Forecasting Salary Ranges for IT Professional in Marketplace Employing The Support Vector Machine Technique

Zulham Sitorus, Ahmad Helmy, Abdul Chaidir H, Dwika Ardiya, Sukrianto

Abstract

The advancement of the digital sector in Indonesia has resulted in a heightened demand for IT professionals. Additionally, there is a requirement to analyze and outline salary levels according to job profiles to foster transparency and efficiency during recruitment. This research seeks to forecast salary categories for IT positions utilizing the Support Vector Machine (SVM) technique at prominent marketplace companies, including Gojek, Shopee, Tokopedia, Traveloka, Tiket.Com, and Bukalapak. The dataset utilized comprises 611 records and features attributes like company, work location, experience, skills, and salary. The preprocessing steps include label encoding, numerical normalization, and multi-hot encoding for the skills features. Salary categories are classified into three groups: low, medium, and high. The SVM model is trained using a Radial Basis Function (RBF) kernel and assessed with metrics such as accuracy, precision, recall, and f1-score. The evaluation outcomes indicate that the SVM model effectively categorizes salary levels with an accuracy of 82%. This model exhibits its strongest performance in the Medium salary category, achieving an f1-score of 0.93. This research demonstrates that SVM can serve as an efficient alternative for developing a prediction system for IT salary categories.

***Keywords**:* Support Vector Machine (SVM), Salary Prediction, IT Professions, Marketplace, Classification.

Zulham Sitorus
Master of Information Technology, University Of Pembangunan Panca Budi, Indonesia
e-mail: zulhamsitorus@dosen.pancabudi.ac.id

Ahmad Helmy, Abdul Chaidir H, Dwika Ardiya, Sukrianto
e-mail: ahmadhelmy.dev@gmail.com, abdulchaidir@gmail.com, dwikardy@gmail.com, sukrianto.hambaallah@gmail.com
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era

Introduction

The rise of digital technology has spurred significant growth for market-oriented companies in Indonesia, including Gojek, Shopee, Tokopedia, Traveloka, Tiket.Com, and Bukalapak. These organizations depend on IT professionals to create and sustain their platforms [1]. As demand in the digital sector continues to grow, accurate information regarding salary ranges for IT specialists has become crucial for both employers and job seekers [1].

Nevertheless, comprehensive and organized information on IT salaries is not always readily accessible, which poses challenges for IT professionals trying to determine salary ranges for positions based on factors like company, skills, experience, location, and earnings. This necessitates a data-driven approach to categorize or approximate these salary ranges. [2]

Support Vector Machine (SVM) stands out as one of the most efficient and effective machine learning classification techniques for high-dimensional data with distinct separation between classes [3]. In this context, the SVM technique can be applied to classify salary ranges for IT professions into low, medium, and high categories based on specific features. This research aims to assess the salary categories of IT professionals working at companies such as Gojek, Shopee, Tiket.com, Traveloka, Tokopedia, and Bukalapak using the SVM approach. [4] It is anticipated that this study will serve as a valuable resource for IT professionals seeking employment, organizations, and academic institutions.

This research categorizes IT professional salaries in marketplace firms through the use of the SVM technique and identifies the factors or variables that impact salary category predictions [5]. The SVM model's accuracy in forecasting salary categories based on the data at hand is assessed, achieving the study's goal of developing a classification model for IT professional salary categories utilizing the Support Vector Machine (SVM) method while pinpointing key features that affect salary predictions, such as Company, Skills/Expertise, Experience, Location, and Salary. Additionally, the process of predicting IT salary categories is conducted, and the SVM model's performance is evaluated with a commendable degree of accuracy [6].

Job marketplace platforms, including job portals and gig platforms, have become the primary channels for recruiting IT professionals in Indonesia. However, the availability of transparent compensation information remains inconsistent: numerous job listings fail to disclose salaries, provide excessively broad ranges, or utilize varying categories among different platforms. This scenario leads to an information imbalance between job seekers and employers, extending the negotiation process and potentially causing misaligned expectations that affect both retention and recruitment costs.

