JoCoSiR
Journal of Computer Science Research
ISSN: 2986-2337(Online)

Analysis and Classification of IT Professions in the Marketplace Using the Support Vector Machine Method

Dwika Ardya , Muhammad Iqbal

[a], Magister Teknologi Informasi, University Panca Budi, Medan, Indonesia
email: [a] dwikardy@gmail.com , [b] wakbalpb@yahoo.co.id

ARTICLEINFO

Keywords: [Keyword heading]
Maximum
Five
Word
Key
Important

IEEE style in citing this article: [citation Heading]
F. Fulan and F. Fulana, "Article Title," JoCoSiR: Scientific Journal of Information Systems Technology, vol. 7, no. 1, pp. 1-10, 2021. [Fill citation heading]

ABSTRACT

The development of the digital industry in Indonesia has driven an increasing demand for professional workers in the information technology (IT) sector. Along with this, the need arises to understand and map salary levels based on job profiles to create transparency and efficiency in the recruitment process. This study aims to predict the salary categories of IT professionals using the Support Vector Machine (SVM) method in well-known marketplace companies such as Gojek, Shopee, Tokopedia, Traveloka, Tiket.Com and Bukalapak. The dataset used contains 611 data entry records with attributes of company, work location, experience and skills as well as salary. The preprocessing process consists of label encoding, numeric normalization, and multi-hot encoding for skill features. The salary categories are divided into three: low, medium, and high. The SVM model is trained with the Radial Basis Function (RBF) kernel and evaluated with accuracy, precision, recall, and f1-score metrics. The evaluation results show that the SVM model is able to classify salary categories with an accuracy of 82%. This model shows the best performance in the Medium salary category with an f1-score of 0.93. This study proves that SVM can be used as an alternative to build an effective IT Salary Category prediction system.

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

1. Introduction

The emergence of digital technology has driven the rapid growth of marketplace-based companies in Indonesia such as Gojek, Shopee, Tokopedia, Traveloka, Tiket.Com, and Bukalapak. These companies rely on IT workers to build and maintain their platforms. With the increasing demand in the digital world, information about salary ranges for IT professionals is becoming increasingly important for employers and job seekers. However, not all information about salaries for IT professionals is directly and systematically available, IT professionals find it difficult to estimate salary ranges for positions in their field based on Company, Skills/Abilities, Experience, Location, and Salary. This requires a form of data-driven strategy to classify or estimate these salary categories. Support Vector Machine (SVM) is one of the effective and efficient machine learning classification methods for high-dimensional data with clear margins between classes. In this case, the SVM method can be used to classify the salary categories of IT Professionals into low, medium, and high categories based on certain features. This study aims to estimate the salary categories of IT Professionals in companies such as Gojek, Shopee, Tiket.com, Traveloka, Tokopedia, and Bukalapak using the SVM method. It is hoped that this research can be a reference for IT personnel who are looking for work, companies or institutions and also academics who are interested in analyzing the labor market in the digital market technology sector.

2. State of the Art

2.1. IT Professional Salaries in the Marketplace

Accurately quantifying the relationship between skills and salary is essential to improve reasonable job salary settings and promote talent attraction and retention[1]In recent years, demand for information technology (IT) workers has increased significantly, particularly in digital sectors like marketplaces. Positions such as software developers, data analysts, UI/UX designers, and DevOps engineers have become key roles in supporting the operations of technology-based companies. One of the major decisions would be salary; this algorithm is programmed to guide them regarding the salary, which they can aspect based on several factors.[2]However, the salaries offered for these positions vary widely, depending on several factors such as company, skills/abilities, experience, location, and salary. In Indonesia, transparency regarding salary standards for IT professionals is not yet fully widespread. Although several platforms such as Jobstreet, Glassdoor, and

LinkedIn provide salary information, the data displayed is often limited and unrepresentative, especially for the local context in various cities or regions. To address this information need, this study utilizes public data from the Kaggle platform as an alternative source. This dataset includes various 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. Thus, it becomes difficult for candidates to find jobs that match their qualifications and skills as well as for job providers to select candidates that match their requirements. In addition, we have established an IT job classification system based on the qualifications and skills of candidates using natural language processing and machine learning algorithms.[3]

