JUSIKOM PRIMA (Jurnal Sistem Informasi dan Ilmu Komputer Prima)
Vol. 6 No. 1, Agustus 2022 E-ISSN : 2580-2879

Laptop Price Prediction with Machine Learning Using Regression Algorithm

Astri Dahlia Siburian, Daniel Ryan Hamonangan Sitompul, Stiven Hamonangan Sinurat
Andreas Situmorang, Ruben, Dennis Jusuf Ziegel, Evta Indra*
Prodi Sistem Informasi, Fakultas Teknologi dan Ilmu Komputer, Universitas Prima Indonesia
Jl. Sampul No.3, Sei Putih Barat, Medan Petisah
E-mail : * evtaindra@unprimdn.ac.id

ABSTRACT- Since the COVID-19 pandemic, many activities are now carried out in a Work From Home (WFH) manner. According to data from the Central Statistics Agency (BPS) of East Java, in 2021, large and medium-sized enterprises (UMB) who choose to work WFH partially are 32.37%, and overall WFH is 2.24% (BPS East Java, 2021). With this percentage of 32.37%, many people need a work device (in this case, a laptop) that can boost their productivity during WFH. WFH players must have laptops with specifications that match their needs to encourage productivity. To prevent buying laptops at overpriced prices, a way to predict laptop prices is needed based on the specified specifications. This study presents a Machine Learning model from data acquisition (Data Acquisition), Data Cleaning, and Feature Engineering for the Pre-Processing, Exploratory Data Analysis stages to modeling based on regression algorithms. After the model is made, the highest accuracy result is 92.77%, namely the XGBoost algorithm. With this high accuracy value, the model created can predict laptop prices with a minimum accuracy above 80%.

Kata kunci : *Machine Learning, Regression Algorithm, AutoML, Price Prediction, Laptop*

1. INTRODUCTION

Since the COVID-19 pandemic, many activities are now carried out in a Work From Home (WFH) manner. According to data from the Central Statistics Agency (BPS) of East Java, in 2021, large and medium-sized enterprises (UMB) who choose to work WFH partially are 32.37%, and overall WFH is 2.24% [1]. With this percentage of 32.37%, many people need a work device (in this case, a laptop) that can boost their productivity during WFH. To encourage work productivity, every WFH actor must have a laptop with specifications that suit his needs, and to prevent buying laptops at inappropriate prices, an appropriate way is needed to predict the price of a laptop based on the specified specifications.

There have been many studies that have the theme of predicting prices. To make price predictions, a regression algorithm is generally used in research [2]–[6]. Research [2] presents a method for predicting used car prices using a machine learning model with a regression algorithm configured with hyper-parameter tuning, while research [6] presents a method for predicting house prices using a machine learning model with a regression algorithm, namely XGBoost.

To overcome the problems mentioned, a Machine Learning method is needed to predict the price of a laptop based on parameters, namely laptop specifications. The author presents several machine learning models using a regression algorithm in this study. After modeling, the algorithm with the highest accuracy value will be used to predict the laptop's price, which will also be configured with AutoML to get a better accuracy value. With this method, it is expected that the model created can predict the price of a laptop with a minimum accuracy of above 80%.

2. METHODOLOGY

This study uses the Google Collaboratory tool to create a regression algorithm model. The workflow of the modeling can be seen in Figure 1. The research began with retrieving datasets from the Kaggle website; After the dataset is downloaded, the pre-process stage will be carried out, which consists of Data Cleaning and Feature Engineering; After the pre-processing stage is complete, the Exploratory Data Analysis (EDA) stage is carried out; And the last stage to do is Model Building where this stage consists of making a model with a default configuration to making a model with the help of AutoML.

2.1 *Data Acquisition*

In this study, the dataset came from Muhammet Varli's "Laptop Price" repository on the Kaggle website [7]. This dataset has 12 columns containing up to 1303 rows of data. Each column is a specification of a laptop in general, such as the brand, CPU, VGA/GPU, the price. Details of this "Laptop Price" dataset can be seen in Figure 2.

