JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

Comparison of Prediction Models: Decision Tree, Random Forest, and Support Vector Regression

Kurnia Ramadhan Putra
Information System, Faculty of Industrial Technology, Institut Teknologi Nasional Bandung, Indonesia
kurniaramadhan@itenas.ac.id

Abstract: The Information Technology (IT) industry continues to grow rapidly, creating challenges in determining fair and competitive salaries for professionals. Accurate salary predictions are essential for companies to attract and retain talent while providing insights for individual career planning. This study aims to evaluate and compare the performance of three machine learning models, such as Decision Tree Regression, Random Forest Regression, and Support Vector Regression in forecasting salaries in the IT sector using demographic and professional factors such as age, gender, education level, job title, and work experience. The study uses a dataset of 6,704 entries from Kaggle, with relationships between variables analyzed through statistical techniques such as Pearson Correlation and ANOVA. The models' performance was evaluated using the R^2 score, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Among the models, Random Forest Regression exhibited the best results, achieving the highest R^2 of 91.49% and an RMSE of 0.058, reflecting strong predictive accuracy with minimal errors. Scatter plot visualizations confirm a strong correlation between actual and predicted salaries, supported by error analysis identifying minimal overestimation and underestimation cases. The research concludes that Random Forest Regression is the most effective model for IT salary predictions. These findings provide practical insights for organizations and individuals, highlighting the potential of data-driven approaches in salary determination. Future studies may focus on hyperparameter optimization and incorporating additional features to improve model performance and generalizability further improve model performance and generalizability.
Keywords: IT Salary Prediction; Machine Learning Models; Random Forest Regression; Predictive Analytics; Demographic and Professional Data

1. INTRODUCING

The Information Technology (IT) industry is a dynamic and rapidly evolving field, offering abundant opportunities but also posing significant challenges. Among these challenges is the determination of competitive and equitable salaries for professionals, which is critical for organizations to attract and retain top talent. Salary determination in the IT sector is influenced by various factors, including age, education level, work experience, gender, job position, and geographical location. These multifaceted factors make it challenging to develop a standardized approach for salary prediction [1]–[3].

Traditional salary estimation methods often rely on descriptive statistical approaches or expert judgment, which may not effectively capture the complex interdependencies among the influencing factors. Recent developments in data analytics and machine learning have provided robust tools for predictive modeling, allowing organizations to make data-driven decisions in salary determination. Among the various machine learning models available, Decision Tree Regression [4], Random Forest Regression [5]–[8], and Support Vector Regression (SVR) [9] have shown promising results in handling structured salary datasets.

Kurnia Ramadhan Putra: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

The research problem in salary prediction using machine learning lies in the challenge of selecting the most appropriate model based on accuracy, interpretability, and computational efficiency. While Decision Tree Regression is favored for its simplicity and ease of interpretation, it often struggles with overfitting when applied to complex datasets. Random Forest Regression, as an ensemble of multiple decision trees, improves prediction accuracy but can lack transparency due to its complexity. Meanwhile, Support Vector Regression is effective in handling high-dimensional data and non-linear relationships but is computationally intensive and requires meticulous parameter tuning [10]. Given these factors, this study aims to conduct a comparative analysis of these models to determine their strengths and weaknesses in IT salary prediction. The research will assess which model delivers the most accurate predictions for IT professionals, how performance metrics such as R^2 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) [11] vary across the models, and what key factors influence salary prediction. Additionally, statistical techniques such as Pearson Correlation and ANOVA will be employed to better understand the relationships between salary-determining variables.

Many studies have explored salary prediction using machine learning, applying various models with different levels of success. Regression models such as Linear Regression and Polynomial Regression have been commonly used but often struggle with capturing complex relationships between factors affecting salary [12]. Artificial Neural Networks (ANNs) have shown better accuracy but require large datasets and careful tuning to work effectively [13]. Other studies have used ensemble methods like Gradient Boosting Machines (GBM) and XGBoost, which improve prediction accuracy by combining multiple models. However, these methods are often more complex and computationally demanding, making them less practical for quick predictions [14]. Some researchers have explored hybrid models or feature selection techniques like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) to improve efficiency and reduce unnecessary data [15]. Despite these advancements, many studies focus only on accuracy without discussing the trade-offs between model complexity, interpretability, and efficiency. Moreover, few studies analyze which model is best suited for salary prediction in the IT industry. This research aims to fill that gap by comparing Decision Tree Regression, Random Forest Regression, and Support Vector Regression based on their accuracy, speed, and ease of interpretation.