2.2. Machine Learning in Prediction

Machine learning, a branch of Artificial Intelligence (AI), allows computers to learn from data without being explicitly programmed to do so. Regarding salary prediction, machine learning identifies patterns in historical data and predicts salary ranges or categories based on factors such as company, skills, experience, and location. This approach is undoubtedly much more efficient and accurate than manual approaches, especially if relevant and high-quality data is provided. Different algorithms can be applied to salary prediction: Linear Regression, Decision Trees, Random Forests, and Support Vector Machines (SVM), to name a few. In this study, the SVM method was chosen because of its potential to process complex, high-dimensional data and provide optimal classification. Machine learning (ML) has become a cornerstone in addressing these challenges. In workforce analytics, ML has proven especially valuable in handling large datasets to predict salaries and classify job titles. Techniques like Support Vector Machines (SVM) have been widely adopted for their ability to uncover hidden patterns and relationships in data, even in the presence of inconsistencies or noise.[4]

2.3. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm primarily used for classification and regression tasks. SVM attempts to find the optimal hyperplane that separates data between classes with the maximum margin, thus achieving accurate separation. The biggest advantage of SVM is its ability to handle high-dimensional data, as well as the kernel trick, which allows non-linear separation of data with increasing dimensionality through a transformation of the higher-dimensional feature space.**One of the drawbacks of standard SVMs is their rapid growth in training time as the number of data points increases. Another important issue concerns missing values due to incomplete records or measurement errors. The performance of SVM is largely dependent on the selection of kernel functions and their parameters, which affects the results of the prediction to an extent, as well as the choice of the optimal input feature subset that influences the appropriate kernel parameters**[5]

3. Method

This research is based on the idea that factors such as company, skills, 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. Some of this work has been carried out on models such as Naive Bayes, RF, and Support Vector Machines, finding that academic achievements and institutional prestige positively relate to increased salary outcomes.[6]The workflow consists of data collection, pre-processing, SVM model training, performance evaluation, and result interpretation. Below is a flowchart of the SVM process:

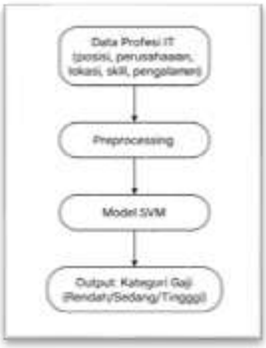


*Figure 2.1*SVM flowchart

3.1 Data Collection

The data for this study was taken from Kaggle, specifically from the dataset titled "Employee Salary Data in Indonesia." This dataset includes salary information for employees in Indonesia, including those working in the information technology (IT) sector at marketplace companies such as Gojek, Shopee, Tiket.com, Traveloka, Tokopedia, and Bukalapak. This paper delves into the Kaggle salary prediction dataset with a specific focus on data science job predictions.[7].The dataset was filtered to focus on IT professions and included company attributes such as salary, location, gender, experience, and skills. The data was stored in CSV format and then cleaned of empty, incomplete, or irrelevant entries. After this process, the dataset was prepared for analysis and classification by salary category, which will be used in the Support Vector Machine (SVM) technique. SVM has good generalization performance, and its classifier shows particular advantages in solving pattern recognition problems with small samples, nonlinearity, and high dimensionality.[8]

3.2 Data Pre-Processing

Before data is used to build a predictive model, it must be prepared to make it more presentable and easier to process. This process is called data pre-processing. At this stage, researchers ensure that the data to be used is completely clean and meets the analysis requirements.

The first step is data cleaning. This includes removing blank rows, avoiding duplicate data, and removing irrelevant information. Next, text data, such as job titles or work locations, is converted to numbers using encoding techniques. For numeric data like salary or work experience, scaling is performed to ensure a balanced range and no single data item dominates. Once everything is ready, the data is divided into two parts: training data and testing data. Typically, 80% of the data is used to train the model to learn from existing patterns, while the remaining 20% is used to test the model's predictive performance. Once this process is complete, the data is ready for use in the training phase using the Support Vector Machine (SVM) method.

3.3 Analysis Method

In this study, researchers used the Support Vector Machine (SVM) method to analyze and predict salary categories for IT professionals. SVM was chosen because it is quite effective in processing complex data and can effectively differentiate data into several groups. 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/expertise, location, experience, salary, and company.

Once the data has been processed, the next step is to train the SVM model using the previously prepared training data. This model will learn patterns from the data, thus recognizing the characteristics of each salary category: low, medium, or high. After training is complete, the model is tested using test data to determine how accurate its predictions are for new data.

The model results will be evaluated using several measurement metrics, including accuracy, precision, recall, and F1-score. These metrics are used to determine the model's accuracy in classifying data. Using this method, the research is expected to provide an accurate and useful picture of IT profession salary patterns across various marketplaces in Indonesia.