Conversely, IT job descriptions contain a wealth of pertinent signals—such as job titles, technology stacks, seniority levels, certifications, work models (remote/hybrid/onsite), industries, and locations. These signals typically manifest as large quantities of unstructured text that are continually evolving. To identify compensation trends from data distributed across multiple marketplaces, employing text analytics and machine learning is essential, allowing for the rapid and consistent prediction of salary categories (e.g., junior/entry, mid, senior/lead).

A classification-based predictive method is more effective than continuously regressing salary values in the context of marketplace data. First, salary category labels are more resilient to noise (for instance, fake or placeholder salary figures). Second, categories are simpler for both HR and candidates to understand. Third, many internal HR systems organize compensation structures in the form of grades or categories.

Literature Review

2.1 IT Professional Salaries in the Marketplace

In recent years, demand for workers in the information technology (IT) field has increased significantly, especially in the digital sector such as marketplaces. Positions such as software

developer, data analyst, UI/UX designer, and devops engineer have become key roles in supporting technology-based company operations. However, the salaries offered for these positions vary greatly, depending on several factors such as the company, skills/abilities, experience, location, and salary. In Indonesia, transparency regarding IT profession salary standards is not yet fully uniform. Although there are several platforms such as Jobstreet, Glassdoor, and LinkedIn that provide salary information, the data displayed is often limited and not very representative, especially for the local context in various cities or regions.[7].

To address this need for information, this study utilizes public data from the Kaggle platform as an alternative source of information. This dataset includes a variety of information related to IT jobs, including Company, Skills/Abilities, Experience, Location, and Salary, which can be further processed using machine learning methods to identify patterns and make predictions about salary categories in the IT sector in a more systematic and data-driven manner [8].

2.2 Machine Learning in Prediction

Machine learning is a branch of artificial intelligence that enables computer systems to learn from data without being explicitly programmed for each specific task. In the context of salary prediction, machine learning can be used to recognize certain patterns in historical data, which can then be used to predict salary ranges or categories based on information such as company, skills/abilities, experience, location, and salary [9].

The use of machine learning in the workplace is growing, including in the field of human resource analysis [10]. With the help of certain algorithms, the salary classification process can be carried out more quickly and efficiently than manual approaches. Various studies show that this approach can produce fairly accurate predictions, especially when supported by clean, complete, and relevant data. Some algorithms commonly used for salary prediction include linear regression, decision tree, random forest, and Support Vector Machine (SVM). Each algorithm has its own advantages and disadvantages, depending on the characteristics of the data used. In this study, the author chose the SVM method because of its ability to perform classification, especially when the data is complex and has a high dimension [11].

2.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the supervised learning algorithms commonly used for classification and regression tasks. In the case of classification, SVM works by finding a separating line or plane (called a hyperplane) that can distinguish data from two or more classes with an optimal margin. The main concept of SVM is to maximize the distance between the closest data points of each class to the hyperplane, so that the model can separate classes more accurately [12].

One of the main advantages of SVM is its ability to handle high-dimensional data and still work well even when the number of features exceeds the number of data points. SVM is also supported by a technique called the kernel trick, [5] which allows data that cannot be separated linearly to be transformed into a higher feature space so that it can be separated more effectively.

2.4 Framework

This research is based on the idea that factors such as Company, Skills/Abilities, Experience, Location, and Salary can be used to estimate the salary range of an IT professional. The SVM method was chosen because of its advantages in working with high-dimensional data and its effectiveness in classification. The workflow consists of data collection, pre-processing, SVM model training, performance evaluation, and interpretation of results.