This study aims to provide a comparative analysis of Decision Tree Regression, Random Forest Regression, and Support Vector Regression for salary prediction in the IT industry. By leveraging a dataset of 6,704 entries sourced from Kaggle, which includes demographic and professional attributes, the research will evaluate and compare the predictive performance of the three models using R^2 Score, MAE, and RMSE. It will also analyze the relationships between salary-influencing factors using Pearson Correlation and ANOVA [16]. Furthermore, the study seeks to identify the most effective model for salary prediction based on accuracy, interpretability, and computational efficiency. Finally, it will provide recommendations on the most suitable machine learning model for salary prediction in different contexts within the IT industry.

By addressing the identified research gaps, this study aims to contribute to the broader discourse on data-driven salary prediction. The findings will help organizations implement more accurate salary estimation methods and assist IT professionals in making informed career decisions. Additionally, this research will serve as a foundation for future studies exploring machine learning applications in salary prediction beyond the IT industry [17].

2. RESEARCH METHODOLOGY

This research employs a quantitative approach by exploring and modeling data using machine learning algorithms to develop and evaluate salary prediction models in the IT sector. The methodology consists of several stages, including:

2.1 Data Collection

The dataset, obtained from credible platforms like Kaggle, consists of 6,704 records with essential attributes such as age, gender, education, job title, work experience, and monthly salary. Exploratory

Kurnia Ramadhan Putra: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

Data Analysis (EDA) was performed to analyze data patterns, detect anomalies, and handle missing values [18]–[20].

2.2 Data Preparation

Data Pre-processing

During data preprocessing, various steps were implemented to prepare the dataset for modeling. Categorical variables, including gender and job title, were converted into numerical values using label encoding to enable their use in machine learning models. Furthermore, numerical features were scaled through normalization or standardization to enhance the models' performance like Support Vector Regression (SVR), which are particularly affected by feature scaling.

Correlation Analysis

A correlation analysis was carried out to gain deeper insights into the relationships and interdependencies among variables. Pearson Correlation was employed to evaluate the relationships between numerical features. Additionally, Analysis of Variance (ANOVA) was utilized to explore the connections between categorical and numerical variables, offering valuable insights into their interrelationships.

2.3 Model Building

Machine Learning Methods

Decision Tree Regression was utilized to build a model that is both easy to interpret and efficient during the training process. Its clear and intuitive visualizations provide a transparent representation of the decision-making process. Random Forest Regression, on the other hand, was employed to enhance accuracy by aggregating the outputs of multiple decision trees. This approach not only improves the model's overall robustness but also helps reduce the risk of overfitting. Finally, SVR was applied to address non-linear patterns in the data. By using an appropriate kernel function, SVR allows the model to effectively capture intricate relationships within the dataset.

Model Optimization

GridSearchCV was used to perform hyperparameter tuning, allowing for the identification of the optimal parameters for each model to improve their performance. By systematically testing various combinations of hyperparameters, this technique helped determine the best configuration for each algorithm, thereby enhancing prediction accuracy [21].

2.4 Model Evaluation

Evaluation Metrics

The R^2 score was utilized to assess the model's ability to explain the variability in the data, where a higher score indicates a better alignment between the model and the data. MAE was used to calculate the average absolute errors in the model's predictions, providing insight into the typical magnitude of the errors without accounting for their direction. RMSE was employed to evaluate the average squared differences between predicted and actual values. This metric is particularly responsive to significant errors, as they disproportionately affect the overall score, making it valuable for identifying outliers.

2.5 Testing

Prepare Test Data

Extract the reserved test set from the dataset, which was not used during model training and validation. Ensure the test data undergoes the same preprocessing steps applied to the training set, such as label encoding for categorical variables and normalization/standardization for numerical variables.

Kurnia Ramadhan Putra: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

Load the Best Model

Select the model with the highest performance based on evaluation metrics (e.g., Decision Tree Regression, Random Forest Regression, or Support Vector Regression). Load the model with the optimal hyperparameters obtained during the GridSearchCV optimization process.

Perform Predictions

Apply the best-performing model to the test set to generate predictions for the target variable (monthly salary). Save these predictions as "Predicted Results then compare with the corresponding "Actual Results" from the test set.

Calculate Error Metrics

To gauge the model's effectiveness on the test set, various metrics were determined. First, the Error Rate was determined by calculating the absolute disparity between the forecasted and observed values for each data point, providing a direct measure of prediction errors. Next, the Mean Absolute Error (MAE) was computed by averaging the absolute differences across all data points, offering a summary of the average magnitude of errors without considering their direction. Lastly, the Root Mean Squared Error (RMSE) was calculated. This involves calculating the square root of the mean of the squared differences between the forecasted and observed. This metric places greater emphasis on more significant errors, making it especially useful for identifying significant deviations in the predictions. Together, these metrics offer a thorough assessment of the model's accuracy and overall performance. The research steps are illustrated in Fig. 1.

