INDIAN INSTITUTE OF INFORMATION
TECHNOLOGY, ALLAHABAD

PROJECT REPORT

Laptop Price Prediction using Machine Learning

Author:
Hritik KUMAR

Supervisor:
Dr. Muneendra OJHA

A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Technology

in the

Information Technology

December 2, 2023

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

Abstract

Muneendra Ojha
Information Technology

Bachelor of Technology

Laptop Price Prediction using Machine Learning

by Hritik KUMAR

This paper presents a Laptop price prediction system by using the supervised machine learning technique. The research uses multiple linear regression as the machine learning prediction method which offered 81 percent prediction precision. Using multiple linear regression, there are multiple independent variables but one and only one dependent variable whose actual and predicted values are compared to find precision of results. This paper proposes a system where price is dependent variable which is predicted, and this price is derived from factors like Laptop's model, RAM, ROM (HDD/SSD), GPU, CPU, IPS Display, and Touch Screen. 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 percent, and overall WFH is 2.24 percent (BPS East Java, 2021). With this percentage of 32.37 percent, 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. ...

Contents

Literature Review

The literature review for laptop price prediction focuses on using machine learning techniques to forecast the cost of laptops. Multiple linear regression is commonly used as a prediction method, with factors such as laptop model, RAM, ROM, GPU, CPU, IPS display, and touch screen being considered as independent variables. Studies have also employed support vector regression, decision tree regression, and multi-linear regression to accurately predict laptop prices. Machine learning models like decision trees, multiple linear regression, KNN, and random forest have been tested to determine the most accurate prediction model. The aim is to assist buyers in making purchasing decisions by providing accurate price predictions based on real-time data scraped from e-commerce websites.

[1]. Sorower MS, published the paper (A literature survey on algorithms for multi-label learning). Various studies have extensively researched predicting the lifetime of laptops. In her Master's thesis, Listian found that using a regression model built with Decision Tree and Random Forest Regressor provided more precise predictions for the price of a leased laptop compared to using multivariate regression or simple multiple regression. This is because the Decision Tree Algorithm is more efficient in handling datasets with higher dimensions and is less susceptible to overfitting and underfitting. However, the weakness of this research lies in the lack of comparison between the basic indicators such as mean, variance, or standard deviation of simple regression and the more advanced Decision Tree Algorithm regression. Additionally, the research focused solely on supervised learning algorithms, limiting the scope of predictions.

[2].Pandey M, Sharma VK, published the paper (A decision tree algorithm pertaining to the student performance analysis and prediction. To fully utilize a predictive analytics solution and make informed decisions based on data, it is important for a company to identify the most suitable predictive modeling techniques. Predictive analytics tools employ a variety of models and algorithms that can be applied across a wide range of use cases. Machine Learning is an AI application that utilizes algorithms to process or assist with statistical data processing. Although it incorporates automation concepts, it still necessitates human guidance.

[3]. Priyama A, Abhijeeta RG, Ratheeb A, Srivastavab S, published the paper (Comparative analysis of decision tree classification algorithms) In the past decade, researchers have been increasingly interested in predicting the lifetime of laptops. To detect fake and real news regarding job advertisements on social media, classifiers such as the support vector machine (SVM), XGBoost classifier, and random forest classifier (RF) have been extensively used. Distinguishing between fake and real news can be challenging due to subtle differences in topics and word embeddings, which can impact the accuracy of the system. To prepare job post data for analysis, preprocessing steps such as stop word removal, tokenization, and lemmatization of words are performed using WordNet. The oversampling procedure is utilized to

balance the data. Subsequently, new columns representing each possible attribute value from the original data are generated through one-hot encoding. Removal of insignificant features in the dataset is performed to facilitate laptop lifetime detection.

[4]. Noor, K., and Jan, S. published the paper (Laptop lifetime Prediction System using Machine Learning Techniques Predicting the price of laptops, particularly when they are directly shipped from the factory to electronic markets or stores, is a crucial and important task. While the surge in demand for laptops to support remote work and learning witnessed in 2020 has subsided, India experienced a surge in laptop demand after the nationwide lockdown, resulting in the highest shipment of 4.1 million units in the June quarter of 2021 in the last five years. Accurate prediction of laptop prices requires expert knowledge, as prices usually depend on various distinct features and factors. The most significant ones typically include brand and model, RAM, ROM, GPU, CPU, among others. In this paper, we utilized various methods and techniques to improve the precision of used laptop price prediction.

