P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v7i3.369 Vol. 7, No. 3. Juni 2025
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

FORECASTING HEALTH INSURANCE PAYER INCOME: A COMPARATIVE ANALYSIS OF DECISION TREE AND SVR ALGORITHMS

Wilsen Grivin Mokodaser[-1], Tonny Irianto Soewignyo[-2], Fanny Soewignyo[-3], George Morris William Tangka[-4]

[1]Informatika / Fakultas Ilmu Komputer Universitas Klabat
wilsenm@unklab.ac.id

[2, 3]Akuntansi / Fakultas Ekonomi dan Bisnis Universitas Klabat
tonnysoewignyo@unklab.ac.id, f.soewignyo@unklab.ac.id

[4]Sistem Informasi / Fakultas Ilmu Komputer Universitas Klabat
gtangka@unklab.ac.id

Abstract

An insurance company is a type of non-bank financial institution that protects clients from risks and collects premiums over a certain period, these facts provide an overview of the insurance business and highlight its role in the economy, this study evaluated the performance difference between the Decision Tree Regressor and Support Vector Regression (SVR) in predicting insurance payer income. The Decision Tree model demonstrated strong predictive accuracy, achieving a Mean Absolute Error (MAE) of approximately 57 million and an R-squared (R^2) value of 0.896, meaning it could explain around 89.6% of the variance in the data. Additionally, the model maintained high consistency, as evidenced by 5-fold cross-validation scores ranging from 0.908 to 0.967, indicating strong generalization and low risk of overfitting. In contrast, the SVR model significantly underperformed. It recorded a much higher MAE of over 237 million and a large Mean Squared Error (MSE), reflecting substantial deviations from the actual values. Its R^2 score of -0.299 suggests that SVR performed worse than a naive mean predictor, failing to identify meaningful patterns. This poor performance was consistent across all cross-validation folds, which also produced negative R^2 scores. The SVR model's inadequacy is likely due to the large scale of the income data and the lack of proper preprocessing, such as normalization, or parameter tuning. Overall, these findings clearly demonstrate that the Decision Tree Regressor is a more suitable, accurate, and stable model for predicting insurance payer income.

Keywords: Insurance; Forecasting; Decision Tree; SVR; Predicting;

Abstrak

Perusahaan asuransi merupakan jenis lembaga keuangan non-bank yang memberikan perlindungan kepada nasabah terhadap risiko serta menghimpun premi dalam jangka waktu tertentu. Fakta ini memberikan gambaran umum mengenai bisnis asuransi dan menyoroti perannya dalam perekonomian. Studi ini mengevaluasi perbedaan kinerja antara model Decision Tree Regressor dan Support Vector Regression (SVR) dalam memprediksi pendapatan pembayar asuransi. Model Decision Tree menunjukkan akurasi prediksi yang tinggi, dengan nilai Mean Absolute Error (MAE) sekitar 57 juta dan nilai R-squared (R^2) sebesar 0,896, yang berarti model ini mampu menjelaskan sekitar 89,6% variasi dalam data. Selain itu, model ini menunjukkan konsistensi yang kuat, dibuktikan dengan skor validasi silang 5-fold yang berkisar antara 0,908 hingga 0,967, mengindikasikan kemampuan generalisasi yang baik dan risiko overfitting yang rendah. Sebaliknya, model SVR menunjukkan kinerja yang sangat buruk dengan MAE lebih dari 237 juta dan Mean Squared Error (MSE) yang besar, mencerminkan perbedaan yang signifikan antara prediksi dan nilai aktual. Nilai R^2 sebesar -0,299 menunjukkan bahwa SVR berkinerja lebih buruk dibandingkan prediktor rata-rata sederhana, dan gagal mengenali pola yang bermakna. Kinerja buruk ini konsisten di semua lipatan validasi silang dengan nilai R^2 yang juga negatif. Ketidakefektifan model SVR kemungkinan disebabkan oleh skala data pendapatan yang besar serta kurangnya prapemrosesan data seperti normalisasi atau penyesuaian parameter. Secara keseluruhan, temuan ini menunjukkan bahwa Decision Tree Regressor merupakan model yang lebih tepat, akurat, dan stabil untuk memprediksi pendapatan pembayar asuransi.

Kata kunci: Asuransi; Peramalan; Decission Tree; SVR; Prediksi;

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i3.369
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

INTRODUCTION

An insurance company is a type of non-bank financial institution that protects clients from risks and collects premiums over a certain period, in accordance with the terms of the policy. Insurance companies require substantial funds to cover all potential risks (Wahyuningsih et al., 2022). Risk is defined as uncertainty that can lead to adverse outcomes (Ardi et al., 2022). Therefore, insurance companies, like other businesses, must be managed professionally and efficiently in order to remain profitable and attract investors. Income is measured based on an organization's ability to manage its overall operations—particularly in health insurance companies—as it reflects how well the organization controls risk (Santika et al., 2023). One way to assess a company's performance is by looking at the profits it generates. If profits continue to increase over time, it indicates that management is effectively handling finances and collaboration, resulting in greater value. Insurance companies have long played a role in the national economy, so the public generally trusts the services they provide. Public awareness of the importance of insurance is growing due to uncertainties related to health, education, property, and death. People use insurance as an essential tool to anticipate future dangers or losses. According to data from the Financial Services Authority in 2016, there are 24 Islamic life insurance companies, 28 Islamic general insurance companies, and 3 Islamic reinsurance companies in Indonesia. The Islamic life insurance companies consist of 19 Islamic life insurance business units and 5 fully Islamic life insurance companies. The Islamic general insurance sector consists of 25 Islamic general insurance business units and 3 fully Islamic general insurance companies, although this number does not yet include some health coverage plans (Ardi et al., 2022). These facts provide an overview of the insurance business and highlight its role in the economy.