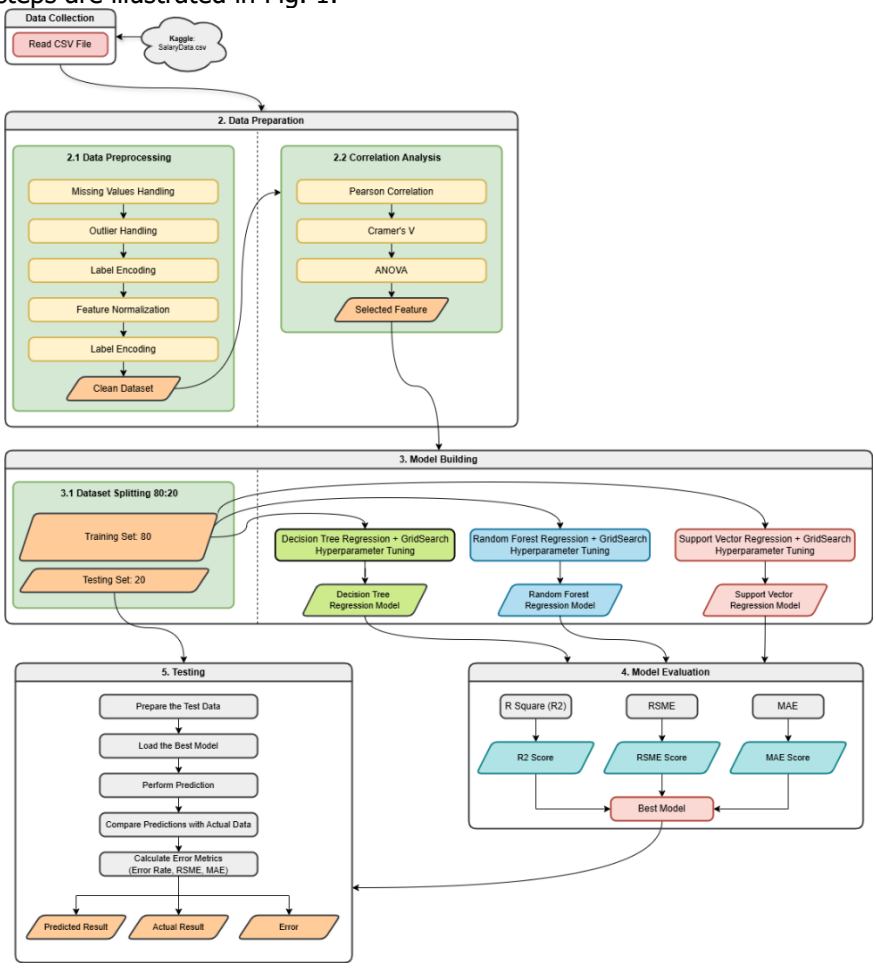

Figure 1. Research Methodology

Kurnia Ramadhan Putra: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

Figure 1 explains the methodology for this research follows a structured approach to developing an accurate salary prediction model using machine learning techniques. The process consists of five main stages: data collection, data preparation, model building, testing, and model evaluation. Each stage is designed to ensure that the data is properly processed, relevant features are selected, and the best-performing model is identified.

The first stage, data collection involves obtaining the dataset then is loaded into the system to be prepared for analysis. In the data preparation stage, the dataset undergoes two key processes: data preprocessing and correlation analysis. Data preprocessing includes handling missing values, detecting and removing outliers, applying label encoding to categorical variables, and normalizing features to ensure consistency. Meanwhile, correlation analysis is performed using Pearson Correlation, Cramér's V, and ANOVA to identify the most relevant features for salary prediction.

In the model building stage, the dataset is split into an 80:20 ratio, where 80% is used for training and 20% for testing. Three machine learning models are implemented: Decision Tree Regression, Random Forest Regression, and Support Vector Regression (SVR). Each model is fine-tuned using GridSearch hyperparameter tuning to optimize performance. The decision tree regression model helps in understanding feature importance, while the random forest regression model leverages ensemble learning for improved accuracy. SVR is included to capture potential non-linear relationships in the data.

The testing phase ensures that the trained models are evaluated on unseen data. The process involves preparing the test data, loading the best-trained model, performing predictions, comparing the predicted results with actual values, and calculating error metrics such as error rate, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics help assess the models' predictive capabilities and generalization performance.

