market, and how can text classification algorithms support this analysis?” The results revealed that the Business/System Analyst role is the most sought-after (34.1%), with critical skills such as ‘management’, ‘server’, ‘security’, ‘application’, and ‘design’ emerging as highly relevant.

To classify job categories from unstructured text data, three machine learning algorithms—SVM, KNN, and NBC—were applied. Among them, SVM showed the highest classification performance, achieving an 87% improvement in evaluation metrics compared to NBC and KNN. This superior result is due to SVM’s robustness in handling high-dimensional, sparse data typical in textual formats, unlike NBC and KNN which struggle under such conditions. The application of NLP techniques significantly enhanced data quality and feature representation, further contributing to model performance.

Additionally, data visualization using Looker Studio enabled clear presentation of findings, uncovering patterns in job demand and skills distribution. This supports data-driven decision-making for both companies (in hiring) and job seekers (in skill development).

In conclusion, this study provides concrete evidence that combining text mining, machine learning, and visualization tools not only improves classification accuracy but also enables insightful analysis of labor market trends. These findings directly answer the research question and offer practical implications for the development of industry-aligned education, training strategies, and job matching systems.

ACKNOWLEDGEMENTS

I am grateful to my supervisors for their unwavering support and invaluable guidance throughout this project. Their mentorship significantly contributed to the successful completion of this research. Additionally, I acknowledge the resources and platforms provided by Kaggle, JobStreet, and Python, which were instrumental in facilitating this study. The combined support of all involved parties has been pivotal in achieving the objectives of this research.

REFERENCES

[1] C. Pete *et al.*, “CRISP-DM 1.0,” in *CRISP-DM Consortium*, 2000, p. 76.
[2] R. Wirth, “CRISP-DM : Towards a Standard Process Model for Data Mining,” no. 24959.