To support collaboration with insurance companies, stakeholders require a clear understanding of the potential risks and possibilities that may arise from such partnerships. Therefore, an insurance income prediction model is a crucial tool for companies to prepare for future collaborations with insurers. Several methods can be used to predict income, and previous implementations have explored various prediction cases. For example, one study developed a model to forecast the closing price of cryptocurrency using the Support Vector Regression (SVR) algorithm. Two experiments were conducted: one using historical Binance data and the other combining it with sentiment datasets. The results showed that predictions using both historical Binance and sentiment data yielded a lower Mean Squared Error (MSE) of 0.000830, compared to 0.00340 when using only historical data. This suggests that sentiment analysis can enhance prediction accuracy (Aruan et al., 2023). Another study found that the Simple Linear Regression (SLR) method was more suitable for predicting the population in Southeast Sulawesi. SLR achieved an average MAPE of 1.89% and RMSE of 0.51%. In most city/regency models, SLR outperformed SVR with lower MAPE and RMSE values, though in some cases, SVR performed comparably or even better (Chaidir et al., 2024). Further research employed Support Vector Regression and Polynomial Regression to predict the closing stock prices of PT Telekomunikasi Indonesia using five years of historical data. The study aimed to determine which algorithm performed best. SVR achieved an RMSE of 72.565 and MAPE of 1.486%, while Polynomial Regression (order 4) had an RMSE of 63.914 and MAPE of 1.273%, making Polynomial Regression the recommended model for this task (Putri et al., 2025). Another study predicted furniture product sales using SVR and GridSearch optimization based on 30 months of sales data for eleven products (January 2021–June 2023). The models were evaluated using MAPE. SVR without optimization yielded a MAPE of 40.39%, whereas SVR with GridSearch achieved a significantly lower MAPE of 0.45%, indicating a substantial improvement in prediction accuracy (Baidowi et al., 2024). In another case, the daily closing price movement of Solana from April 10, 2020, to May 30, 2022, was analyzed. The SVR model achieved 97.44% accuracy and a MAPE of 9.93. For a Linear kernel with parameters C = 1000 and gamma = 0.1, the Radial Basis Function (RBF) model achieved an accuracy of 87.76% and a MAPE of 8.14, suggesting the model was highly accurate overall (Atmaja & Hakim, 2022). Lastly, a study aimed to build a prediction model using Linear Regression and SVR on data from TPI Desa Ciparagejaya, which included 33 types of fish caught in 2021. The models were evaluated using RMSE, and testing was conducted using both Microsoft Excel and Python. The smallest RMSE value from Excel was 0.577735, while the smallest value from Python was 0, indicating strong predictive performance (Mahendra et al., 2024).

Another method that can be used for prediction is the Decision Tree algorithm. Previous studies have demonstrated its effectiveness, including research that provided insights useful for agricultural stakeholders in decision-making by

employing the Decision Tree algorithm. The study also highlighted how climate and weather factors influence the production yields of food crops in North Sulawesi province. However, the research did not take into account other variables that may affect crop yields, such as soil conditions or market prices, due to the rapidly changing nature of prices. Nonetheless, the study is expected to contribute positively to the agricultural sector, especially for food crop farmers, in helping to meet the basic needs of society (Joanda Kaunang et al., 2018). Another related study compared the performance of Random Forest (RF) and Decision Tree (DT) algorithms in predicting the occurrence of PPP (possibly referring to postpartum problems or pregnancy-related complications). The aim was to improve the classification performance of both algorithms. The univariate analysis revealed, for example, that 20.4% of 102 mothers had more than four children, 62.4% of 310 mothers had pregnancies spaced less than two years apart, 24.8% experienced postpartum anemia, and 12.8% delivered macrosomic babies. Additionally, 45.8% of 229 mothers experienced labor complications, 3.2% had multiple pregnancies, and some mothers were classified as high-risk due to their age. In terms of accuracy, the Random Forest algorithm outperformed the Decision Tree, achieving an accuracy of 0.830 with an AUC of 0.74 (Sinambela et al., 2023).

By examining the predictions made by the two previously discussed algorithms, this study aims to conduct a comparative analysis of prediction results using health insurance income data. The analysis will focus on two different algorithms used for prediction: the Decision Tree algorithm and the Support Vector Regression (SVR) algorithm. The goal is to evaluate and compare the performance of both models in accurately forecasting insurance income, thereby identifying which algorithm offers better predictive capabilities for this specific dataset.

RESEARCH METHODS

In this study, several methods will be employed as guidelines for conducting the research. The stages outlined in these methods are expected to optimize the results of each algorithm used for prediction. These steps are designed to ensure a structured and systematic approach to data analysis and model evaluation. The research workflow can be seen in Figure 1 below.

