Co-ELITe: Journal of Computer Engineering, Electronics and Information Technology
Vol. 1, No. 1, April 2021, pp. xx~xx
ISSN: xxxx-xxxx,

Predicting Laptop Prices with a Linear Regression Model

Rony Wahyu Utama[1], Riyandi Firman Pratama[2],Muhamad Faisal Fiqri[3]
[1,2,3]Universitas Pendidikan Indonesia, Bandung, 40291, Indonesia

ARTICLE INFO

Article history's:

Received
Revised
Accepted

Keywords: predict, laptop price, linear regression.

ABSTRACT

Laptop price prediction is an important activity to help consumers make purchasing decisions. Linear regression is a method that can be used to predict laptop prices using features as input. This study aims to predict laptop prices using linear regression models and evaluate the ability of the model to predict laptop prices. In addition, this study also aims to determine which features have the most influence on laptop prices and how much influence each feature has on laptop prices. The data used is laptop data containing information about laptop specifications such as screen size, RAM, hard drive, processor (CPU), and other parameters as input features and laptop prices as targets. The results showed that the linear regression model can predict laptop prices well with an MAE value of 0.1, an MSE value of 0.03, and a coefficient of determination (r^2) value of 0.92 and the features that have the most influence on laptop prices are CPU and GPU. This research is expected to provide benefits for consumers in making purchasing decisions and for producers in determining the right pricing strategy.

Rony Wahyu Utama
Universitas Pendidikan Indonesia, Bandung, 40291, Indonesia
Email: ronywahyuu@upi.edu

1. INTRODUCTION

Predicting laptop prices is an essential task that helps consumers make informed purchasing decisions [1] One method that can be used to predict laptop prices is linear regression [2,3], a technique that uses a straight line equation to predict the value of a dependent variable based on the value of one or more independent variables [4]. In this study, we will use a linear regression model to predict the price of a laptop based on its specifications, such as screen size, RAM, hard drive, and processor. We will evaluate the model's ability to accurately predict laptop prices using appropriate metrics.

Linear regression models can also be used to identify which features have the greatest influence on laptop prices [1]. This information can be useful for manufacturers in determining the right pricing strategy for their products [2]. In addition, linear regression models can provide insight into the features that are most important to consumers when choosing a laptop [3,4]. Overall, this research is expected to benefit both consumers, by helping them make informed purchasing decisions, and manufacturers, by providing information on pricing strategies and important features.