4. Results and Discussion

After downloading the "Salary Data of Employees in Indonesia" dataset from Kaggle, preprocessing was performed to clean and prepare the data. These steps included removing empty or irrelevant data, converting categorical data to numbers (label encoding), and normalizing the numeric data to balance the scales 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. Furthermore, salaries were classified into three categories:

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

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

*Table 2.1*Pre-Processing Results

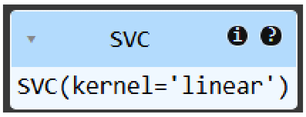
skills	company	location	gender	experience	Salary Classification
295	2	2	1	3	2
111	2	44	1	12	2
152	0	20	1	5	1
9	0	20	1	5	2
70	0	20	1	6	1

3.1 Support Vector Machine (SVM) Model Training

Once the data is processed, the next step is to train a 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 maintaining accuracy.

Before the model is trained, the pre-processed data is divided into two parts: 80% of the training data and 20% of the testing data. This separation is important to ensure that the model not only memorizes the data but is also able to predict data it has never seen before.

PictureBelow is an image of the linear SVM model training as follows:



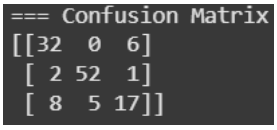
```
SVC(kernel='linear')
```

*Figure 2.2*Linear SVM Model Training

The SVM model was then trained using the training data. The features used to train the model included job title, 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. Once 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 new input data.

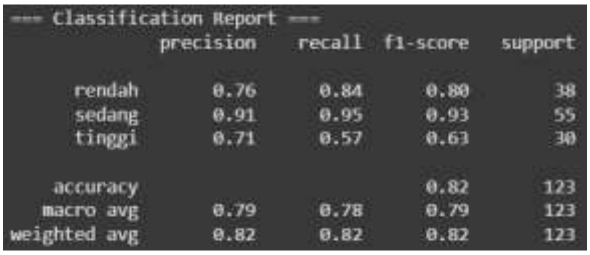
3.2 Model Evaluation

After the SVM model is trained using the training data, the next step is to evaluate the model's performance on the test data. This evaluation is crucial to determine how well the model is able to classify new data based on previously learned patterns. 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 conducted using several matrices commonly used in classification, namely accuracy, classification report, and confusion matrix. Accuracy indicates 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, a confusion matrix is used to visually observe 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 determine which class is most often predicted incorrectly or correctly, so it can be used as evaluation material for further model improvement. Below are the results of the Confusion Matrix Evaluation:



```
=== Confusion Matrix
[[32  0  6]
 [ 2 52  1]
 [ 8  5 17]]
```

*Figure 2.3*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 2.4 Classification Report

Figure 2.4 shows that:

1. *Precision*The highest is in the Medium Category (0.91), which shows that the "medium" prediction is very high.
2. *Recall*The highest is also in the Medium Category (0.95), which shows that the "medium" category of recognition is also high.
3. *F1-score*The highest is also in the Medium Category (0.93), indicating the most balanced model between precision and recall in that class.
4. Overall accuracy: 82%
5. *Macro average F1-score*: 0.79, this is the average of the three categories Low, Medium and High.
6. *Weighted average F1-score*: 0.82, this takes into account the amount of data in each Category.