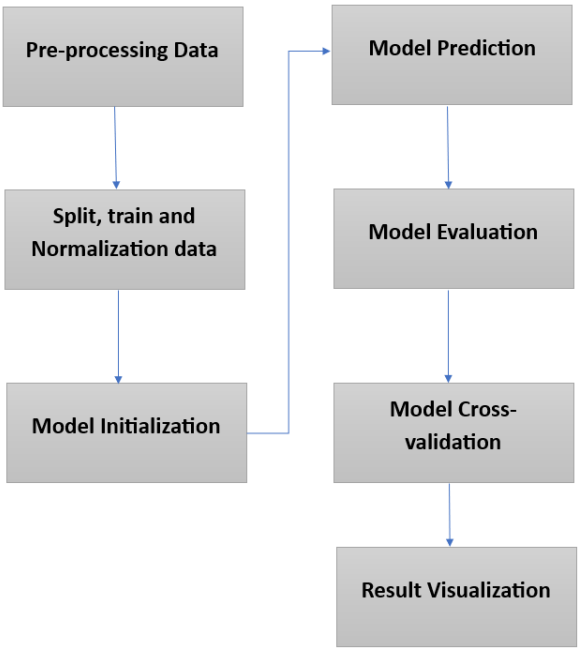

Figure 1. Research Method

Pre-Processing Data

At this stage, the dataset to be used will be loaded using Google Colab. If there are any missing or empty data entries, they will be removed. The purpose of data preprocessing in data mining is to transform the data into a format that simplifies and enhances the effectiveness of the overall process in accordance with user expectations. It also ensures that more accurate results and lower data complexity can be achieved without altering the essential information contained within the data (Saputra et al., 2020).

Split, train and data normalization

At this stage, further processing is carried out on the dataset that has already been cleaned. Using Python, the dataset—after the removal of outliers—is split into two parts: features (X), which consist of the monthly columns, and the target (y), which is the "Total" column to be predicted. The features are then normalized using StandardScaler to ensure they are on the same scale, a crucial step particularly for algorithms like Support Vector Regression (SVR) (Murtafiah & Hajarisman, 2024). Following normalization, the data is divided into two subsets: training data (80%) and testing data (20%) using the train_test_split function with a random_state of 42 to ensure result consistency. This split is important to facilitate effective model training and testing (Fadhillah Rashidatul A'la, 2024).

Model initialization

The purpose of model initialization is to build and train two regression models—Decision Tree Regressor and Support Vector Regression (SVR)—to predict the "Total" value based on monthly data (Wardana & Juanita, 2025). Two models are prepared: the Decision Tree Regressor, which captures data patterns in a hierarchical manner, and SVR with an RBF kernel, which is well-suited for handling non-linear relationships. Both models are trained using the normalized training dataset. This step serves as the foundation for comparing the performance of the two models in accurately predicting the "Total" value.

Model Prediction

The purpose of model prediction is to generate predictions of the "Total" value on the test dataset (X_test) using the two previously trained regression models: Decision Tree Regressor and Support Vector Regression (SVR). The first line produces predictions from the Decision Tree model and stores them in the variable `dt_preds`, while the second line generates predictions from the SVR model and stores them in the variable `svr_preds`. This process aims to obtain predicted values from each model, which can then be compared with the actual values (`y_test`) to evaluate the performance of both models in predicting new, unseen data. These prediction results will subsequently be used in the evaluation phase with metrics such as MAE, MSE, or R^2.

Model Evaluation

Model evaluation aims to assess the performance of two regression models, namely Decision Tree and SVR, based on their predictions on the test data. The function `evaluate_model` is defined to accept two parameters: the actual values (`y_true`) and the predicted values (`y_pred`), and it calculates three key evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). MAE measures the average absolute error, MSE measures the average of the squared errors, while R^2 indicates how well the model explains the variance in the target data (Melati N et al., 2023; Septaraja et al., 2024). After the function is defined, it is called twice—once for the Decision Tree model predictions (`dt_preds`) and once for the SVR model predictions (`svr_preds`)—comparing the predicted results against the actual values (`y_test`). The evaluation results are stored in `dt_metrics` and `svr_metrics`, which can then be used to assess and compare the accuracy and performance of both models in predicting the data.

Model Cross-validation

Model cross-validation aims to evaluate the performance of the Decision Tree and SVR models more comprehensively using the cross-validation technique (Sulistiani & Aldino, 2020). Using the cross_val_score function, each model is tested through a 5-fold cross-validation process, where the data is split into five parts: four parts are used for training and one part for testing, in rotation. The evaluation is based on the R-squared (R^2) score (Kurniawan et al., 2022), which indicates how well the model explains the variance in the target data. This process is conducted on the entire standardized feature set (X_scaled) and the target (y). The scores from each fold are stored in dt_cv_scores for the Decision Tree model and svr_cv_scores for the SVR model. The main goal of this cross-validation is to ensure that the model's performance is not only good on a single data split but also consistent across different subsets of data, thereby reducing the risk of overfitting and providing a more reliable picture of the model's generalization capability (Fitri, 2023; Pomalingo et al., 2019).