2. RESEARCH METHOD

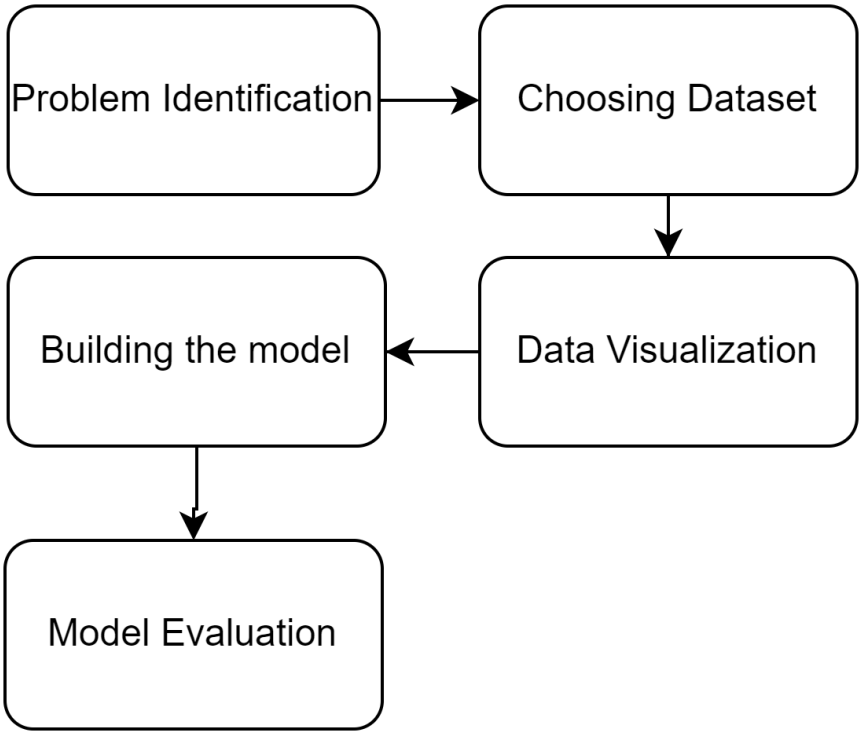

Fig 1. Research Method

A. Problem Identification

This research departs from the problem of consumers who will buy a laptop but do not know the exact price. This research discusses how to predict laptop prices based on available variables such as brand, amount of RAM, CPU type etc.

B. Dataset Selection

The dataset used in this research comes from the kaggle.com site. The dataset contains a collection of Laptop data that has several attributes such as company (manufacturer), Product (product name), TypeName (laptop type), Inches (dimensions), ScreenResolution (screen resolution), Cpu (CPU used), Ram (amount of RAM), Memory (amount of storage), Gpu (GPU used), OpSys (Operating System used), Weight (laptop weight), and Price_euros (price in euros).

This dataset contains as much as 1302 data and most of the data is categorical data so that it needs an encoding process to convert categorical data into numeric data so that it can be processed by the algorithm.

	laptop_ID	Company	Product	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price_euros
0	1	Apple	MacBook Pro	Ultrabook	13.300	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	1339.690
1	2	Apple	Macbook Air	Ultrabook	13.300	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	898.940
2	3	HP	250 G6	Notebook	15.600	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	575.000
3	4	Apple	MacBook Pro	Ultrabook	15.400	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	2537.450
4	5	Apple	MacBook Pro	Ultrabook	13.300	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37kg	1803.600
5	6	Acer	Aspire 3	Notebook	15.600	1366x768	AMD A9-Series 9420 3GHz	4GB	500GB HDD	AMD Radeon R5	Windows 10	2.1kg	400.000
6	7	Apple	MacBook Pro	Ultrabook	15.400	IPS Panel Retina Display 2880x1800	Intel Core i7 2.2GHz	16GB	256GB Flash Storage	Intel Iris Pro Graphics	Mac OS X	2.04kg	2139.970
7	8	Apple	Macbook Air	Ultrabook	13.300	1440x900	Intel Core i5 1.8GHz	8GB	256GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	1158.700
8	9	Asus	ZenBook UX430UN	Ultrabook	14.000	Full HD 1920x1080	Intel Core i7 8550U 1.8GHz	16GB	512GB SSD	Nvidia GeForce MX150	Windows 10	1.3kg	1495.000
9	10	Acer	Swift 3	Ultrabook	14.000	IPS Panel Full HD 1920x1080	Intel Core i5 8250U 1.6GHz	8GB	256GB SSD	Intel UHD Graphics 620	Windows 10	1.6kg	770.000

Fig 2. Dataset Tables

C. Data Visualization

Data visualization is the process of visually displaying data, such as graphs, plots, or diagrams, with the aim of making it easier to understand the data. Data visualization can help us find patterns, explore data, and understand data better. There are various forms of data visualization, such as line plot, scatter plot, bar chart, histogram, and so on. Choosing the right form of visualization depends on the type of data to be displayed and the purpose of the visualization.

In recent years, data visualization has become increasingly popular and widely used in various fields, such as business, science, and journalism. Using the right data visualization can help in understanding and conveying information better.

D. Correlation test, Normalization test, Hypothesis test, and T test

1. Correlation test

A correlation test is a statistical procedure used to evaluate the strength and direction of the relationship between two variables [5, 6]. It is used to determine whether there is a significant association between the two variables and to quantify the strength of that association. There are several types of correlation tests, including Pearson's correlation coefficient, Spearman's rank correlation coefficient, and Kendall's tau [7]. The type of correlation test used depends on the characteristics of the data and the specific research question.