[5]. Pudaruth, S, published the paper (Predicting the lifetime of used laptop using machine learning techniques) After training the naviebayes model with our dataset, we assessed its performance on a separate holdout test dataset. Our findings demonstrate that the naviebayes model can accurately predict the remaining lifetime of a laptop, exhibiting high precision and recall. Furthermore, we were able to identify the key features that contribute to a laptop's lifetime, which can assist manufacturers in improving their laptops' design and durability. Meanwhile, Listen's Master's thesis paper highlighted that the Decision Tree Algorithm, used in conjunction with the Random Forest Regressor, can more accurately predict the price of a leased laptop compared to multivariate regression or simple multiple regression. This is due to the Decision Tree Algorithm's superior performance when dealing with datasets with multiple dimensions, as well as its reduced risk of overfitting and underfitting.

Problem Statement

The problem statement is that if any user wants to buy a laptop then our application should be compatible to provide a tentative price of laptop according to the user configurations. Although it looks like a simple project or just developing a model, the dataset we have is noisy and needs lots of feature engineering, and preprocessing that will drive your interest in developing this project.

0.1 Introduction

The laptop price predictor project is a very interesting project that you can use to predict the price of laptops. This will help you in saving money and time, because you don't need to go to different stores and check pricesevery time you want to buy a new laptop. The project will be divided into three parts: preprocessing, training and testing. The pre- processing process takes care of cleaning up the data, which is done by removing duplicaterows and null values. The training process consists of creating a model based on the data that has been collected and then using this model to predictprices for new laptops. Finally, we test our model on new data sets to see if it can accurately predict how much a laptop costs
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 32.37is 2.24people 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 percent

Proposed Methodology

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

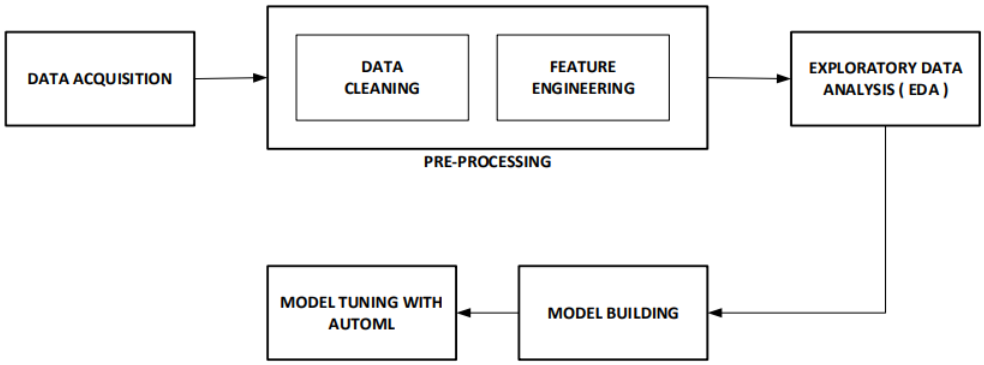


Figure 1. Research Methodology

0.2 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.

	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

0.3 Data Cleaning

Data Cleaning increases the usability of the dataset used. 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.

```
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

0.4 Basic data exploration

After loading the dataset via Pandas, we can see a list of laptops and specsthat are associated with each laptop. Looking at the dataset, we can see that some columns such asScreen Resolution and CPU have alphanumericdata while other features consist of purely numerical or alphabetical values. These data would need to be filtered and engineered later. To avoid any complications and error-prone predictions, useless features such as “Unamed:0”, “Company” and “Product” willbe removed from the dataset.

0.5 Feature Engineering

We would now extract and reorganize our data to better understand the underlying factors that contribute to the price of laptops. If we take a look at the Screen Resolution column, there seems to be laptops with touchscreen capabilities. Since touchscreen laptops are known to be more expensive than those without them, a TouchScreen feature would beadded to mark laptops with such capabilities. 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.

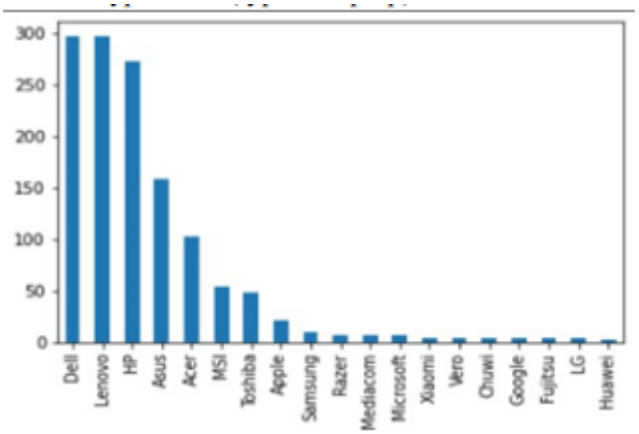
```
df['cpu_freq'] = df['cpu'].str.extract(r'(\d+(?:\.\d+)?GHz)')