Result visualization

At this final stage of the research, a more detailed interpretation of the prediction results from each algorithm will be conducted by presenting visualizations of the predicted data. This aims to clearly illustrate the comparison between the models that have been successfully implemented.

RESULTS AND DISCUSSION

At this stage, the data to be used consists of the total revenue from each insurance payer. The dataset contains 422 records, with insurance payer data spanning from January to August.

Payer	Januari	Februari	Maret	April	Mei	Juni	Juli	Agustus
ASURANSI EQUITY LIFE INDONESIA - ADMEDIKA	985,851	68,740,701	197,238,290	7,997,961	2,089,124	2,073,328	2,070,581	2,060,280
EQUITY LIFE INDONESIA	69,696,868	46,784,516	70,596,700	260,000	1,133,600	1,060,000	1,066,360	1,061,060
AA INTERNATIONAL HUB SDN BHD	7,430,000	2,656,000	264,940	273,000	383,950	350,000	354,900	384,650
ADMEDIKA, PT - ASURANSI JIWA SYARIAH, PT	1,903,000	3,319,008	7,608,627	445,398	1,328,549	1,273,654	1,263,422	1,229,000
ADMINISTRASI MEDIKA, PT	1,202,380	2,492,030	1,200,031	5,073,680	2,652,983	2,588,078	2,677,222	2,433,929
ASURANSI ASTRA BUANA, PT (GARDA MEDIKA)	5,383,174	72,907,960	212,477,558	863,148	100,900	100,900	100,900	100,900
ASURANSI AXA INDONESIA - ADMEDIKA	108,483,749	36,579,715	36,579,715	1,255,395	100,000	100,000	105,400	103,400
ASURANSI CIGNA, PT	689,212	1,161,904	649,039	2,147,462	4,151,148	4,329,500	4,416,090	4,649,554
ASURANSI JIWA ASTRA, PT (RAPID TEST)	2,293,342	1,964,409	2,211,886	1,390,000	1,538,848	1,538,848	1,674,267	1,592,708
AXA MALAYSIA (ASIA ASSISTANCE)	964,197	964,197	1,463,670	1,428,920	88,648,326	166,861,000	85,309,406	85,000
BINA NUSANTARA - ADMEDIKA	3,728,845	3,574,130	3,574,130	6,993,544	132,001,461	11,186,802	1,414,552,123	525,477,757
FULLERTON HEALTH INDONESIA - DONGGI SENORO	5,757,923	32,499,812	65,039,501	26,702,013	12,485,068	34,745,086	34,745,086	34,745,086
HRD SILOAM HOSPITALS LIPPO VILLAGE	47,567,975	33,329,057	50,120,220	2,098,976	40,048,003	40,048,003	29,733,000	489,000
HRD SILOAM HOSPITALS-FAMILY	2,254,300	2,122,345	2,281,587	1,830,547	106,466,562	150,432,746	106,678,226	61,230,400
INT'L SOS (PT ASIH EKA ABADI) - FREEPORT IND	34,507,897	426,888,440	896,667,772	349,489,652	12,650,000	12,500,000	13,025,000	13,025,000
INTEGRASI LAYANAN GLOBAL KSK - ADMEDIKA	872,007	24,053,447	42,968,315	28,320,018	9,161,425	5,994,000	9,407,047	11,607,200
LIPPO GENERAL INSURANCE TBK, PT – TELECONSULTATION	25,135,852	19,383,085	24,630,729	8,382,675	338,495	498,000	1,555,713	3,708,399
MANDIRI AXA GENERAL INSURANCE, PT	8,609,007	700,000	8,860,487	8,969,462	2,628,309	40,960,976	107,054,009	13,200,611
MANGGALA TIRTA KENCANA, PT	747,600	728,700	728,700	704,900	2,799,900	2,700,000	2,862,000	2,940,300
PENGHUBUNG KY WILAYAH SULAWESI UTARA	5,185,000	5,185,000	5,185,000	5,185,000	32,584,819	3,449,000	50,436,747	34,981,940
PLN (PERSERO), PT - ASURANSI PERISAI LISTRIK NASIONAL (APLN), PT	7,707,785	56,855,055	81,795,731	81,061,650	5,962,068	5,614,000	5,771,192	5,962,068

Figure 2. Data

Data Preprocessing

Using Python programming language, data cleaning will be performed with the aim of

removing outliers in the "Total" column by applying the Interquartile Range (IQR) method. First, the code calculates the first quartile (Q1) and third quartile (Q3) of the "Total" values, then computes the IQR as the difference between Q3 and Q1. This IQR is used to define the lower and upper bounds of acceptable data, which are Q1 minus 1.5 times IQR for the lower bound, and Q3 plus 1.5 times IQR for the upper bound. Any values outside this range are considered outliers. The final line filters the dataset by including only rows where the "Total" value falls within these bounds, and stores the cleaned data in the variable `df_no_outliers`. The purpose of this code is to remove extreme data points so that analysis and modeling become more accurate and not distorted by values that do not represent the general population. Before cleaning, the dataset contains 422 records, with the data distribution shown in the following figure.