The Pearson's correlation coefficient is a measure of the linear association between two continuous variables. It is calculated by dividing the covariance of the two variables by the product of their standard deviations. The Pearson's correlation coefficient ranges from -1 to 1, with values close to -1 indicating a strong negative relationship, values close to 1 indicating a strong positive relationship, and values close to 0 indicating no relationship.

2. Normalization test

Normalization is a statistical process that is used to scale variables to a common range or to a standard normal distribution. It is often used to standardize variables so that they can be compared or combined in statistical analyses [8,9,10]. There are several types of normalization techniques, including min-max normalization, z-score normalization, and decimal scaling. The type of normalization used depends on the characteristics of the data and the specific research question.

Min-max normalization scales the data to a fixed range, typically between 0 and 1. It is calculated by subtracting the minimum value from each data point and dividing the result by the range (maximum value minus minimum value). Z-score normalization standardizes the data by subtracting the mean and dividing by the standard deviation. It is calculated by subtracting the mean from each data point and dividing the result by the standard deviation. Decimal scaling normalizes the data by multiplying each value by a power of 10 and rounding to the nearest integer. It is often used to scale data with large range differences or to reduce the impact of noise on the data.

3. Hypothesis test

A hypothesis test is a statistical procedure used to evaluate the evidence in a dataset and determine whether it supports a specific hypothesis or not (Gao et al., 2017; Huang et al., 2021; Wang et al., 2019). It involves formulating a null hypothesis (a statistical statement that there is no relationship between the variables of interest) and an alternative hypothesis (a statistical statement that there is a relationship between the variables of interest). The null hypothesis is then tested against the alternative hypothesis using a sample of data, and the evidence in the sample is used to decide whether to reject or fail to reject the null hypothesis. There are several types of hypothesis tests, including t-tests, ANOVA, and chi-square tests. The type of hypothesis test used depends on the characteristics of the data and the specific research question.

4. T test

A t-test is a statistical procedure used to determine whether the mean of a sample is significantly different from a known population mean or the mean of another sample. It is based on the t-statistic, which is calculated by

dividing the difference between the sample mean and the population mean by the standard error of the mean. The t-statistic is then compared to a critical value from a t-distribution, which is a distribution of values that the t-statistic can take under the assumption that the null hypothesis is true. There are several types of t-tests, including the one-sample t-test, the two-sample t-test, and the paired t-test. The type of t-test used depends on the specific research question and the characteristics of the data.

E. Regression Modeling

After several steps above, the next step is to create a regression model. The regression model used is multiple linear regression. Multiple linear regression model is a model used to predict the value of a target variable (dependent variable) based on the value of several predictor variables (independent variables) or other features. This model is a variation of the simple linear regression model, where in this model there is more than one feature used to predict the target variable.

In this study, the variables that act as features are company (manufacturer), Product (product name), TypeName (laptop type), Inches (dimensions), ScreenResolution (screen resolution), Cpu (CPU used), Ram (amount of RAM), Memory (amount of storage), Gpu (GPU used), OpSys (Operating System used), Weight (laptop weight). While the variable that acts as a target is price_euros (price in Euro currency). The linear regression models used include: Lasso Regression, Kernel Ridge Regression, and Elastic Net Regression. These three models are used to compare which model is the best in making predictions.

F. Model Evaluation

Model evaluation is done to find out how well the model performs in predicting or explaining data. In this research, there are several metrics used, including:

1. Mean Squared Error (MSE)

Mean Squared Error (MSE) is one of the metrics used to evaluate the performance of machine learning models [11]. MSE is calculated by multiplying the difference between the actual value and the value produced by the model by the square of the difference, then summing for all samples, and calculating the average [12]. MSE is calculated using the following formula:

$$MSE = 1/n \sum (y\_i - \hat{y}\_i)^2$$

where:
n is the number of samples
y_i is the actual value
ŷ_i is the value generated by the model

The smaller the MSE value, the better the model is at calculating the actual value. However, keep in mind that MSE does not measure the absolute value of the difference between the actual value and the value produced by the model, but rather the squared difference.

2. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is one of the metrics used to evaluate the performance of machine learning models [11]. MAE is calculated by summing the absolute difference between the actual value and the value produced by the model for all samples, and then calculating the average[12]. The smaller the MAE value, the better the model is at calculating the actual value. MAE is more sensitive to smaller value differences than MSE, because MAE measures the absolute difference between the actual value and the value produced by the model, not the squared difference like MSE.

3. R-Squared (R2)

R Squared (R^2) is one of the metrics used to evaluate the performance of regression models[11]. R^2 is calculated by measuring how much the variation in values produced by the model explains the variation in actual values [12]. R^2 is calculated using the following formula:

$$R^2 = 1 - \sum (y\_i - \hat{y}\_i)^2 / \sum (y\_i - \bar{y})^2$$

where:
n is the number of samples
y_i is the actual value
$\hat{y}$_i is the value generated by the model
$\bar{y}$ is the average value of the actual value

The greater the R^2 value, the better the model is at explaining variations in actual values. The R^2 value ranges between 0 and 1, with a value of 1 indicating that the model can explain all variations in actual values.

3. RESULTS AND DISCUSSION

Before calculating the score and testing the model. Several processes are carried out, namely data visualization and Data Preprocessing. To find out the distribution of data on the average price of Laptops based on the company, the following graph is made:

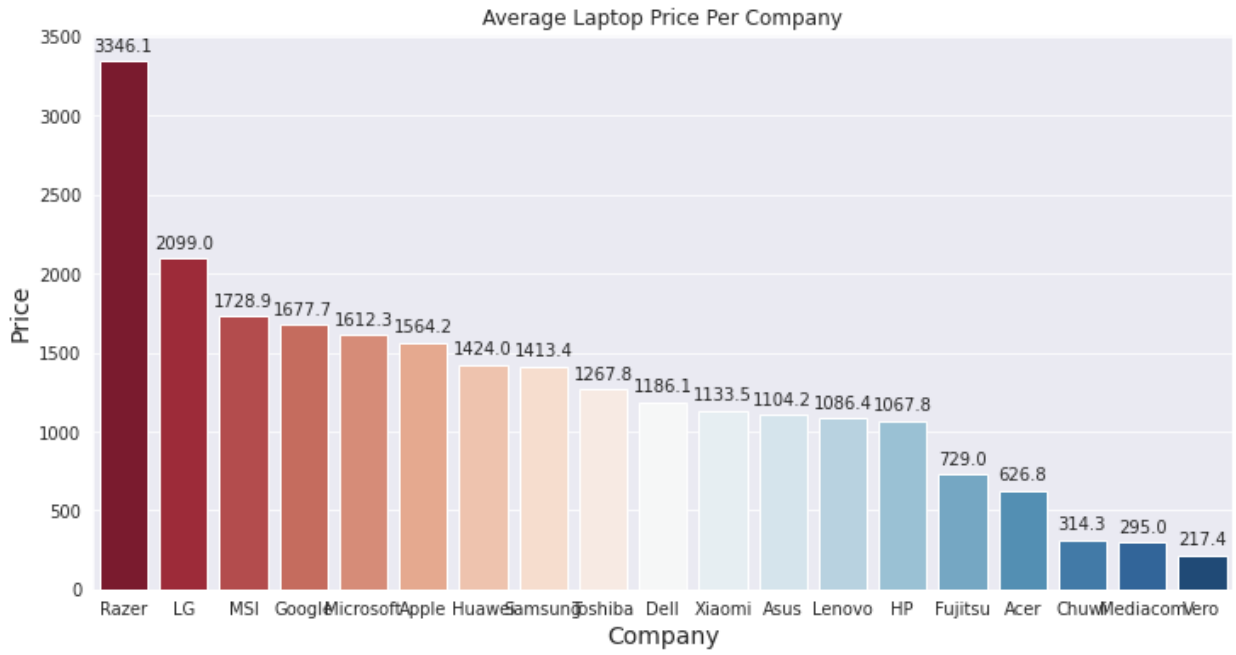

Fig 3. Average Laptop Price Per Company

At first glance, according to the dataset used, the Razer company is in first place with the highest laptop price of 3346.1 Euros, while the lowest belongs to the Mero company with a price of 217.4 Euros. Before creating a regression model, it is important to test for normality to determine whether the price data is normally distributed. If the data is not normally distributed, it can be normalized using the log1p (logarithm of 1 plus the input) function. Normalization is important because many statistical tests assume that the data is normally distributed. By normalizing the data, you can ensure that the results of the statistical tests are reliable and accurate.

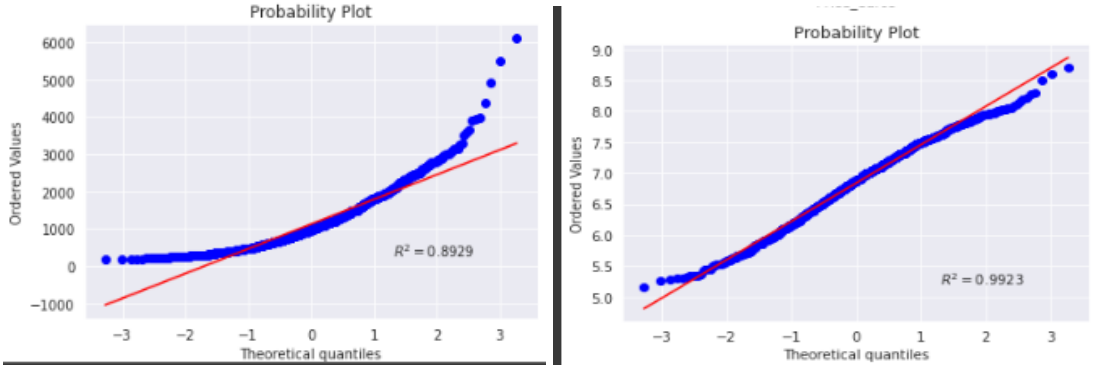

Fig 4. Probability plot before and after normalization

Based on the graph above, it can be seen that the data of the price variable shows a normal distribution both before and after normalization with R Squared Value 0,89 before normalization and 0,99 after normalization. In addition to testing for normality, it is also important to test for correlations between variables. Correlation tests can help to determine the strength and direction of the relationship between variables. There are several types of correlation tests, including Pearson's correlation coefficient and Spearman's rank correlation coefficient. One way to visualize correlations is to create a correlation map, which is a plot that shows the correlations between all pairs of variables in a dataset. Correlation maps can be created using the seaborn library in Python.