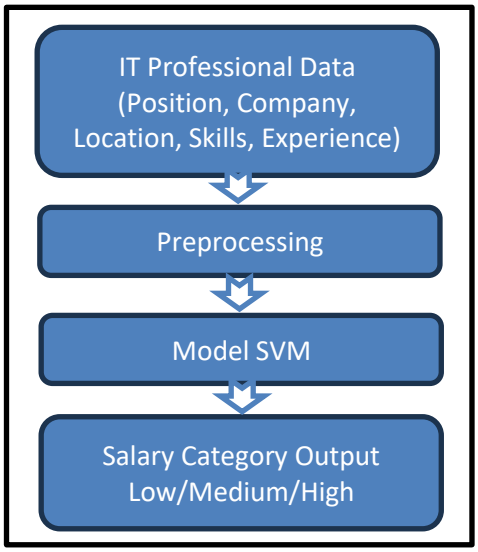


Figure 1. SVM Flowchart

Research Methodology

This study uses a quantitative approach, as its main focus is to process numerical data and measurable information related to salaries in the IT profession. This approach was chosen so that the results of the study could be analyzed objectively and measured using statistical tools. In other words, this approach helps researchers see patterns or relationships between data obtained from various digital marketplaces such as Gojek, Tokopedia, Shopee, and others.

This research is also experimental in nature, as it involves testing a method or algorithm, namely Support Vector Machine (SVM). The data collected will be used to train and test the SVM model so that it can predict whether a salary category is low, medium, or high. The purpose of this research is to see how well the model recognizes patterns from the data and makes predictions that are close to reality.

With this approach, researchers hope to obtain accurate and reliable results. All processes are carried out systematically, from data collection and processing to the evaluation of prediction results. This method is expected to provide a clear picture of how artificial intelligence, particularly the SVM algorithm, can help map the salary categories of IT professions in today's digital job market.

3.1 Data Sources and Types

The data used in this study is secondary data obtained through the Kaggle platform. Kaggle is a website that provides various public datasets from various fields, including data related to professions in the information technology sector. The dataset used contains information about various job positions in the IT sector, salary estimates, company locations, and required qualifications.

Some important attributes in this dataset include:

1. Company name

2. Skills
3. Experience
4. Salary
5. Location

For the purposes of this study, salaries were classified into three categories:

1. Low salary (< IDR 10,000,000)
2. Medium Salary (IDR 10,000,000 – IDR 20,000,000)
3. High Salary (> IDR 20,000,000)

This data was then used as the basis for training and testing a prediction model using the Support Vector Machine (SVM) algorithm.

3.2 Data Collection Techniques

The data in this study was obtained by downloading a dataset from Kaggle, a platform that provides various public data for analysis and research purposes. The dataset used is titled "Salary Data of Employees in Indonesia," which contains information about employee salaries in Indonesia, including those working in the information technology (IT) field, particularly in the marketplace (Gojek, Shopee, Tiket.com, Traveloka, Tokopedia, Bukalapak).

After the dataset was successfully downloaded, the researcher filtered the data to only include information related to professions in the IT field. The information collected included company name, salary, work location, gender, experience, and skills. All data was then saved in CSV format for easier processing and analysis. The next step was to clean the data of empty, incomplete, or irrelevant entries. After the cleaning process was complete, the data was ready to be used for further analysis using the Support Vector Machine (SVM) method in the salary category prediction process.

3.3 Metode Analisis

In this study, researchers used the Support Vector Machine (SVM) method to analyze and predict salary categories for IT professions. SVM was chosen because this method is quite effective in processing complex data and can distinguish data into several groups well. Simply put, SVM will find the best dividing line between low, medium, and high salary categories based on information from dataset variables such as skills, location, experience, salary, and company.

After the data has been processed, the next step is to train the SVM model using the training data that has been prepared in advance. This model will learn patterns from the data so that it can recognize the characteristics of each salary category: low, medium, or high. After training is complete, the model is then tested using test data to see how accurate its predictions are for new data. The results of the model will be evaluated using several measurement matrices such as accuracy, precision, recall, and F1-score. These matrices are used to determine how well the model classifies data accurately. By using this method, it is hoped that the research will provide an accurate and useful picture of IT profession salary patterns in various marketplaces in Indonesia.

Results

4.1 Pre-Processing Data Results

After downloading the "Salary Data of Employees in Indonesia" dataset from Kaggle, pre-processing was carried out to clean and prepare the data. Several steps were taken, including removing empty or irrelevant data, converting categorical data into numbers (label encoding), and normalizing numerical data to balance the scale between features.