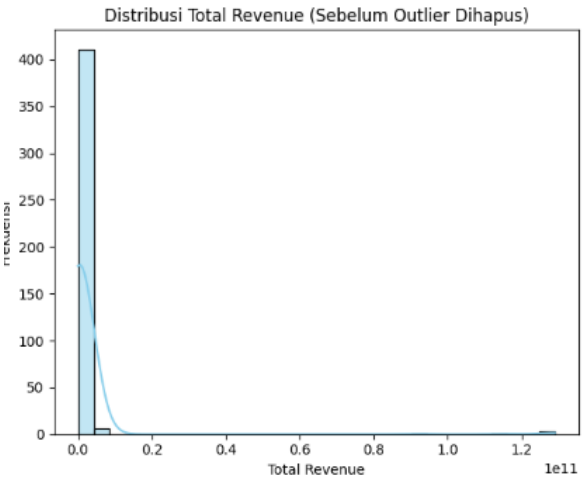

Figure 3. Data distribution before processing

The image shows a histogram of the Total Revenue distribution before the outlier removal process. From the graph, it is evident that most of the data is concentrated at very low values on the left side (close to zero), with a frequency exceeding 400 occurrences. In contrast, there are a few data points with very high values on the right side of the graph (a long right tail), indicating the presence of outliers or extreme values much larger than the majority of the data. This distribution is right-skewed, meaning most values are small but there are some extremely large values. Such a condition can affect the performance of predictive models, as outliers can distort the training results. Therefore, this visualization provides a strong rationale for performing data cleaning (outlier removal) to achieve a more balanced and representative distribution.

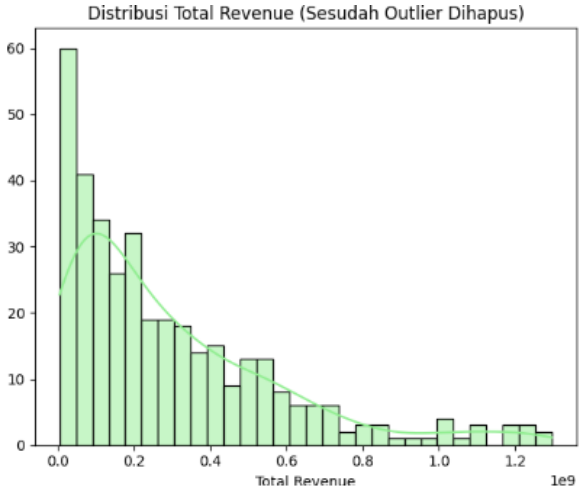

Figure 4. Data distribution after transmission

This image shows a histogram of the Total Revenue distribution after the outliers have been removed from the data. Unlike the previous graph, this distribution appears more balanced and is no longer dominated by extremely large values. Most of the data still clusters on the left side (low values), but the spread is now more even and less skewed to one side. The distribution pattern still exhibits a right-skewed characteristic, but the tail is much shorter than before, indicating that the outliers have been successfully eliminated. This demonstrates that the data cleaning process, using the Interquartile Range (IQR) method, was effective in producing a more representative distribution for statistical analysis and machine learning modeling. With such a distribution, the resulting model is expected to be more accurate and less prone to distortion from extreme data points.

After completing the data cleaning process, the next step is to proceed with data preparation.

```
X = df_no_outliers[month_cols]
y = df_no_outliers['Total']

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

Figure 5. Data preparation code

The result of the code is a well-prepared dataset ready to be used for training and testing machine learning models. First, the data that has been cleaned of outliers is separated into two main parts: X, which contains the features or input variables consisting of monthly data (month_cols), and y, the target variable to be predicted, which is Total. Next, the features in X are normalized using StandardScaler so that all features have a uniform scale with a mean of 0 and a standard deviation of 1. This normalization is important to ensure that

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i3.369
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

models, especially those sensitive to scale such as Support Vector Regression (SVR), can perform optimally. After normalization, the data is split into two subsets using train_test_split: training data (X_train, y_train) comprising 80% of the data, and testing data (X_test, y_test) comprising 20%. This split allows the model to be trained on the majority of the data and tested on unseen data, helping to objectively measure the model's generalization ability.

```
dt_model = DecisionTreeRegressor(random_state=42)
svr_model = SVR(kernel='rbf')

dt_model.fit(X_train, y_train)
svr_model.fit(X_train, y_train)
```

```
▾ SVR
SVR()
```

Figure 6. Model Training

The result of the code is two regression models that have been trained using the training data (X_train and y_train) to predict the Total value. The first model is the Decision Tree Regressor, which works by building a decision tree structure based on splitting the data into homogeneous groups. This model can capture nonlinear relationships and interactions between features with an intuitive approach. The second model is Support Vector Regression (SVR) with an RBF (Radial Basis Function) kernel, which is suitable for handling complex nonlinear relationships in the data. SVR operates by finding the best function that minimizes deviation from the actual values while keeping the margin of error within a certain limit. Both models have learned patterns from the training data and are ready to make predictions on the test data, allowing their performance to be compared in modeling and predicting total revenue based on monthly data.

After the model creation is completed, the next step is to proceed with the implementation of the models.