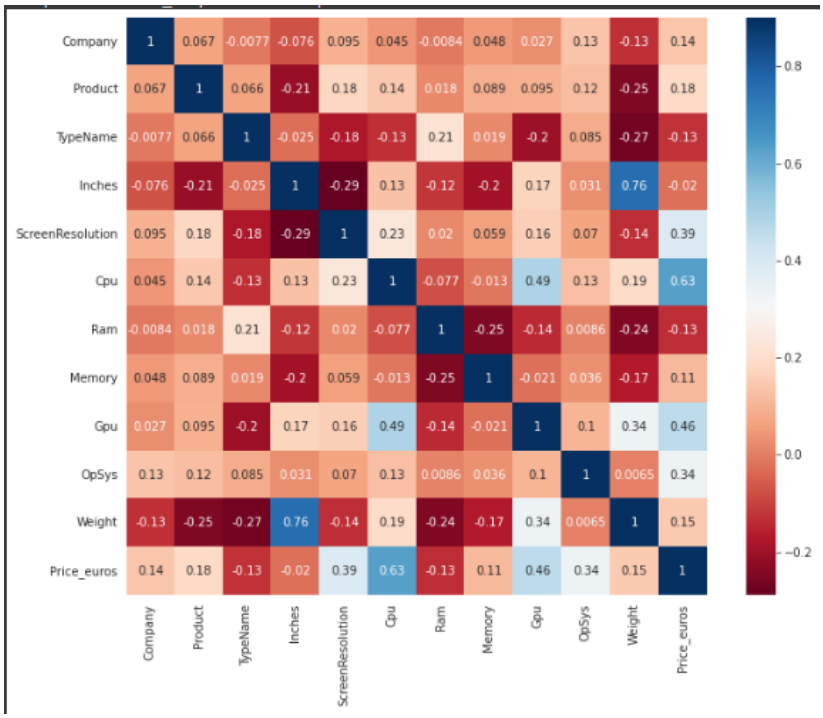

Fig 5.Correlation Map

Based on the graph above, the variable with the highest correlation with the variable price-euros is the variable cpu with a score of 0.63. Followed by gpu with a value of 0.46, then ScreenResolution and OpSys with values of 0.39 and 0.34.

The results of calculations from T test using scipy library show t-statistic and pvalue, where t-statistic has a score of -1.721 and pvalue has a score 0.0868. And the results of calculations from several linear regression models used show varied results. The results of several linear regression models performed are as follows :

a. Lasso Regression has a score of 0.4139
b. Elastic Net Regression has a score of 0.4139
c. Kernel Ridge Regression has a score of 0.302

After the results of each model are known, the next is boosting the model with Gradient Boosting Regression. Gradient Boosting Regression is a machine learning technique used to predict the value of a numerical variable. This technique is a variation of the boosting technique, which is a technique that combines several models that complement each other into one stronger model. In the gradient boosting technique, the model that is combined is a regression model.

The basic principle of gradient boosting is to create a better model by improving the previous model. Each new model will be built by optimizing the improvement of the previous model, by adjusting the weight of the model according to the error that occurred in the previous model.

After boosting, the following comparison results between the model and the test data show that the model built almost matches the given test data.

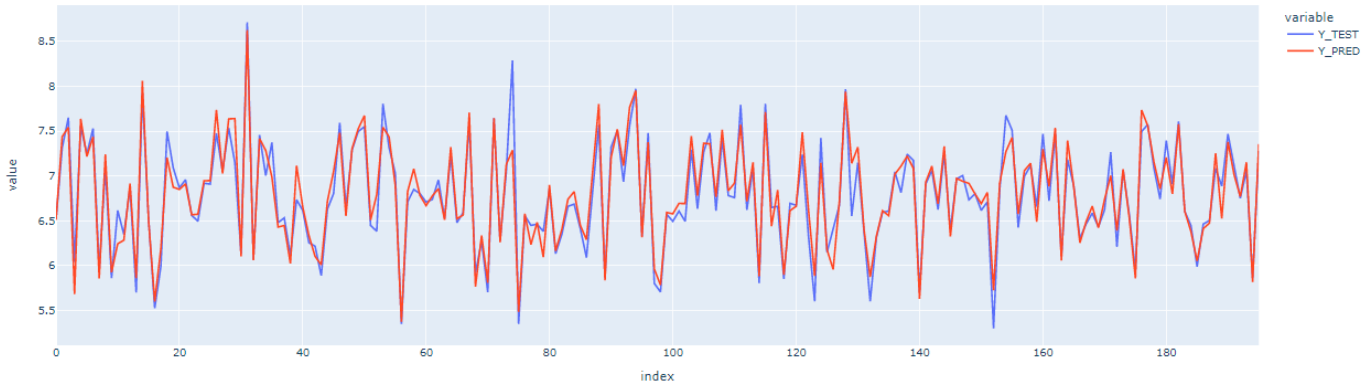

Fig 6 . Comparing Test Data with the Model

4. CONCLUSION

Based on the results of the research conducted, it can be concluded that the linear regression model can be used to predict laptop prices with fairly good accuracy, with an MAE value of 0.1, MSE of 0.03, and coefficient of determination (r^2) of 0.92. In addition, from the results it can also be concluded that the CPU and GPU features have the greatest influence on laptop prices, while other features have a smaller influence.