Finally, in the model evaluation stage, the models are assessed based on three key performance metrics: R-Square (R^2 Score), RMSE, and MAE. The R^2 Score measures the proportion of variance explained by the model, while RMSE and MAE quantify prediction errors. The best model is selected based on the highest R^2 Score and the lowest RMSE and MAE values, ensuring optimal accuracy and reliability for salary prediction in the IT sector. This structured methodology ensures that the research is data-driven, interpretable, and optimized for performance.

3. RESULT AND DISCUSSIONS

3.1 Stages of Applying Method

The stage begins with the data collection phase, where salary data is obtained from Kaggle (SalaryData.csv). This dataset forms the foundation for predictive modeling, requiring preprocessing to ensure data quality before applying machine learning algorithms.

In the data preparation stage, the dataset undergoes preprocessing to handle missing values, detect and manage outliers, apply label encoding for categorical variables, and normalize numerical features to ensure consistency. The data is then cleaned to remove inconsistencies and redundant entries. Additionally, correlation analysis is performed to select the most relevant features for prediction. Pearson correlation is used to analyze linear relationships between numerical variables, Cramer's V evaluates associations between categorical variables, and ANOVA determines the significance of different feature categories. The most impactful features are then selected for model training.

Next, the model building process begins with splitting the dataset into an 80:20 ratio, where 80% of the data is used for training, and 20% is reserved for testing. Three machine learning algorithms are implemented: Decision Tree Regression, which creates a tree-based predictive model; Random Forest Regression, an ensemble learning method that combines multiple decision trees for improved performance; and Support Vector Regression (SVR), which captures complex relationships between variables using kernel functions. Each model undergoes GridSearch hyperparameter tuning to optimize performance.

The model evaluation phase assesses each trained model using three performance metrics. R-Square (R^2) measures how well the independent variables explain the variance in salary predictions. Root Mean Square Error (RMSE) quantifies the average deviation between predicted and actual salaries, while Mean

Kurnia Ramadhan Putra: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

Absolute Error (MAE) represents the mean of absolute differences between predictions and actual values. The model with the best overall performance is selected for deployment.

In the testing and validation phase, the final model is validated using unseen test data. The test set undergoes preprocessing similar to the training data. The best-performing model is then loaded, and predictions are generated. The predicted salary values are compared with actual salaries from the dataset, and error calculations are performed using R^2, RMSE, and MAE. These results provide insights into IT salary prediction trends, highlighting the most effective model for accurate salary estimation is shown in Figure 2.

Figure 2. Result of prediction

3.2 Model Selection Result

Table 1 presents the results of hyperparameter tuning for each model, detailing their respective parameters and the corresponding Negative Mean Squared Error (-MSE) scores. In Scikit-learn, negative MSE values are used for optimization, with a higher score (closer to zero) indicating better model performance. The results reveal that Random Forest Regression achieved the best performance, with a -MSE score of -3.901513 when using n_estimators = 80, demonstrating its ability to minimize prediction errors effectively and making it the most accurate model among those evaluated. Support Vector Regression (SVR) achieved a -MSE score of -5.930423 with parameters C = 10, gamma = "scale", and kernel = "rbf", performing well in capturing non-linear relationships, though slightly less effective than Random Forest Regression. On the other hand, Decision Tree Regression performed the least well, with a -MSE score of -5.944865 and parameters max_depth = 15, min_samples_split = 10, and random_state = 0. While Decision Tree Regression is computationally efficient and interpretable, its susceptibility to overfitting resulted in lower performance compared to the other models.

These results confirm that Random Forest Regression is the most effective model for predicting salaries in the IT sector, based on its ability to minimize error. The superior performance of this model can be attributed to its ensemble nature, which reduces overfitting by aggregating predictions from multiple decision trees.

Table 1. Best Model Selection

Model	Params	-MSE
Random Forest Regression	n_estimators = 80	-3.901513
SVR	C = 10, gamma: "scale", kernel: "rbf"	-5.930423
Decision Tree Regression	max_depth=15, min_samples_split=10, random_state=0	-5.944865

Scoring functions are defined as metrics that should be maximized to indicate the best performance. Mean Squared Error (MSE), on the other hand, is an evaluation metric where lower values indicate

Kurnia Ramadhan Putra: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

better performance (lower is better) because it measures prediction errors. To align MSE with the optimization pattern, it must be measured using Negative MSE is show in Equation (1).

$$Negative\ MSE = -1 * MSE \quad (1)$$

By using the negative value, scikit-learn can identify the highest (maximized) score, which corresponds to the smallest MSE value approaching zero. It was observed that the smallest MSE value was achieved by the Random Forest Regressor. To further confirm this result, training was conducted for each model to validate and compare their performance thoroughly. Meanwhile, MSE is shown in Equation (2).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

Where,
n = total number of data points
y_i = actual value of the i^{th} data point
$\hat{y}_i$ = predicted value of the i^{th} data point
$(y_i - \hat{y}_i)^2$ = squared difference between actual and predicted values.