[3] E. Kristoffersen, O. O. Aremu, F. Blomsma, P. Mikalef, and J. Li, *Exploring the Relationship Between Data Science and Circular Economy: An Enhanced CRISP-DM Process Model*, vol. 11701 LNCS. Springer International Publishing, 2019. doi: 10.1007/978-3-030-29374-1_15.
[4] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.
[5] N. Caetano, P. C. B, and R. M. S. Laureano, "Using Data Mining for Prediction of Hospital Length of Stay : An Application of the CRISP-DM Methodology," vol. 2, pp. 149–166, doi: 10.1007/978-3-319-22348-3.
[6] F. Sholekhah, A. D. Putri, R. Rahmaddeni, and L. Efrizoni, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbors untuk Klasifikasi Metabolik Sindrom," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 507–514, 2024, doi: 10.57152/malcom.v4i2.1249.
[7] N. S. Wardani, A. Prahutama, and P. Kartikasari, "Analisis Sentimen Pemindahan Ibu Kota Negara Dengan Klasifikasi Naïve Bayes Untuk Model Bernoulli Dan Multinomial," *J. Gaussian*, vol. 9, no. 3, pp. 237–246, 2020, doi: 10.14710/j.gauss.v9i3.27963.
[8] A. C. Khotimah *et al.*, "Comparison Naive Bayes Classifier, K-Nearest Neighbor And Support Vector Machine In The Classification of Individual On Twitter Account," vol. 3, no. 3, 2022.
[9] A. Hermawan, I. Jowensen, J. Junaedi, and Edy, "Implementasi Text-Mining untuk Analisis Sentimen pada Twitter dengan Algoritma Support Vector Machine," *JST (Jurnal Sains dan Teknol.*, vol. 12, no. 1, pp. 129–137, 2023, doi: 10.23887/jstundiksha.v12i1.52358.
[10] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model," *Procedia CIRP*, vol. 79, no. March, pp. 403–408, 2019, doi: 10.1016/j.procir.2019.02.106.
[11] Y. Nurdiansyah, A. Andrianto, and L. Kamshal, "New book classification based on Dewey Decimal Classification (DDC) law using tf-idf and cosine similarity method," *J. Phys. Conf. Ser.*, vol. 1211, no. 1, 2019, doi: 10.1088/1742-6596/1211/1/012044.
[12] A. D. Adhi Putra, "Sentiment Analysis on User Reviews of the Bibit and Bareksa Application with the KNN Algorithm," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 636–646, 2021.
[13] Y. A. Singgalen, "Analisis Sentimen Wisatawan terhadap Kualitas Layanan Hotel dan Resort di Lombok Menggunakan SERVQUAL dan CRISP-DM," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 1870–1882, 2023, doi: 10.47065/bits.v4i4.3199.
[14] A. Géron, *Hands-On Machine Learning with Scikit-Learn*, 1st editio. Sebastopol: O'Reilly Media, Inc., 2019. doi: dl.acm.org/doi/10.5555/3378999.
[15] H. Sulistiani, *Implementasi Berbagai Metode Kecerdasan Buatan (Artificial Intelligence) Pada Masalah Gangguan Kepribadian (Narcissistic Personality Disorder: NPD)*. Bandarlampung, 2024.
[16] V. Nurcahyawati and Z. Mustaffa, "Improving sentiment reviews classification performance using support vector machine-fuzzy matching algorithm," *Bull. Electr. Eng. Informatics*, vol. 12, no. 3, pp. 1817–1824, 2023, doi: 10.11591/eei.v12i3.4830.
[17] J. T. Sri Sumantyo, "Development of circularly polarized Synthetic Aperture Radar onboard Unmanned Aerial Vehicle (CP-SAR UAV)," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2012, pp. 4762–4765. doi: 10.1109/IGARSS.2012.6352549.
[18] E. Fujisaki and T. Okamoto, "Secure integration of asymmetric and symmetric encryption schemes," in *Annual International Cryptology Conference*, Springer, 1999, pp. 537–554.
[19] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE Trans. Inf. theory*, vol. 31, no. 4, pp. 469–472, 1985.
[20] Y. Arta, E. A. Kadir, and D. Suryani, "KNOPPIX: Parallel computer design and results comparison speed analysis used AMDAHL theory," in *Information and Communication Technology (ICoICT), 2016 4th International Conference on*, IEEE, 2016, pp. 1–5.
[21] M. Hofmann and R. Klinkenberg, *Data Mining and Knowledge Discovery Series*. 2014.
[22] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 6382–6388, 2019, doi: 10.18653/v1/d19-1670.
[23] A. Downey, J. Elkner, and C. Meyers, "Think Python: How to Think Like a Computer Scientist," p. 304, 2014.
[24] R. Ribeiro, A. Pilastri, C. Moura, F. Rodrigues, R. Rocha, and P. Cortez, "Predicting the tear strength of woven fabrics via automated machine learning: An application of the CRISP-DM methodology," *ICEIS 2020 - Proc. 22nd Int. Conf. Enterp. Inf. Syst.*, vol. 1, pp. 548–555, 2020, doi: 10.5220/0009411205480555.
[25] S. Y. Feng *et al.*, "A Survey of Data Augmentation Approaches for NLP," *Find. Assoc. Comput.*

Linguist. ACL-IJCNLP 2021, pp. 968–988, 2021, doi: 10.18653/v1/2021.findings-acl.84.
[26] Y. Kang, Z. Cai, C. W. Tan, Q. Huang, and H. Liu, "Natural language processing (NLP) in management research: A literature review," *J. Manag. Anal.*, vol. 7, no. 2, pp. 139–172, 2020, doi: 10.1080/23270012.2020.1756939.
[27] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.
[28] I. Analytics, D. S. M. Media, and A. R. Reserved, "A complete guide to cleaning and preparing data for analysis using Excel™ and Google Sheets™," 2019.
[29] A. Zhang, *Data Analytics: Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life*. North Charleston: CreateSpace Independent Publishing Platform, 2017. doi: https://dl.acm.org/doi/book/10.5555/3153180.
[30] H. Wiemer and L. Drowatzky, "A Holistic Extension to the applied sciences Data Mining Methodology for Engineering Applications (DMME)," 2019.