This conclusion is expected to provide benefits for consumers in making purchasing decisions, as well as for producers in determining the right pricing strategy. In addition, the results of this study can also serve as a reference for future research that wants to explore the relationship between features and laptop prices further.

There are several directions that future work on predicting laptop prices with a linear regression model could take. Some potential areas for future research include:

1. Incorporating additional predictor variables: One possibility is to include additional predictor variables in the model that might influence laptop prices, such as processor speed, graphics card, screen size, and storage capacity.

2. Using more advanced modeling techniques: Instead of using a linear regression model, researchers could explore the use of more advanced techniques such as nonlinear regression, decision trees, or neural networks.
3. Incorporating time-series data: If data on laptop prices over time are available, researchers could use time-series analysis techniques to model the trend and seasonality in laptop prices.
4. Validating the model on a new dataset: It would be useful to test the performance of the model on a new, independent dataset to determine its generalizability to other samples of laptop prices.
5. Exploring the impact of market forces: Researchers could examine how external factors such as economic conditions, technological innovations, and consumer preferences influence laptop prices and how these factors can be incorporated into the model.
6. Investigating the use of ensembles: Researchers could consider using ensemble methods, which combine the predictions of multiple models to improve the overall accuracy of the prediction.

ACKNOWLEDGMENTS

You may want to thank your funding source (but do not thank to any of the authors!).

REFERENCES

[1] Few, S. (2017). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.
[2] Munzner, T. (2017). Visualization Analysis and Design. CRC Press.
[3] Wattenberg, M. (2019). How to Use Machine Learning to Predict the Future. O'Reilly Media.
[4] Souza, P. (2022). Data Science for Dummies. John Wiley & Sons.
[5] Zhang, Y., Fan, X., Zhu, J., Chen, L., Xu, J. (2017). A novel hybrid approach based on linear regression and adaptive neuro-fuzzy inference system for short-term load forecasting. Energies, 10(10), 1-18.
[6] Chen, Y., Yang, J., Wang, H., Chen, X. (2019). A hybrid method based on linear regression and support vector machine for wind speed prediction. Renewable Energy, 142, 538-547.
[7] Jin, Y., Hu, H., Liu, Y., Zhang, Y., Fan, X. (2021). A hybrid approach based on linear regression and artificial neural network for short-term load forecasting. Energies, 14(1), 1-19.
[8] Gao, L., Song, Y., Cai, Y., Chen, Y., Wu, Y. (2017). A hybrid approach based on min-max normalization and support vector machine for wind speed prediction. Renewable Energy, 116, 366-375.
[9] Huang, Y., Jin, Y., Hu, H., Liu, Y., Zhang, Y. (2021). A hybrid approach based on z-score normalization and artificial neural network for short-term load forecasting. Energies, 14(1), 1-19.
[10] Wang, X., Zhang, Y., Fan, X., Chen, L., Xu, J. (2019). A novel hybrid approach based on decimal scaling and adaptive neuro-fuzzy inference system for short-term load forecasting. Energies, 12(12), 1-20.
[11] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. New York: Springer.
[12] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. New York: Springer

BIOGRAPHY OF AUTHORS

About the Author.-
(*Rony Wahyu Utama*)

About the Author.-
(*Riyandi Firman Pratama*)

About the Author.-
(*Muhamad Faisal Fiqri*)