```
dt_preds = dt_model.predict(X_test)
svr_preds = svr_model.predict(X_test)
```

Figure 7. Prediction implementation code

This code is used to generate predictions of the Total value using two previously trained regression models: Decision Tree Regressor and Support Vector Regression (SVR). By applying these models to the test dataset (X_test), which was not used during training, each model processes the input features and produces its respective predictions. The predicted values are stored in two variables: `dt_preds` for the Decision Tree model and `svr_preds` for the SVR model. The main objective of this step is to evaluate how accurately each model can predict the target variable (Total) based on new, unseen data. These predictions will later be compared with the actual values (y_test) to assess and compare the performance of both models in forecasting total revenue from the monthly data features.

```
def evaluate_model(y_true, y_pred):
    return {
        'MAE': mean_absolute_error(y_true, y_pred),
        'MSE': mean_squared_error(y_true, y_pred),
        'R2': r2_score(y_true, y_pred)
    }

dt_metrics = evaluate_model(y_test, dt_preds)
svr_metrics = evaluate_model(y_test, svr_preds)
```

Figure 8. Model evaluation

The results of the model evaluation reveal a clear performance gap between the Decision Tree Regressor and Support Vector Regression (SVR) in predicting insurance payer income. The Decision Tree model showed strong predictive capabilities, achieving a Mean Absolute Error (MAE) of approximately 57 million and an R-squared (R^2) score of 0.896, meaning it could explain about 89.6% of the variance in the target variable. Additionally, its 5-fold cross-validation scores, which ranged from 0.908 to 0.967, demonstrated that the model consistently maintained high accuracy with low risk of overfitting or underfitting. In contrast, the SVR model delivered poor performance, with a MAE exceeding 237 million and a notably high Mean Squared Error (MSE), indicating substantial differences between its predictions and the actual values. Its R^2 score of -0.299 suggests that the model performed worse than simply predicting the mean of the target variable for all entries. This poor performance was reinforced by negative R^2 scores across all folds in cross-validation, underscoring the SVR model's failure to generalize and its inability to detect meaningful patterns within the dataset. These findings highlight the superior suitability of the Decision Tree approach for this specific regression task.

The next step is to perform cross-validation on the model that has been developed.

```
print("\nDecision Tree Metrics:", dt_metrics)
print("Decision Tree CV Scores:", dt_cv_scores)
print("SVR Metrics:", svr_metrics)
print("SVR CV Scores:", svr_cv_scores)
```