3.3 Model Evaluation Result

In this study, three salary prediction models for the IT sector were developed using Decision Tree Regression, Random Forest Regression, and Support Vector Regression (SVR). The dataset consisted of 6,704 records, including features such as age, gender, education level, job position, work experience, and monthly salary. The effectiveness of each model was assessed through metrics including the R^2 Score, MAE, and RMSE, with a summary of the results provided in Table 2.

Table 2. Model Evaluation

Model	R^2 Score (%)	MSE	MAE	RSME
Random Forest Regression	91.49	0.003338	0.038295	0.057779
SVR	85.12	0.005836	0.060341	0.076399
Decision Tree Regression	89.20	0.004238	0.045427	0.065104

The R^2 Score

The R^2 score indicates the proportion of variability in the dependent variable that is accounted for by the independent variables. Ranging from 0 to 100%, a higher R^2 score means the model is better at explaining the variation in the target variable. Among the models evaluated, Random Forest Regression achieved the highest R^2 score of 91.49%, meaning it accounts for 91.49% of the variance in the target variable, demonstrating strong predictive accuracy and a good fit to the data. The Support Vector Regression (SVR) model, with an R^2 score of 85.12%, explains 85.12% of the variance, which is a respectable result but slightly less precise compared to Random Forest Regression. Decision Tree Regression, with an R^2 score of 89.20%, explains 89.20% of the variance, positioning it between Random Forest and SVR in terms of predictive accuracy.

Mean Squared Error (MSE)

MSE measures the average squared difference between predicted and actual values, where a lower MSE indicates better model performance. Among the evaluated models, Random Forest Regression achieved the lowest MSE of 0.003338, indicating its superior ability to minimize prediction errors. In contrast, Support Vector Regression (SVR) recorded the highest MSE at 0.005836, reflecting relatively weaker performance. Decision Tree Regression, with an MSE of 0.004238, performed moderately, falling between Random Forest Regression and SVR in terms of accuracy.

Kurnia Ramadhan Putra: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

Mean Absolute Error (MAE)
MAE determines the average absolute differences between the predicted values and the true values, disregarding their direction, where a lower MAE indicates greater accuracy. Random Forest Regression demonstrated the best performance with the lowest MAE of 0.038295, signifying its superior precision in minimizing absolute errors. On the other hand, Support Vector Regression (SVR) exhibited the highest MAE at 0.060341, indicating larger average errors compared to the other models. Decision Tree Regression achieved a moderate MAE of 0.045427, placing its accuracy between that of Random Forest Regression and SVR.

Root Mean Squared Error (RMSE)
MSE gives an indication of the average error size, measured in the same units as the target variable, and helps to understand the standard deviation of the residuals. Among the evaluated models, Random Forest Regression achieved the lowest RMSE of 0.057779, underscoring its superior predictive accuracy. Support Vector Regression (SVR), on the other hand, recorded the highest RMSE at 0.076399, reflecting larger prediction errors relative to the other models. Decision Tree Regression demonstrated a moderate RMSE of 0.065104, positioning its performance between Random Forest Regression and SVR.

3.4 Strenghts and Weaknesses of the Models

Random Forest Regression
Random Forest Regression offers several advantages, including its ability to combine multiple decision trees, which effectively reduces overfitting and enhances accuracy, particularly when dealing with complex datasets. However, a notable drawback is its longer training time compared to Decision Tree Regression. Despite this, the findings indicate that Random Forest Regression delivered superior performance, reaching an R^2 Score of 91.49%, an MSE of 0.003338, an MAE of 0.038295, and an RMSE of 0.057779. These results highlight its capability to produce highly accurate and consistent predictions, making it the most effective model among those evaluated.

Support Vector Regression
Support Vector Regression is particularly strong in handling non-linear relationships within data by employing kernel functions, making it well-suited for capturing complex patterns. However, its limitations include the need for extensive hyperparameter tuning, which can significantly increase training time, and its reduced interpretability compared to simpler models like Decision Tree Regression. The findings indicate that SVR performed reasonably well, reaching an R^2 Score of 85.12%, an MSE of 0.005836, an MAE of 0.060341, and an RMSE of 0.076399. While the model effectively captured non-linear patterns, its overall performance was slightly inferior to that of Random Forest Regression.