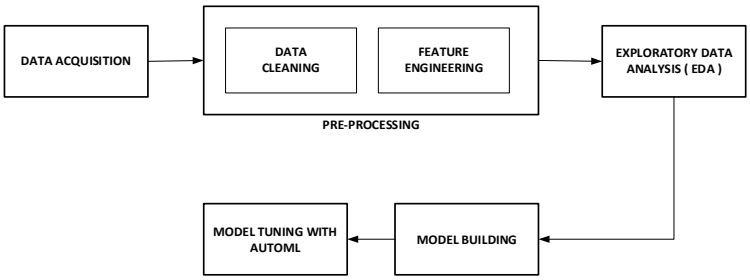

Figure 1. Research Methodology
Sumber [20]

	laptop_ID	Company	Product	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price_euros
0	1	Apple	MacBook Pro	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	1339.69
1	2	Apple	Macbook Air	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	898.94
2	3	HP	250 G6	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	575.00
3	4	Apple	MacBook Pro	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	2537.45
4	5	Apple	MacBook Pro	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37kg	1803.60

Figure 2. Dataset Detail

2.2 *Data Cleaning*

Data Cleaning increases the usability of the dataset used [8], [9]. For example, in this study, all column names in the dataset will be changed to lowercase. The syntax for performing Data Cleaning can be seen in Figure 3.

```
[ ]  # Melakukan Cek Nama Kolom Yang Ada Pada File CSV

     df = df.rename(columns=str.lower)
     df.columns

     Index(['laptop_id', 'company', 'product', 'typename', 'inches',
            'screenresolution', 'cpu', 'ram', 'memory', 'gpu', 'opsys', 'weight',
            'price_euros'],
           dtype='object')
```

Figure 3. Data Cleaning

2.3 *Feature Engineering*

Feature Engineering serves to create more features from existing datasets [10], [11]. In this study, the laptop specification column, such as CPU, will be broken down into several new columns, namely CPU Brand and CPU Clock. The syntax for performing Feature Engineering can be seen in Figure 4, and the overall results can be seen in Table 1.

```
[ ]  # MEMISAHKAN FREKUENSI CPU KE KOLOM LAIN

     df['cpu_freq'] = df['cpu'].str.extract(r'(\d+(?:\.\d+)?GHz)')

     df['cpu_freq'].value_counts()

     2.5GHz    290
     2.7GHz    165
     2.8GHz    165
     1.6GHz    133
     2.3GHz     86
     1.8GHz     78
     2.6GHz     76
     2GHz       67
     1.1GHz     53
     2.4GHz     52
     2.9GHz     21
     3GHz       19
     2.0GHz     19
     1.2GHz     15
```

Figure 4. Feature Engineering Syntax

Table 1. Entire Feature Engineering Process

Kolom	Setelah Data Cleaning
laptop_id	Column dropped
screenresolution	Addition of column resolution, screentype, touchscreen
cpu	Addition of column cpu_freq
ram	Removal of GB from data and Updating column name to ram(GB)
memory	Addition of column memory_type and memory_size
weight	Removal of Kg from data and Updating column name to weight(kg)

2.4 *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) is an approach to analyzing a set of data to infer its main characteristics. Generally, EDA is done by visualizing data with statistical charts [12], [13]. This study conducted EDA to see the correlation between columns and the average laptop price by brand. Visualization of the average laptop price by the brand can be seen in Figure 5.