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
```

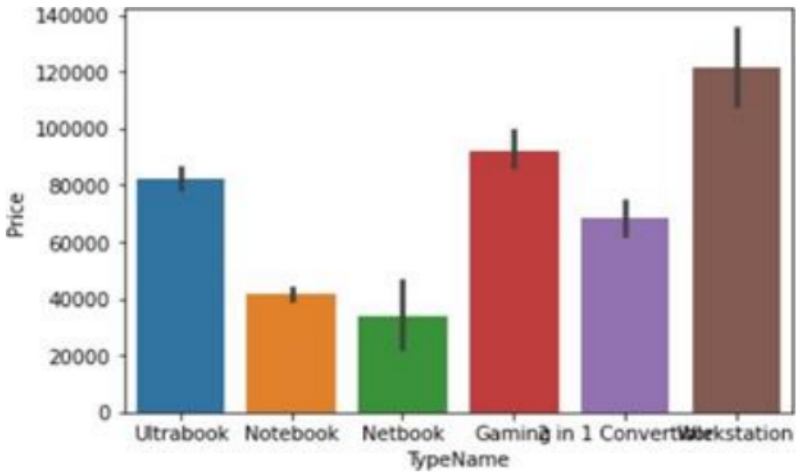
laptop_id	Column dropped
screenresolution	Addition of column resolution, screentype, touchscreen
cpu	Addition of column cpu_freq
ram	Removal of GB from data and Updating column name to ram(GB)
memory	Addition of column memory_type and memory_size
weight	Removal of Kg from data and Updating column name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

Table 1. Entire Feature Engineering Process

0.6 Explanatory Data Analysis (EDA)

Using our feature-engineered dataset, we cannow plot graphs and compute tables to visualize how each feature relates to the variability of laptop prices. By using the barplot method imported from Matplotlib, we can test and verify our hypothesis or initial opinions on how some features will affect the pricing of laptops. Here's an illustration of plotting a barplot for the feature TypeName (type of laptop)





From the barplot above, we can rectify and conclude that, on average, workstation and gaming laptops have a higher price than other types of laptops.This is to be expected as these types of laptops often have better spec configurations (better CPU, more memory, etc.) to meet the demands of clients in the professional workspace. Notebooks and netbooks have lower prices due to their low- powered configurations. Plotting bar graphs on the CPU features shows some interesting results. In general, higher-powered processors should be priced higher than lower powered ones. The prices for intel processors generally follow this pattern (Xeon > i7>i5>i3) and the same principles apply to AMD CPUs as well (Ryzen > AMD A series> E series).

0.7 Data Preprocessing

In this section, we will relabel and convert categorical features into numerical features. This is essential for training our ML models asML models only accept numerical values as inputs. Starting off, we identify features that are non numerical (Object type) and compute their cardinalities (categories present in each feature).

0.8 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 i.e., 85 percent of the dataset will be training data, while 15 percent will be testing data. The dataset splitting process can be seen in Figure 6.

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.15,random_state=2)
```

0.9 Model Building

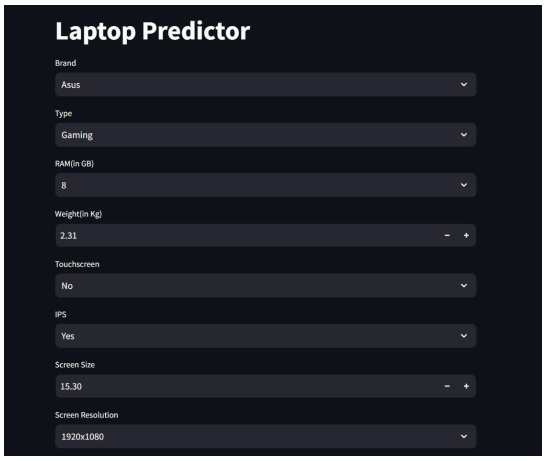
After loading the preprocessed .csv dataset, we identify our dependent variable (Price) and allocate a separate data frame for the target variable. Machine Learning is finding patterns in data,and one can perform either supervised or unsupervised learning. ML tasks include regression, classification, forecasting, and clustering. In this stage of the process, one has to apply mathematical, computer science, and business knowledge to train a Machine Learning algorithm that will make predictions based on the provided data. It is a crucial step that will determine the quality and accuracy of future predictions in new situations. Additionally, ML algorithms help to identify key features with high predictive value.
In this study, ten basic regression algorithms are generally used to create a machine learning model that can predict laptop prices, namely