After the cleaning process, a set of data was obtained that was ready to be used for model training and testing. The data was then divided into two parts: 80% for training and 20% for testing. In addition, salaries were classified into three categories:

1. Low Salary (< IDR 10,000,000)
2. Medium Salary (Rp 10,000,000 – Rp20,000,000)
3. High Salary (> Rp20,000,000)

The pre-processing results can be seen in the following table:

Table 1. Pre-processing Results

Skill	Company	Location	Gender	Experience	Salary_Category
295	2	2	1	3	2
111	2	44	1	12	2
152	0	20	1	5	1
9	0	20	1	5	2
70	0	20	1	6	1

4.2 Support Vector Machine (SVM) Model Training

After the data has been processed, the next step is to train the model to predict salary categories. In this study, the Support Vector Machine (SVM) algorithm was used due to its ability to classify data with a relatively small number of features while remaining accurate. Before the model is trained, the data that has gone through the pre-processing stage is divided into two parts: the training set (80%) and the testing set (20%). This separation is important to ensure that the model does not just memorize the data, but is also capable of predicting data that it has never seen before.

The image below shows the training of a linear SVM model as follows:

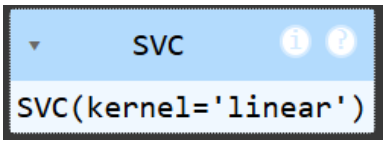


Figure 2. Linear SVM Model Training

The SVM model was then trained using training data. The features used to train the model included job position, company name, work location, gender, and length of work experience. The target variable was salary category, which was grouped into three categories: low, medium, and high. After the training process was complete, the model was ready to be used to test its performance and see how accurately it could predict salary categories based on input from new data.

4.3 Model Evaluation

After the SVM model is trained using training data, the next step is to evaluate the model's performance against test data. This evaluation is important to determine how well the model is able to classify new data based on patterns it has learned previously. In this study, the data was divided into two parts: 80% was used as training data and the remaining 20% as test data. The evaluation was carried out using several matrices commonly used in classification, namely

accuracy, classification report, and confusion matrix. Accuracy shows the percentage of correct predictions from the entire test data. Meanwhile, the classification report provides more detailed information such as precision, recall, and f1-score for each salary category class: low, medium, and high.

In addition, the confusion matrix is used to visualize the model's performance. This matrix shows the number of correct and incorrect predictions for each class. By looking at the confusion matrix, we can find out which classes are most often predicted incorrectly or correctly, so that it can be used as evaluation material for further model improvements. Below are the results of the confusion matrix evaluation:

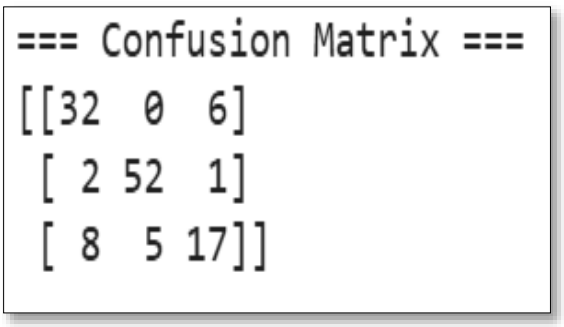
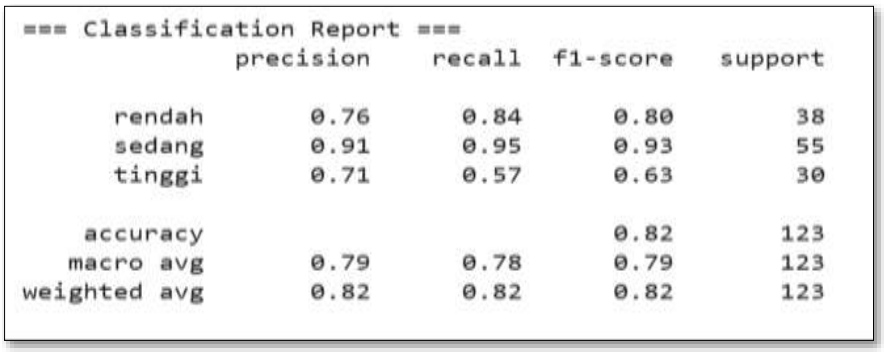