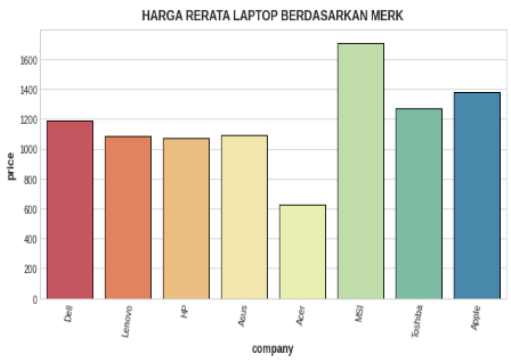

Figure 5. Harga Rerata Laptop Berdasar Merk

2.5 *Dataset Splitting*

The dataset will be split twice before model generation into train-testing data. In this study, the configuration of the dataset splitting used is 7:3, i.e., 70% of the dataset will be training data, while 30% will be testing data. The dataset splitting process can be seen in Figure 6.

```
[ ]  # TRAIN TEST SPLIT

    X_train, X_test, y_train, y_test = train_test_split(X,
                                                        y,
                                                        test_size = 0.3,
                                                        random_state = seed)
```

Figure 6. Proses Splitting Dataset

2.6 *Model Building*

In this study, three basic regression algorithms are generally used to create a machine learning model that can predict laptop prices, namely Random Forest, Gradient Boosting, and XGBoost.

2.6.1 *Random Forest Regressor*

Random Forest is an Ensemble Learning-based Machine Learning algorithm that operates by creating multiple Decision Trees for training. For regression, the average value of the predictions from each tree will be used [14], [15]. In this study, the Random Forest algorithm was created using the configuration estimator = 100, max_depth = 100, and max_features = 15. The detail can be seen in Figure 7.

```
[119] # MODEL RANDOM FOREST

     rf = RandomForestRegressor(n_estimators=100,
                                max_depth=100,
                                max_features=15)
     rf.fit(X_train,y_train)
     y_pred_rf = rf.predict(X_test)
```

Figure 7. Proses Pembuatan Model *Random Forest*

2.6.2 *Gradient Boosting Regressor*

Gradient Boosting Regressor (GBR) is an algorithm consisting of several decision trees to perform classification and regression tasks. Gradient Boosting generally works by creating a sequence of decision trees in which each tree focuses on the rest of the previous tree [16], [17]. The GBR algorithm is made with the default configuration in this study. The detail can be seen in Figure 8.

```
[133] # MODEL GRADIENT BOOSTING

      gbr = GradientBoostingRegressor()
      gbr.fit(X_train,y_train)
      y_pred_gbr = gbr.predict(X_test)
```

Figure 8. Proses Pembuatan Model *GBR*

2.6.3 *XGBoost Regressor*

XGBoost Regressor is Gradient Boosting optimized and designed to be more efficient, flexible, and portable. XGBoost presents a parallel tree that can solve Data Science problems quickly and accurately [18], [19]. In this study, the XGBoost Regressor model was created with AutoML configuration, where the AutoML Optuna library will automatically determine the model parameters. The detail can be seen in Figure 9.

```
[ ]  # XGBOOST OPTIMIZATION DENGAN OPTUNA

     xgb = XGBRegressor(reg_lambda = lambda_opt,
                        alpha = alpha_opt,
                        colsample_bytree = colsample_bytree_opt,
                        subsample_opt = subsample_opt,
                        learning_rate = learning_rate_opt,
                        n_estimators = n_estimators_opt,
                        max_depth = max_depth_opt,
                        min_child_weight = min_child_weight_opt)

[ ]  xgb.fit(X_train, y_train, eval_set=[(X_val, y_val)],
             early_stopping_rounds=50,
             verbose=0)
```

Figure 9. Proses Pembuatan Model XGBoost

3. RESULT AND DISCUSSION

3.1 Model Accuracy

After the model is made, a comparison will be made from the model. In this study, what is compared to the Random Forest, Gradient Boosting, and XGBoost models is the R2 value and Real Mean Squared Error (RMSE). Model comparison can be seen in Table 2.