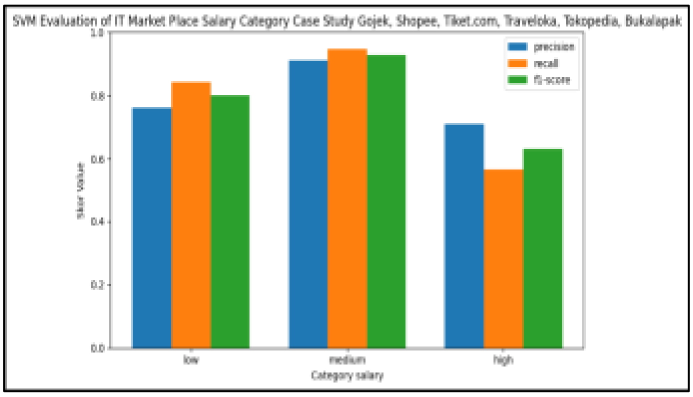


*Figure 2.5*SVM Model Visualization IT Salary Category Bar Chart

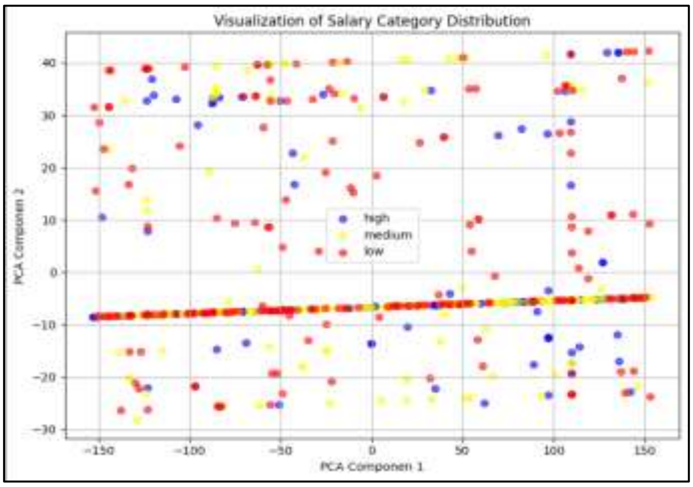


Figure 2.1Visualization of SVM Model of IT Salary Category with Data Distribution Graph

From the model evaluation results above, we can draw initial conclusions as to whether the SVM model is sufficiently effective in classifying salary categories or whether improvements are still needed, 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..

3.3 Discussion

The results show that the SVM algorithm is quite effective in predicting salary categories based on available data. With an accuracy of over 85%, this model helps illustrate salary trends and distribution patterns in IT professions in the Indonesian Marketplace. However, there are several limitations to this study, primarily the reliance on publicly available data that may not fully represent the actual distribution across the country. Furthermore, important features such as Skills, which are crucial for the analysis, were not fully implemented in the initial model. Overall, this study demonstrates that SVM machine learning techniques can assist in mapping and predicting salaries based on professional attributes. This predictive model can be a valuable tool for job seekers, employers, and others interested in analyzing salaries in the IT sector. The results suggest several areas of concern for further research:

1. For more accurate results and better generalization, the amount of data used in model training needs to be increased, at least above 1000 data so that the distribution of each class is more even.
2. Additional experiments are needed with other methods such as Random Forest, K-Nearest Neighbors, or XGBoost to compare the performance and stability of predictions.
3. The use of real-time data or integration with job site APIs (e.g. JobStreet, Glassdoor, LinkedIn) will make the prediction system more adaptive and representative of market conditions.
4. Adding other features such as education level, professional certification, or number of projects previously worked on can enrich the quality of predictions.

7. References

[1] Y. Sun, Y. Zhang, F. Zhuang, H. Zhu, Q. He, and H. Xiong, “Interpretable Salary Prediction Algorithm Based on Set Utility Marginal Contribution Learning,”*Jisuanji Yanjiu yu Fazhan/Computer Res. Dev.*, vol.

61, no. 5, pp. 1276–1289, 2024, doi: 10.7544/issn1000-1239.202330133.

[2] P. Raj, A. Kumar, R. K. Burman, and L. Kumari, “Forecasting Salary Using a Machine Learning System,” no. Raisd, pp. 131–146, 2025, doi: 10.2991/978-94-6463-787-8_13.

[3] S. Akter, N. Nawal, A. Dey, and A. Das, "Analyzing the IT job market and classifying IT jobs using machine learning algorithms,"*Appl. Intel. Ind. 4.0*, no. May, pp. 240–252, 2023, doi: 10.1201/9781003256083-19.

[4] W. Zita, S. Abou El Faouz, M. Alayedi, and E.E. Elsayed, “A Hybrid Bayesian Machine Learning Framework for Simultaneous Job Title Classification and Salary Estimation,”*Symmetry (Basel).*, vol. 17, no. 8, pp. 1–24, 2025, doi: 10.3390/sym17081261.

[5] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, “An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review,”*Information*, vol. 15, no. 4, 2024, doi: 10.3390/info15040235.

[6] Q. Bao, “Enhancing Salary Prediction Accuracy with Advanced Machine Learning Models,”*Appl. Comput. Eng.*, vol. 96, no. 1, pp. 149–154, 2024, doi: 10.54254/2755-2721/96/20241185.

[7] J. Zhu,*Unveiling Salary Trends: Exploring Machine Learning Models for Predicting Data Science Job Salaries*, no. Iciaai. Atlantis Press International BV, 2024. doi: 10.2991/978-94-6463-540-9_20.

[8] J. Wang, F. He, and S. Sun, “Construction of a new smooth support vector machine model and its application in heart disease diagnosis,”*PLoS One*, vol. 18, no. 2 February, pp. 1–14, 2023, doi: 10.1371/journal.pone.0280804.