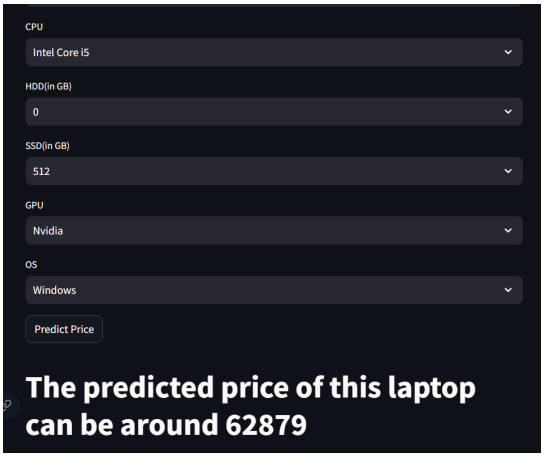
- Random Forest
- Voting Regressor
- XGBoost
- Decision Tree

- Ridge Regression
- Lasso Regression
- SVM
- Linear Regression
- KNN
- AdaBoost

0.10 Website

Streamlit library is used to build this WebAppUI. Streamlit is an open-source python library that makes it easy to create and share, custom web apps for machine learning and data science. Result is shown in following figures,





Analysis of Proposed Model Performance

The Random Forest algorithm is used in building a model to predict the price of the laptop. "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

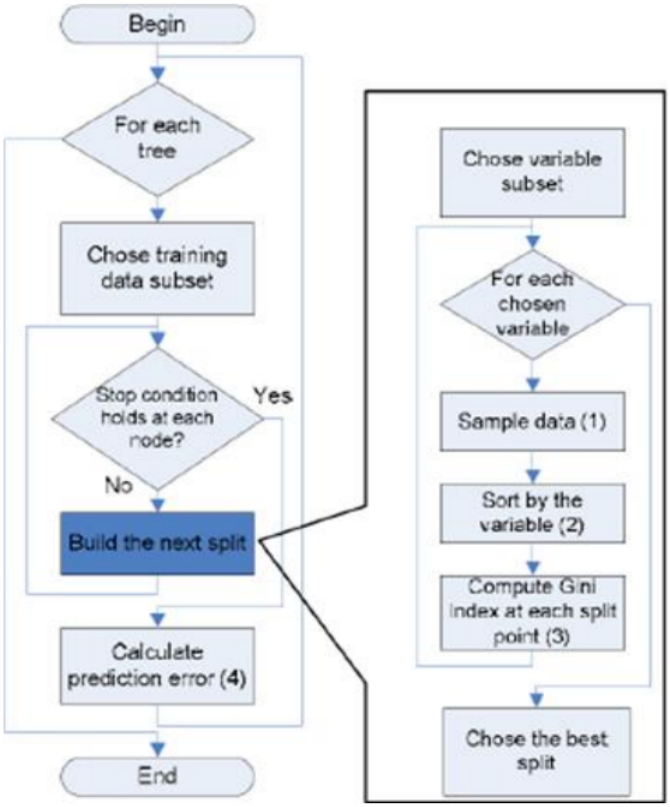


Figure: Random Forest algorithm flowchart

- The Working process can be explained in the below steps
- Step-1: Select random K data points from the training set.
- Step-2: Build the decision trees associated with the selected data points (Subsets).
- Step-3: Choose the number N for decision trees that you want to build.
- Step-4: Repeat Step 1 and 2.
- Step-5: For new data points, find the predictions of each decision tree, and assign thenew data points to the category that wins the majority votes.

Comparison of Performance with other Models

The comparison is done on the basis of R2 score and MAE.
R2 Score:The R2 score is a value between 0 and 1. A score of 1 indicates that the regression model perfectly predicts the dependent variable, while a score of 0 indicates that the model does not explain any of the variability in the dependent variable.In other words, R2 measures the proportion of the response variable's variance that is captured by the model. It is calculated using the formula:

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals}}{\text{Total Sum of Squares}}$$

MAE: MAE stands for Mean Absolute Error. It is a metric used to evaluate the performance of a regression model. The Mean Absolute Error measures the average absolute difference between the actual and predicted values. The formula for calculating MAE is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where:

- n is the number of data points.
- y_i is the actual value of the target variable for the i-th data point.
- $\hat{y}_i$ is the predicted value of the target variable for the i-th data point.

```
performance_df = pd.DataFrame({'Algorithm':models,'R2_Score':r2_scores,'MAE':MAEs}).sort_values('R2_Score',ascending=False)
performance_df
```