Table 2. Model Comparison

Model	R2 Score
RF Regressor	80.79%
GB Regressor	90.55%
XGB Regressor	92.77%

From the table above, it can be concluded that XGBoost has the highest R2 value and the lowest RMSE. So XGBoost can be said to be better than other algorithms used to predict the price of laptops.

3.2 *Feature Importance Model*

After making the model, to help employees when buying a laptop, the model can also determine the essential laptop specifications. This study uses the three algorithms to determine that Random Access Memory (RAM) is an essential feature in a laptop for work purposes. Visualization of Feature Importances can be seen in Figure 10.

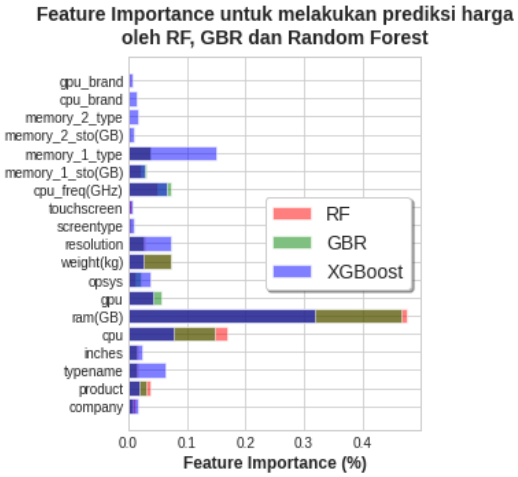

Figure 10. Feature Importance Model

3.3 Model Prediction

In this study, a prediction test is carried out on the validation dataset to see whether the model that has been made can also predict valid data. The three algorithms will be used in this case, and the scatter plot will be used to visualize the predictions. Visualization can be seen in Figure 11.

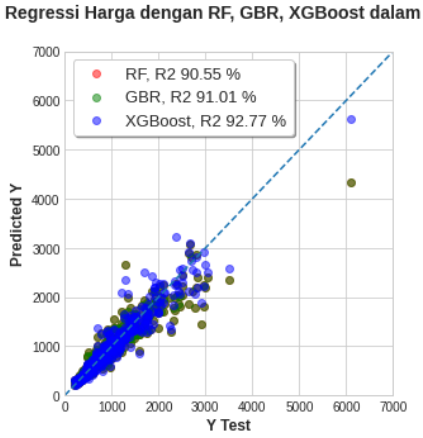

Figure 11. Scatter Plot Model

The visualization above shows that the three models made are close to the original value (price) contained in the validation dataset. So, the three models that have been made can also be used to predict laptop prices.

4. CONCLUSSION

With a model that can predict the price of a laptop, employees can more easily determine the laptop that suits their needs. In this study, the model with the highest R2 value and the lowest RMSE is XGBoost, with an R2 value of 92.77% and an RMSE of $221.66. Therefore, XGBoost can be said to be better than Random Forest and Gradient Boosting. The XGBoost model that has been created can also be used to make predictions in real-time using a web-based Machine Learning application.

BIBLIOGRAPHY

[1] BPS Jawa Timur, "Hasil Survei Kegiatan Usaha Pada Masa Pandemi COVID-19 Provinsi Jawa Timur," Mar. 2021.

[2] M. Radhi, D. Ryan Hamonangan Sitompul, S. Hamonangan Sinurat, and E. Indra, "PREDIKSI HARGA MOBIL MENGGUNAKAN ALGORITMA REGRESSI DENGAN HYPER-PARAMETER TUNING," *Jurnal Sistem Informasi dan Ilmu Komputer Prima*, vol. 4, no. 2, 2021.

[3] C. C. Emioma and S. O. Edeki, "Stock price prediction using machine learning on least-squares linear regression basis," *Journal of Physics: Conference Series*, vol. 1734, no. 1, p. 012058, Jan. 2021, doi: 10.1088/1742-6596/1734/1/012058.