Decision Tree Regression
Decision Tree Regression stands out for its simplicity and interpretability, thanks to its tree-based structure, making it easy to understand and visualize. It also offers fast training times, particularly when dealing with large datasets. However, the model is susceptible to overfitting, especially when using deep trees, which can limit its generalization ability. Its performance tends to fall short compared to more robust models like Random Forest Regression and SVR. In this study, Decision Tree Regression achieved an R^2 Score of 89.20%, reflecting good accuracy. Nonetheless, its higher MAE and RMSE values suggest that its predictions were less precise than those of the other models.

3.5 Analysis of Significant Factors

Correlation analysis using Pearson Correlation and ANOVA revealed several key factors influencing salary predictions. Age and work experience showed a strong positive correlation with salary, indicating that individuals with greater age and experience tend to earn higher wages. Job position emerged as the most significant factor, as it directly impacts salary variations due to the distinct pay scales associated with different roles. Education level also positively affects salary, although its influence is weaker compared to work experience and job position, suggesting that while qualifications matter, practical experience and role responsibilities play a more critical role in determining salary.

Kurnia Ramadhan Putra: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

Analysis of Significant Factor Using Pearson Correlation

The analysis of significant factors using Pearson Correlation, see Fig. 2.

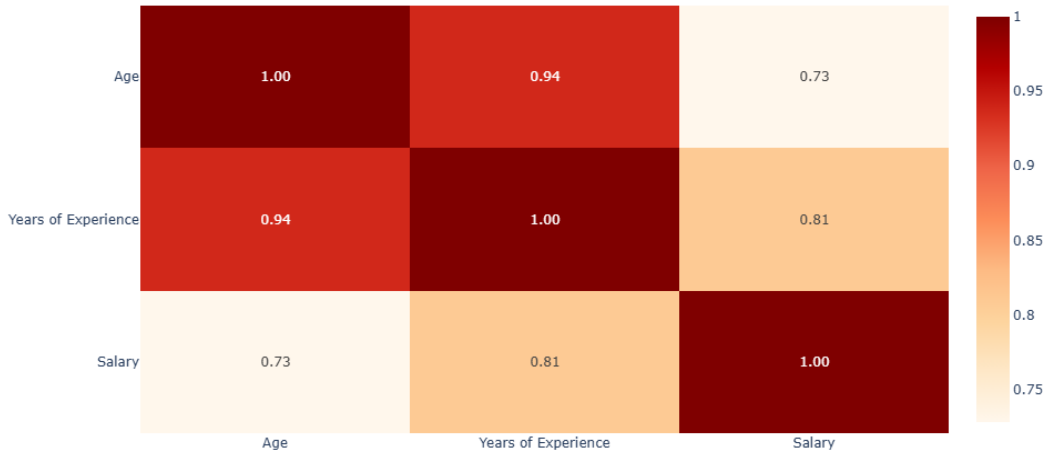

Figure 3. Pearson Correlation

The explanation of the figure describes how the Pearson correlation coefficient (r) is used to measure the strength and direction of the relationship between two variables.

Interpretation of Correlation Strength

The Pearson correlation coefficient (r) quantifies the degree and direction of the linear association between two variables, with values ranging from -1 to +1. A coefficient of r = -1 indicates a perfect negative correlation, where an increase in one variable corresponds to a predictable decrease in the other. Conversely, a coefficient of r = +1 signifies a perfect positive correlation, where both variables increase predictably together. A value of r = 0 denotes the absence of any linear association between the variables.

The strength of the correlation is typically categorized as follows: very weak ($0.00 \leq |r| < 0.30|$), weak ($0.30 \leq |r| < 0.50$), moderate ($0.50 \leq |r| < 0.70$), strong ($0.70 \leq |r| < 0.90$), and very strong ($0.90 \leq |r| \leq 1.00$). These classifications provide insight into how closely two variables are linearly associated, with stronger correlations indicating a more reliable and predictable relationship.

Correlation Analysis in the Example

The correlation analysis reveals key relationships among Salary, Age, and Experience based on their Pearson correlation coefficients. The correlation between Salary and Age (r = 0.73) indicates a strong positive relationship, as the value falls within the range $0.70 \leq r < 0.90$. This implies that an increase in Age is significantly associated with a rise in Salary. Similarly, the correlation between Salary and Experience (r = 0.81) also reflects a strong positive relationship, suggesting that higher levels of Experience are strongly linked to increased Salary.

Furthermore, the relationship between Age and Experience exhibits a very strong positive correlation with r = 0.94, which falls within the range $0.90 \leq r \leq 1.00$. This nearly perfect linear relationship indicates that as Age increases, Experience almost invariably increases as well, highlighting their close connection in the dataset.