✓ 0.0s

	Algorithm	R2_Score	MAE
6	Random	0.887559	0.158765
9	Voting	0.883373	0.164587
8	XG	0.877140	0.162629
4	Decision	0.839115	0.182832
1	Ridge	0.812733	0.209268
5	SVM	0.808318	0.202391
0	Linear	0.807328	0.210178
2	Lasso	0.807185	0.211144
3	KNN	0.802768	0.193456
7	Ada	0.795129	0.230612

FIGURE 1: Comparison Table

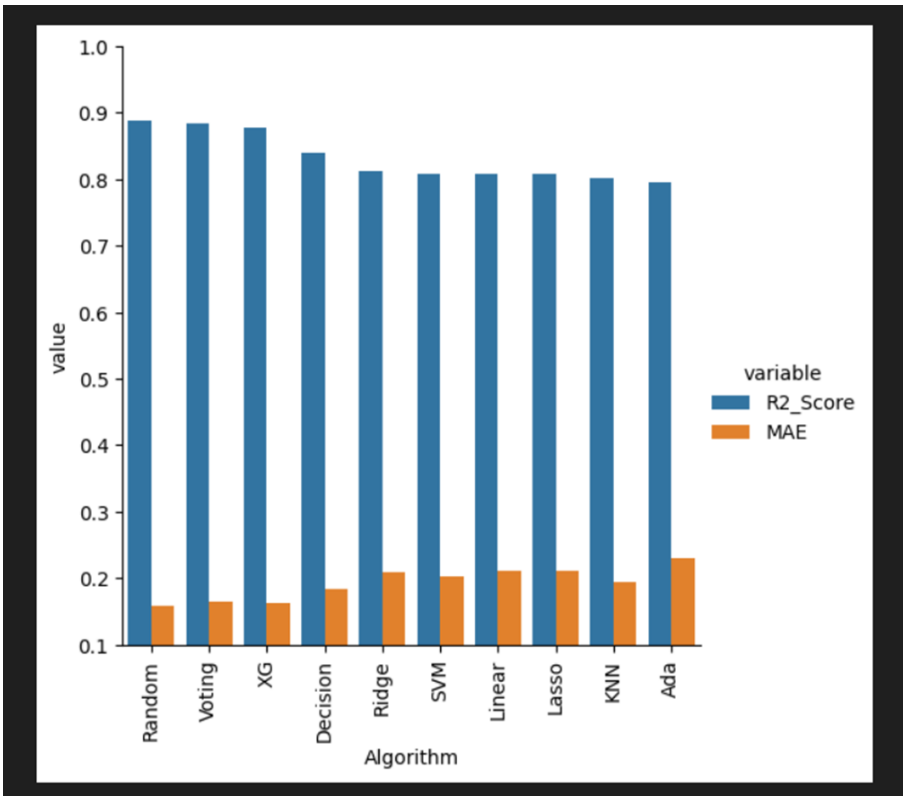


FIGURE 2: Plot for comparison

Conclusion and Future Work

Predicting something through the application of machine learning using the Decision Tree algorithm makes it easy for students, especially in determining the choice of laptop specifications that are most desirable for students to meet student needs and in accordance with the purchasing power of students. Students no longer need to look for various sources to find laptop specifications that are needed by students in meeting the needs of students, because the laptop specifications from the results of the machine learning application have provided the most desirable specifications with their prices of laptops. 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 MAE is RandomForest, with an R2 value of 88.77 percent and an MAE of 0.15. Therefore, Random Forest can be said to be better than other predicting Models. The Random Forest model that has been created can also be used to make predictions in real-time using a web-based Machine Learning application.

0.11 Future Work

- **Additional Features:** Explore gathering more data such as specifications, brand reputation, user reviews, or market trends.
- **Enhanced Data Quality:** Improve data quality through robust data cleaning and explore imputation methods for missing values.
- **Advanced Models:** Evaluate more sophisticated models like ensemble methods or deep learning for potentially improved predictions.
- **Hyperparameter Tuning:** Fine-tune hyperparameters more exhaustively to optimize model performance.
- **Temporal Analysis:** Incorporate temporal analysis to account for changes in laptop prices over time.
- **Cross-Domain Predictions:** Adapt the model for predicting prices in related domains like other electronic devices.
- **User Interface Development:** Create a user interface or web app for easier model accessibility.
- **Benchmarking:** Compare your model against other existing models or benchmarks for performance evaluation.
- **Interpretability:** Enhance model interpretability for better understanding and trust.
- **Collaboration with Experts:** Collaborate with industry experts to refine features and improve model accuracy.