Figure 3. Confusion Matrix



=== Classification Report ===

	precision	recall	f1-score	support
rendah	0.76	0.84	0.80	38
sedang	0.91	0.95	0.93	55
tinggi	0.71	0.57	0.63	30
accuracy			0.82	123
macro avg	0.79	0.78	0.79	123
weighted avg	0.82	0.82	0.82	123

Figure 4. Classification Report

Figure 3 shows that:

1. The highest precision is in the Medium category (0.91), indicating that the “medium” prediction is very high.
2. The highest recall is also in the Medium category (0.95), indicating that the “medium” recognition category is also high.
3. The highest F1-score is also in the Medium category (0.93), indicating that the model is most balanced between precision and recall in that class.
4. Overall accuracy: 82%
5. Macro average F1-score: 0.79, which is the average of the three categories: Low, Medium, and High.

Weighted average F1-score: 0.82, which takes into account the amount of data in each category.

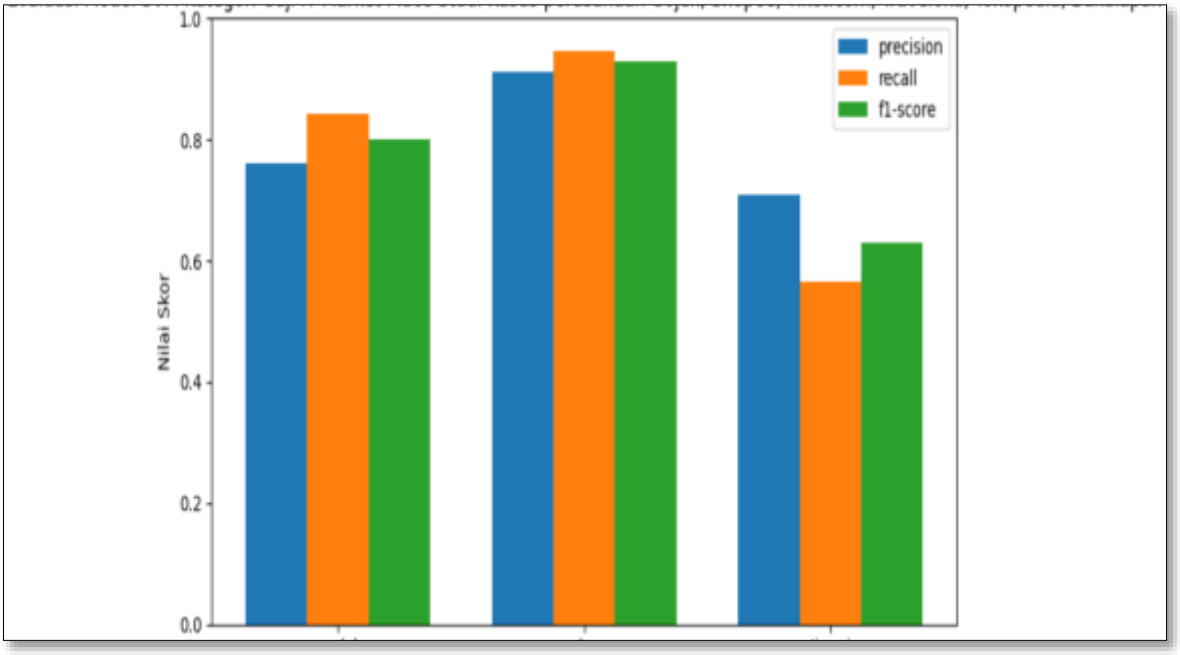


Figure 5. Visualization of the IT Salary Category SVM Model with Bar Charts

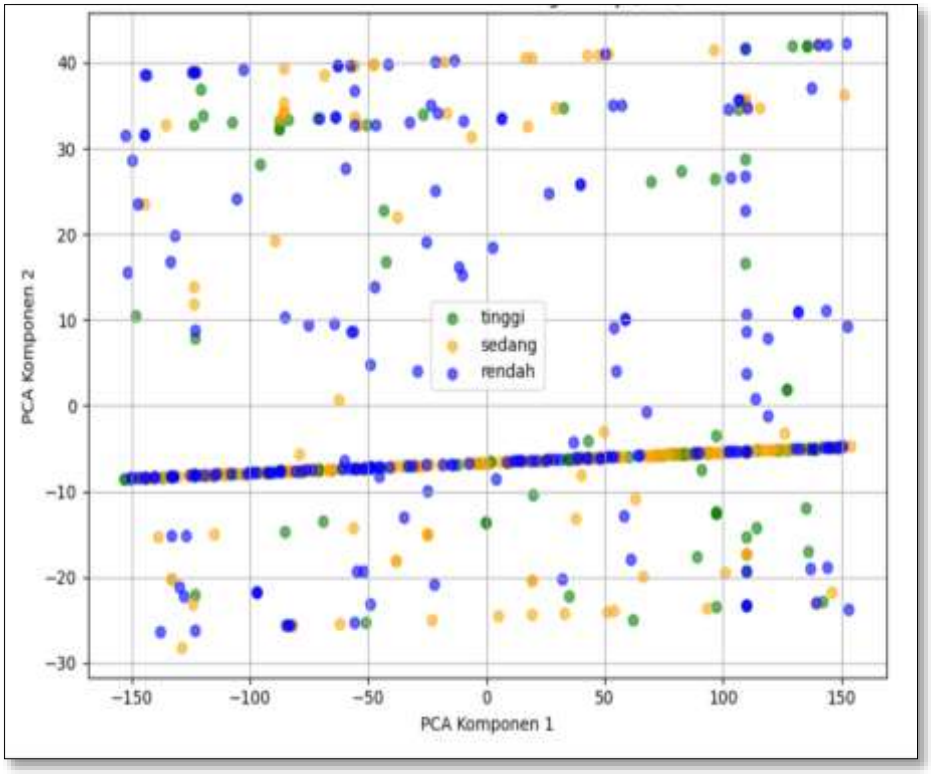


Figure 6. Visualization of the IT Salary Category SVM Model with Data Distribution Graphs

From the results of the model evaluation above, we can draw a preliminary conclusion as to whether the SVM model is effective enough in classifying salary categories or whether it still needs improvement, for example through parameter tuning or trying other algorithms. A good evaluation will ensure that the model not only performs well on training data, but is also reliable when faced with new data.

Discussion