The analysis indicates that as one variable rises, the other also tends to increase. The strength of these correlations varies, ranging from strong to very strong, depending on the specific pair of variables

Kurnia Ramadhan Putra: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

being examined. Notably, the relationship between Age and Experience stands out as the most significant, with a correlation coefficient of 0.94, suggesting an almost perfect linear association between the two variables.

Analysis of Significant Factors Using ANOVA

This analysis assesses the relationship between categorical variables, included Gender, Job Title, and Education Level and a numerical variable like Salary using ANOVA. Additionally, the analysis of significant factors through Pearson Correlation is provided in Table 3.

Table 3. Summary of Correlations

Correlation	F-statistic	p-value	Significance
Gender and Salary	55.3276	0.0000	significant
Job Title and Salary	67.7184	0.0000	significant
Education Level and Salary	1045.1541	0.0000	significant

The ANOVA analysis highlights key relationships between categorical variables and salary. Gender has a significant effect on salary, with the F-statistic of 55.3276 and a p-value of 0.0000, indicating that average salaries vary based on gender. Similarly, Job Title also plays a crucial role, as evidenced by an F-statistic of 67.7184 and a p-value of 0.0000, suggesting that salaries differ across various job titles. Furthermore, Education Level demonstrates the strongest impact on salary, with an F-statistic of 1045.1541 and a p-value of 0.0000, confirming substantial salary differences across education levels.

4. CONCLUSION

This research provides a comparative analysis of three machine learning models—Decision Tree Regression, Random Forest Regression, and Support Vector Regression (SVR)—to predict IT sector salaries based on demographic and professional attributes. The findings indicate that Random Forest Regression achieves the highest accuracy, with an R^2 Score of 91.49% and the lowest RMSE of 0.0578, demonstrating its ability to generalize well across diverse data points. Decision Tree Regression and SVR also performed well, with R^2 Scores of 89.20% and 85.12%, respectively, though their predictive performance was slightly lower than that of Random Forest Regression.

The study highlights the practical application of machine learning in salary prediction, aiding organizations in formulating equitable compensation policies and assisting individuals in career planning. Moreover, the findings serve as a baseline for similar predictive tasks in other industries, contributing to advancements in data-driven decision-making. However, the research has certain limitations. The dataset, while comprehensive, was limited to a specific time frame and geographic scope, which may affect the generalizability of the results. Additionally, some relevant features, such as industry trends, specific skill sets, or company size, were not included, and incorporating these factors could enhance prediction accuracy.

For future research, several improvements can be considered to enhance the model's performance and applicability. First, feature expansion should include additional variables like geographic location, skill demand trends, and company-specific attributes to provide a more comprehensive analysis. Second, hyperparameter tuning should be optimized to improve model accuracy and robustness. Expanding the dataset to encompass larger and more diverse samples would enhance the model's generalizability across different regions and time periods. Finally, exploring advanced or hybrid machine learning algorithms could potentially yield even better results, surpassing the performance of the current models.

Research contribution of this study provides valuable insights into the comparative performance of machine learning models for salary prediction, bridging the gap between accuracy, interpretability, and computational efficiency. The findings can assist businesses in refining salary structures, support HR professionals in making informed compensation decisions, and serve as a foundation for future advancements in salary estimation methodologies. By addressing the identified limitations and exploring new avenues, future studies can further refine and expand the applications of this approach.

Kurnia Ramadhan Putra: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Number 1, March 2025, Page 39-49
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v6i1.18

5. REFERENCES

[1] O. Dilip Dsouza *et al.*, "Salary Estimator using Machine Learning," *Int. J. All Res. Educ. Sci. Methods*, vol. 12, no. 1, pp. 2455–6211, 2024, [Online]. Available: https://www.researchgate.net/publication/377776848

[2] D. M. Lothe, P. Tiwari, N. Patil, S. Patil, and V. Patil, "Salary Prediction Using Machine Learning," *Int. J. Adv. Sci. Res.*, vol. 6, no. 5, p. 199, 2021.

[3] Y. GÖRMEZ, H. ARSLAN, S. SARI, and M. DANIŞ, "SALDA-ML: Machine Learning Based System Design to Predict Salary In-crease," *Adv. Artif. Intell. Res.*, vol. 2, no. 1, pp. 15–19, 2022, doi: 10.54569/aair.1029836.

[4] D. Nyoman, M. Cahyani, N. Putu, and K. Indah, "Comparison Of Decision Tree, Linear Regression, and Random Forest Regressor Models for Predicting House Prices," vol. 12, no. 1, pp. 62–71, 2024.