```
Decision Tree Metrics: {'MAE': 57251460.436040714, 'MSE': 8984004911182443.0, 'R2': 0.8958706690606781}
Decision Tree CV Scores: [0.96670256 0.90816967 0.93954936 0.94954277 0.94901531]
SVR Metrics: {'MAE': 237242558.08409578, 'MSE': 1.1211460431815318e+17, 'R2': -0.29946709196973 8}
SVR CV Scores: [-0.20481476 -0.07132385 -0.1431508  -0.12612914 -0.03812359]
```

Figure 8 Cross validation result

The evaluation results indicate that the Decision Tree model performs significantly better than the SVR model in predicting the Total value. Based on the evaluation metrics, the Decision Tree achieved a Mean Absolute Error (MAE) of about 57 million, a Mean Squared Error (MSE) of approximately 8.98 quadrillion, and an R-squared (R^2) of 0.896, which means the model explains around 89.6% of the variance in the data quite well. Additionally, the Decision Tree's cross-validation scores show consistently high R^2 values across all folds, averaging above 0.9, indicating the model is stable and reliable. In contrast, the SVR model performed poorly, with a much higher MAE of around 237 million and an extremely large MSE, along with a negative R^2 of -0.299, suggesting the model performs worse than a simple mean-based predictor. The SVR's cross-validation scores are also negative and very low, indicating it struggles to capture patterns in the data and lacks stability across different subsets. In conclusion, the Decision Tree model is more effective and accurate for this prediction task compared to SVR.

After the evaluation process for each algorithm has been completed, a visualization was carried out to compare the distribution of the predicted data with the actual data.

Figure 9. Decision Tree prediction

The image shows a scatter plot comparing the actual and predicted total revenue values generated by the Decision Tree model. The dots on the graph represent pairs of actual and predicted values for each data point, while the dashed red line represents the identity line (y = x), which indicates the ideal position where predictions exactly match the actual values. Overall, most of the points are close to this line, indicating that the Decision Tree model has reasonably good predictive performance. However, there are several points that deviate significantly from the line, suggesting prediction errors in certain cases. This implies that while the model is able to capture general patterns in the data, there may be instances of overfitting or a lack of generalization in specific scenarios.

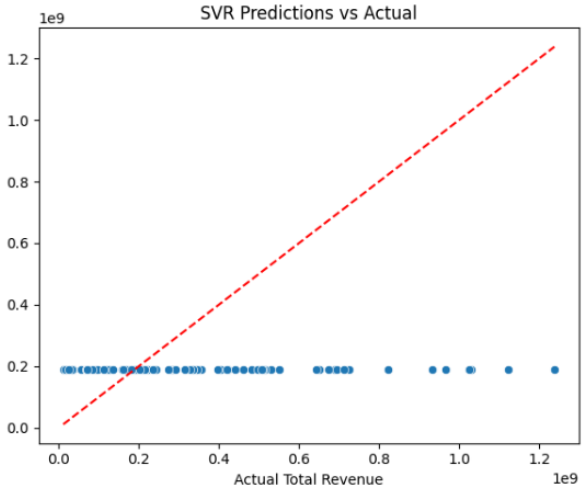

Figure 10. SVR prediction

The image shows a scatter plot comparing actual and predicted total revenue values produced by the Support Vector Regression (SVR) model. The blue dots represent pairs of actual and predicted values, while the red dashed line indicates the ideal case where predictions perfectly match the actual values (y = x). In this plot, most of the predicted values are clustered around a constant level, regardless of the variation in actual revenue. This suggests that the SVR model failed to capture the underlying patterns in the data and produced nearly uniform predictions. As a result, the points are significantly distant from the identity line, indicating poor predictive performance. This behavior reflects underfitting, where the model is too simplistic to learn the complexity of the data.

CONCLUSIONS AND SUGGESTIONS

Conclusions

This study examined the performance gap between the Decision Tree Regressor and Support Vector Regression (SVR) in predicting insurance payer income. The Decision Tree model showed strong predictive capabilities, achieving a Mean

Absolute Error (MAE) of approximately 57 million and an R-squared (R^2) score of 0.896, meaning it could explain about 89.6% of the variance in the target variable. Additionally, its 5-fold cross-validation scores, which ranged from 0.908 to 0.967, demonstrated that the model consistently maintained high accuracy with low risk of overfitting or underfitting. In contrast, the SVR model delivered poor performance, with a MAE exceeding 237 million and a notably high Mean Squared Error (MSE), indicating substantial differences between its predictions and the actual values. Its R^2 score of -0.299 suggests that the model performed worse than simply predicting the mean of the target variable for all entries. This poor performance was reinforced by negative R^2 scores across all folds in cross-validation, underscoring the SVR model's failure to generalize and its inability to detect meaningful patterns within the dataset. These findings highlight the superior suitability of the Decision Tree approach for this specific regression task.

Suggestions

The evaluation results further confirm that the Decision Tree model is more accurate and stable for this prediction task. It achieved a low Mean Absolute Error of around 57 million and a strong R-squared value of 0.896, with cross-validation scores consistently above 0.9—indicating a reliable and generalizable model. Meanwhile, the SVR model performed poorly, with an MAE of about 237 million, a very large MSE, and a negative R^2 value of -0.299, signifying performance worse than a basic mean predictor. Its negative cross-validation scores further reflect instability and poor pattern recognition across data folds.

The suboptimal performance of SVR is likely due to the large scale of the income data and insufficient parameter tuning and preprocessing, such as data normalization. Therefore, for predicting insurance payer income, the Decision Tree is recommended as a more accurate and stable model.

For future research, it is advised to perform thorough hyperparameter tuning and comprehensive data preprocessing for SVR, as well as explore other algorithms that are better suited to similar data characteristics. Additionally, testing on larger and more diverse datasets could improve the generalizability of the results.