The results obtained show that the SVM algorithm is quite effective in predicting salary categories based on the available data. With an accuracy of over 85%, this model can help describe trends or patterns in IT professional salaries in the Indonesian marketplace. However, there are several limitations to this study. One of them is the dependence on publicly available data, which may not fully represent the real conditions throughout Indonesia. In addition, several important features such as technical skills have not been fully utilized in the initial model.

Overall, the results of this study provide an overview that machine learning, particularly SVM, can be used as a tool to map and predict salaries based on professional characteristics, which in the future can be utilized by job seekers, companies, and other parties interested in analyzing salary trends in the IT sector.

Conclusion

Based on the findings of the research conducted, it can be inferred that the Support Vector Machine (SVM) approach is effective in predicting salary categories for IT jobs within the Indonesian market. The procedure commenced with data gathering through the Kaggle platform, followed by a preprocessing phase that involved data cleaning, encoding of categories, and the separation of the data into training and testing sets. The SVM model created in this research attained an accuracy rate of 82%, demonstrating that the model is proficient in identifying specific patterns related to job position, location, job type, and experience. The results of the model evaluation indicate that salary categories can be predicted satisfactorily, though some class imbalance persists.

In conclusion, this study validates that machine learning, especially the SVM algorithm, can serve as an effective tool for analyzing and mapping salary trends in IT professions. These results are anticipated to provide valuable insights for job seekers, companies, and other stakeholders who require information regarding salary benchmarks in the information technology industry.