[5] S. Wijaya and F. Fauziah, "Analysis of the Comparison Between Linear Regression, Random Forest, and Logistic Regression Methods in Predicting Crude Palm Oil (CPO) Price," *Brill. Res. Artif. Intell.*, vol. 3, no. 2, pp. 343–350, 2023, doi: 10.47709/brilliance.v3i2.3334.

[6] F. Özen, "Random forest regression for prediction of Covid-19 daily cases and deaths in Turkey," *Heliyon*, vol. 10, no. 4, pp. 1–19, 2024, doi: 10.1016/j.heliyon.2024.e25746.

[7] D. Doz, M. Cotič, and D. Felda, "Random Forest Regression in Predicting Students' Achievements and Fuzzy Grades," *Mathematics*, vol. 11, no. 19, 2023, doi: 10.3390/math11194129.

[8] M. Mao, "A Comparative Study of Random Forest Regression for Predicting House Prices Using," *Highlights Sci. Eng. Technol.*, vol. 85, pp. 969–974, 2024, doi: 10.54097/bdfe8032.

[9] Wulan Septya Zulmawati, Nonong Amalita, Syafriandi Syafriandi, and Admi Salma, "Evaluation of Support Vector Regression Methods inPredictions Bitcoin's Close Price," *UNP J. Stat. Data Sci.*, vol. 1, no. 5, pp. 488–495, 2023.

[10] T. Yu and H. Zhu, "Hyper-Parameter Optimization: A Review of Algorithms and Applications," pp. 1–56, 2020, [Online]. Available: http://arxiv.org/abs/2003.05689

[11] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.

[12] S. I. Ayua, Y. M. Malgwi, and J. Afrifa, "Salary Prediction Model for Non-Academic Staff Using Polynomial Regression Technique," *Artif. Intell. Appl.*, no. 2021, pp. 1–11, 2023, doi: 10.47852/bonviewaia3202795.

[13] C. Magazzino, M. Mele, and M. Mutascu, "An artificial neural network experiment on the prediction of the unemployment rate," *J. Policy Model.*, no. xxxx, pp. 1–21, 2025, doi: 10.1016/j.jpolmod.2024.10.004.

[14] Abdullah-All-Tanvir, I. Ali Khandokar, A. K. M. Muzahidul Islam, S. Islam, and S. Shatabda, "A gradient boosting classifier for purchase intention prediction of online shoppers," *Heliyon*, vol. 9, no. 4, p. e15163, 2023, doi: 10.1016/j.heliyon.2023.e15163.

[15] H. Aminu, B. Imam Yau, F. Umar Zambuk, E. Ramsom Nanin, A. Abdullahi, and I. Zahraddeen Yakubu, "Salary Prediction Model using Principal Component Analysis and Deep Neural Network Algorithm," *Int. J. Innov. Sci. Res. Technol.*, vol. 8, no. 12, pp. 1–11, 2023, [Online]. Available: www.ijisrt.com

[16] F. Zinzendoff Okwonu, B. Laro Asaju, and F. Irimisose Arunaye, "Breakdown Analysis of Pearson Correlation Coefficient and Robust Correlation Methods," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 917, no. 1, 2020, doi: 10.1088/1757-899X/917/1/012065.

[17] J. Kotary, V. Di Vito, J. Cristopher, P. Van Hentenryck, and F. Fioretto, "Learning Joint Models of Prediction and Optimization," vol. 2, no. d, 2024, doi: 10.3233/FAIA240775.

[18] E. D. Wahyuni, A. A. Arifiyanti, and M. Kustyani, "Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining," *Pros. Nas. Rekayasa Teknol. Ind. dan Inf. XIV Tahun 2019*, vol. 2019, no. November, pp. 263–269, 2019, [Online]. Available: http://journal.itny.ac.id/index.php/ReTII

[19] M.- Mambang, "Exploratory Data Analysis of Exact Science and Social Science Learning Content on Digital Platform," *Walisongo J. Inf. Technol.*, vol. 4, no. 2, pp. 87–94, 2022, doi: 10.21580/wjit.2022.4.2.12676.

[20] A. S. Rao, B. V. Vardhan, and H. Shaik, "Role of Exploratory Data Analysis in Data Science," *Proc. 6th Int. Conf. Commun. Electron. Syst. ICCES 2021*, no. August, pp. 1457–1461, 2021, doi: 10.1109/ICCES51350.2021.9488986.

[21] I. Muhamad Malik Matin, "Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware," *Multinetics*, vol. 9, no. 1, pp. 43–50, 2023, doi: 10.32722/multinetics.v9i1.5578.

Kurnia Ramadhan Putra: *Corresponding Author