[4] A. Varma, A. Sarma, S. Doshi, and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Apr. 2018, pp. 1936–1939. doi: 10.1109/ICICCT.2018.8473231.

[5] CH. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," in *2019 International Conference on Smart Structures and Systems (ICSSS)*, Mar. 2019, pp. 1–5. doi: 10.1109/ICSSS.2019.8882834.

[6] M. Ahtesham, N. Z. Bawany, and K. Fatima, "House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan," in *2020 21st International Arab Conference on Information Technology (ACIT)*, Nov. 2020, pp. 1–5. doi: 10.1109/ACIT50332.2020.9300074.

[7] Muhammet Varli, "Laptop Price," 2021. https://www.kaggle.com/datasets/muhammetvarl/laptop-price (accessed Apr. 19, 2022).

[8] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, "Data Cleaning," in *Proceedings of the 2016 International Conference on Management of Data*, Jun. 2016, pp. 2201–2206. doi: 10.1145/2882903.2912574.

[9] K. H. Tae, Y. Roh, Y. H. Oh, H. Kim, and S. E. Whang, "Data Cleaning for Accurate, Fair, and Robust Models," in *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning - DEEM'19*, 2019, pp. 1–4. doi: 10.1145/3329486.3329493.

[10] Z. Li, X. Ma, and H. Xin, "Feature engineering of machine-learning chemisorption models for catalyst design," *Catalysis Today*, vol. 280, pp. 232–238, Feb. 2017, doi: 10.1016/j.cattod.2016.04.013.

[11] M. Uddin, J. Lee, S. Rizvi, and S. Hamada, "Proposing Enhanced Feature Engineering and a Selection Model for Machine Learning Processes," *Applied Sciences*, vol. 8, no. 4, p. 646, Apr. 2018, doi: 10.3390/app8040646.

[12] N. Verbeeck, R. M. Caprioli, and R. van de Plas, "Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry," *Mass Spectrometry Reviews*, vol. 39, no. 3, pp. 245–291, May 2020, doi: 10.1002/mas.21602.

[13] T. Milo and A. Somech, "Automating Exploratory Data Analysis via Machine Learning: An Overview," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, Jun. 2020, pp. 2617–2622. doi: 10.1145/3318464.3383126.

[14] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer, "Robust and Accurate Shape Model Fitting Using Random Forest Regression Voting," 2012, pp. 278–291. doi: 10.1007/978-3-642-33786-4_21.

[15] A. Khalyasmaa *et al.*, "Prediction of Solar Power Generation Based on Random Forest Regressor Model," in *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, Oct. 2019, pp. 0780–0785. doi: 10.1109/SIBIRCON48586.2019.8958063.

[16] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308–324, Sep. 2015, doi: 10.1016/j.trc.2015.02.019.

[17] X. Li, W. Li, and Y. Xu, "Human Age Prediction Based on DNA Methylation Using a Gradient Boosting Regressor," *Genes (Basel)*, vol. 9, no. 9, p. 424, Aug. 2018, doi: 10.3390/genes9090424.

[18] S.-E. Ryu, D.-H. Shin, and K. Chung, "Prediction Model of Dementia Risk Based on XGBoost Using Derived Variable Extraction and Hyper Parameter Optimization," *IEEE Access*, vol. 8, pp. 177708–177720, 2020, doi: 10.1109/ACCESS.2020.3025553.

[19] S. Li and X. Zhang, "Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm," *Neural Computing and Applications*, vol. 32, no. 7, pp. 1971–1979, Apr. 2020, doi: 10.1007/s00521-019-04378-4.

[20] A. Amalia, M. Radhi, S. H. Sinurat, D. R. H. Sitompul, and E. Indra, "PREDIKSI HARGA MOBIL MENGGUNAKAN ALGORITMA REGRESSI DENGAN HYPER-PARAMETER TUNING", JUSIKOM PRIMA, vol. 4, no. 2, pp. 28 -32, Feb. 2022.