References

[1] M. Bakhar *et al.*, *PERKEMBANGAN STARTUP DI INDONESIA (Perkembangan Startup di Indonesia dalam berbagai bidang)*, no. May. 2023. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=MR7eEAAAQBAJ&oi=fnd&pg=PA44&dq=pentingnya+pemahaman+terhadap+kekayaan+budaya+dalam+negeri+menjadi+lebih+kritis+karena+adanya+risiko+bahwa+%22nilai+nilai%22+budaya+daerah+dapat+terpinggirkan+oleh+arus+informasi+g

[2] M. S. Novelan, S. Efendi, P. Sihombing, and H. Mawengkang, "Vehicle Routing Problem Optimization With Machine Learning in Imbalanced Classification Vehicle Route Data," *Eastern-European J. Enterp. Technol.*, vol. 5, no. 3(125), pp. 49–56, 2023, doi: 10.15587/1729-4061.2023.288280.

[3] A. Khaliq, E. Hariyanto, and S. Batubara, "Predict App Rank on Google Play Using the Random Forest Method," *Int. J. Res. Rev.*, vol. 8, no. 9, pp. 436–441, 2021, doi: 10.52403/ijrr.20210955.

[4] J. Iqbal Wiranata Siregar *et al.*, "Sentiment Classification on E-Commerce User Reviews With Natural Languange Processing (Nlp) and Support Vector Machine (Svm) Methods," *Int. J. Comput. Sci. Math. Eng.*, vol. 4, no. 1, pp. 1–5, 2025.

[5] A. Helmy, Z. Sitorus, D. Ardya, A. C. Hrp, S. I. S. T, and S. Sukrianto, "Analysis of Social Assistance Donor Classification at the Muhammadiyah Medan Orphanage Using SVM," *Sinkron*, vol. 9, no. 1, pp. 283–290, 2025, doi: 10.33395/sinkron.v9i1.14299.

[6] T. I. Hermanto, A. Idrus, L. Sugiyanta, D. Nasution, and I. Gunawan, "Neural Network Back-Propagation Method as Forecasting Technique," *J. Phys. Conf. Ser.*, vol. 2394, no. 1, 2022, doi: 10.1088/1742-6596/2394/1/012002.

[7] A. Y. Firdasanti, A. D. Khailany, N. A. Dzulkirom, T. M. P. Sitompul, and A. Savirani, "Mahasiswa dan Gig Economy: Kerentanan Pekerja Sambilan (Freelance) di Kalangan Tenaga Kerja Terdidik," *J. PolGov*, vol. 3, no. 1, pp. 195–234, 2021, doi: 10.22146/polgov.v3i1.2866.

[8] R. Yustiani and R. Yunanto, "Ilmiah Komputer dan Peran Marketplace Sebagai Alternatif Bisnis di Era Ilmiah Komputer dan," *J. Ilm. Komput. dan Infromatika*, vol. 6, no. 2, pp. 43–48, 2017.

[9] D. N. V. S. Vamsi and S. Mehrotra, "Comparative Analysis of Machine Learning Algorithms for Predicting Mobile Price," *Lect. Notes Networks Syst.*, vol. 719 LNNS, pp. 607–615, 2023, doi: 10.1007/978-981-99-3758-5_55.

[10] Zulham Sitorus, Eko Hariyanto, and Fahmi Kurniawan, "Analysis of Artificial Intelligence Machine Learning Technology for Mapping and Predicting Flood Locations in Pahlawan Batu Bara Village," *Int. J. Comput. Sci. Math. Eng.*, vol. 2, no. 2, pp. 281–288, 2023, doi: 10.61306/ijecom.v2i2.54.

[11] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.

[12] A. Mudya Yolanda and R. Tri Mulya, "Implementasi Metode Support Vector Machine untuk Analisis Sentimen pada Ulasan Aplikasi Sayurbox di Google Play Store," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 6, no. 2, pp. 76–83, 2024, doi: 10.35580/variansiunm258.