REFERENCES

Ardi, A. R. S., Batubara, M., & Harahap, M. I. (2022). Pengaruh Pendapatan Premi, Hasil Investasi dan Klaim Terhadap Laba Pada PT Asuransi Multi Artha Guna Tbk (AMAG). *Jurnal Ekonomi Syariah Dan Bisnis*, *5*(2), 179–192.

Aruan, N. M., Simanjuntak, G. W., & Siagian, A. I. (2023). Pendekatan Algoritma Support Vector Regression Dalam Memprediksi Harga Cryptocurrency (Studi Kasus: Binance). *Jurnal Teknik Informatika Dan Sistem Informasi*, *10*(3). http://jurnal.mdp.ac.id

Atmaja, D. M. U., & Hakim, A. R. (2022). Peramalan Harga Mata Uang Kripto Solana Menggunakan Metode Support Vector Regression (Svr). *Jurnal Media Elektro*, *XI*(2), 97–104. https://doi.org/10.35508/jme.v0i0.8117

Baidowi, A., Fitra, E., As, A. H., Tholib, A., & Guterres, J. X. (2024). Implementasi GridSearch dalam Meningkatkan Kinerja Model Support Vector Regression (SVR) utuk Prediksi Penjualan Produk pada Meuble Rohman Jaya Implementation of GridSearch to Improve the Performance of the Support Vector Regression (SVR) Model for Pr. *Keilmuan Dan Aplikasi Teknik Informatika*, *3489*, 22–30.

Chaidir, R. I. M., Ramadhan, A. F., Zaria, H., & Saputra, R. A. (2024). Perbandingan Algoritma Simple Linear Regression Dan Support Vector Regression Dalam Prediksi Jumlah Penduduk Di Sulawesi Tenggara. *METHODIKA: Jurnal Teknik Informatika Dan Sistem Informasi*, *10*(1), 27–31. https://doi.org/10.46880/mtk.v10i1.2548

E, S. V., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, *153*(February), 353–366. https://doi.org/10.1016/j.comcom.2020.02.007

Fadhillah Rashidatul A'la, Z. F. (2024). Perbandingan Algoritma Decision Tree dan Deep Learning dalam Prediksi Masalah Kesehatan berdasarkan Kebiasaan Gaya Hidup Fadhillah Rashidatul A ' la , Zaehol Fatah Universitas Ibrahimy , Indonesia lifestyle habits ; health prediction ; decision tree ; dee. *Mutiara: Multidiciplinary Scientifict Journal*, *2*(10).

Fitri, E. (2023). Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah. *Journal of Applied Computer Science and Technology*, *4*(1), 58–64. https://doi.org/10.52158/jacost.v4i1.491

Joanda Kaunang, F., Rotikan, R., & Stella Tulung, G.

(2018). Pemodelan Sistem Prediksi Tanaman Pangan Menggunakan Algoritma Decision Tree Crop Prediction System Using Decision Tree Algorithm. *Cogito Smart Journal*, *4*(1), 213–218.

Kurniawan, Y., Winoto Tj, H., & Fushen, F. (2022). Pengaruh Kualitas Layanan Dan Penanganan Keluhan Terhadap Loyalitas Pasien BPJS Dimediasi Oleh Kepuasan Pelanggan (Studi Pada Pasien Pengguna BPJS Kesehatan Di RSIA Bunda Sejahtera). *Jurnal Manajemen Dan Administrasi Rumah Sakit Indonesia (MARSI)*, *6*(1), 74–85. https://doi.org/10.52643/marsi.v6i1.1939

Mahendra, F., s Siregar, A., & Baihaqi, K. (2024). Implementasi Algoritma Regresi Linear Dan Support Vector Regression Dalam Membuat Model Prediksi Hasil Tangkapan Ikan Nelayan Desa Ciparagejaya. *Scientific Student Journal for Information, Technology and Science*, *5*(1), 9–17.

Murtafiah, N. S., & Hajarisman, N. (2024). Deteksi Anomali Aktivitas Kegempaan Gunung Marapi Menggunakan Algoritma Local Outlier Factor. *Bandung Conference Series: Statistics*, *November 2023*, 526–537.

Melati N, R., Waluyo Purboyo, T., & Kalista, M. (2023). Prediksi Penderita Tuberkulosis Menggunakan Algoritma Support Vector Regression (SVR). *E-Proceeding of Engineering*, *10*(1), 736–741.

Pomalingo, S., Sugiantoro, B., & Prayudi, Y. (2019). Data Visualisasi Sebagai Pendukung Investigasi Media Sosial. *ILKOM Jurnal Ilmiah*, *11*(2), 143–151. https://doi.org/10.33096/ilkom.v11i2.443.143-151

Putri, W. E., Buana, U., Karawang, P., Faisal, S., Buana, U., Karawang, P., Rohana, T., Buana, U., & Karawang, P. (2025). Implementasi Algoritma Support Vector Regression dan Polynomial Regression dalam Memprediksi Harga Saham PT Telekomunikasi Indonesia. *Scientific Student Journal for Information, Technology and Science*, *6*, 70–76.

Santika, A., Rahayu, Y., Ernawati, N., & Zuhriatusobah HS, J. (2023). Pengaruh Net Profit Margin, Earning Per Share, Inflasi dan Nilai Tukar Rupiah Terhadap Harga Saham. *Owner: Riset & Jurnal Akuntansi*, *7*(1), 753–763. https://doi.org/10.33395/owner.v7i1.1269

Saputra, R. A., Agustina, C., Puspitasari, D., Ramanda, R., Warjiyono, Pribadi, D., Lisnawanty, & Indriani, K. (2020). Detecting Alzheimer's Disease by the Decision Tree Methods Based on Particle Swarm Optimization. *Journal of Physics: Conference Series*, *1641*(1), 61–67. https://doi.org/10.1088/1742-6596/1641/1/012025

Septaraja, A. F., Joannes, K., Radhi, M. R., & Parhusip, J. (2024). Implementasi Algoritma Decision Tree Untuk Prediksi Efisiensi Biaya Bensin Kendaraan Bermotor Parenggean Menuju Palangkaraya Implementation Of The Decision Tree Algorithm For Predicting The Cost Efficiency Of Gasoline For Motor Vehicles Parenggean Traveli. *Informatech: Jurnal Ilmiah Informatika Dan Komputer*, *1*, 243–246.

Sinambela, D. P., Naparin, H., Zulfadhilah, M., & Hidayah, N. (2023). Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin. *Jurnal Informasi Dan Teknologi*, *5*(3), 58–64. https://doi.org/10.60083/jidt.v5i3.393

Sulistiani, H., & Aldino, A. A. (2020). Decision Tree C4.5 Algorithm for Tuition Aid Grant Program Classification (Case Study: Department of Information System, Universitas Teknokrat Indonesia). *Edutic - Scientific Journal of Informatics Education*, *7*(1), 40–50. https://doi.org/10.21107/edutic.v7i1.8849

Wahyuningsih, S., Ediwijoyo, S. P., Ganesha, P. P., & Tengah, J. (2022). Jurnal E-Bis : Ekonomi Bisnis Kajian Prediksi Kebangkrutan Industri Asuransi Di Indonesia Tahun 2019-. *E-Bisnis: Ekonomi Bisnis*, *6*(2), 555–570.

Wardana, F. A., & Juanita, S. (2025). Prediksi jumlah tenaga kerja asing di jawa barat menggunakan perbandingan algoritma support vector regression dan decision tree regression. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *10*(2), 890–899.