IJMSIT

International Journal of Multidisciplinary Studies and Innovative Technologies

e-ISSN: 2602-4888
dergipark.org.tr/en/pub/ijmsit
Research Article

2022, 8 (1), 40-45
DOI: 10.36287/ijmsit.8.1.7
Received: January 7, 2024; Accepted: July 1, 2024

Laptop Price Range Prediction with Machine Learning Methods

Yasin Karakuş[1*], Turgay Tugay Bilgin[2]

[1]Department of Computer Engineering, Graduate School, Bursa Technical University, Bursa, Turkey (22435004005@ogrenci.btu.edu.tr) (ORCID: 0000-0002-4534-0151)*

[2]Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Bursa Teknik University, Bursa, Türkiye (turgay.bilgin@btu.edu.tr) (ORCID: 0000-0002-9245-5728)

Abstract – Prices forecasting, and price range estimation studies are very important for laptops, which have a very wide usage area, number of users and a large market share. Most existing price prediction studies use regression-based methods to estimate a concrete value for price. However, for many real-world applications, it is much more practical to predict a price class (or range). Although there are many studies on laptop price prediction in the literature, there is only one study on laptop price range prediction. The fact that the prices are divided into three different classes in this study does not overlap much with the laptop price range prediction problem in the real world. In addition, very few machine learning methods have been tested on the laptop price range prediction problem. To overcome these problems and contribute to the literature, a dataset previously used for laptop price prediction was adapted to be used for laptop price range prediction and the dataset was optimized for laptop price range prediction by applying preprocessing steps such as data cleaning, feature engineering and label encoding. Then, price range predictions were produced with machine learning methods such as random forest, histogram-based boosting, extra trees and catboost classifiers. When the success of the classifiers was tested, the best classifier was histogram-based boosting classifier with 70% accuracy.

Keywords *– Laptop Price, Price Range, Prediction, Machine Learning, Classification*

Citation: Karakuş, Y., Bilgin, T. T. (2024). Laptop Price Range Prediction with Machine Learning Methods. International Journal of Multidisciplinary Studies and Innovative Technologies, x(x): xx-xx.

I. Introduction

In Computers have been an important tool for human life since the first moment they were invented. Computers, which were used only for mathematical calculations in the early days and were quite rare, have decreased in size and weight over time and increased in capacity and speed. This has led to the widespread use of computers. Today, computers are used in workplaces, in educational institutions such as universities to make education easier and more efficient, in government institutions and defense industry to be used in official affairs, and in personal use as a means of socializing, shopping, and having fun. Computers are used for a myriad of tasks such as text processing, tabulation, database management, viewing websites, and performing mathematical and statistical calculations. Laptops are the portable version of computers and were introduced in the 1980s. Laptops can do all the work and perform all the tasks that computers do. The global laptop market reached a value of approximately USD 140.83 billion in 2022 and is expected to grow further at a CAGR of approximately 4.7% between 2023 and 2028 [1].

A. Related Work

Machine learning can be divided into 3 categories: supervised learning, unsupervised learning, and reinforcement learning [2]. Regression and classification are frequently used tasks in supervised learning. Price prediction studies are usually handled as regression problems in supervised learning. Price prediction range studies are usually handled as classification problems in supervised learning.

Reedy et al. [3] used support vector regression, decision tree regression and multiple linear regression models to predict laptop prices using real-time data from websites. Decision tree regression showed the best result with 0.078 mean squared error (MSE), 0.1167 mean absolute error (MAE) and 92.7274% r2 score.

Siburian et al. [4] used random forest regressor, gradient boosting regressor and XGBoost regressor models to predict laptop price using the dataset obtained from Kaggle. XGBoost regressor showed the best result with an r2 score of 92.77%.

Shaik et al. [5] used decision trees, multiple linear regression, k-nearest neighbors (KNN), and random forest models to predict laptop price using the data set obtained from Kaggle. Random forest showed the best result with 0.1587 MAE and 88.75% R2-score.

Syed et al. [6] used super vector machines (SVM), artificial neural networks, decision Tree, and multinomial logistic regression models to predict the laptop price range using laptop price data set obtained from online environment. The SVM model with linear kernel parameters gave the best result with an accuracy score of 0.99.

Ma et al. [7] proposed a cost-sensitive deep forest method to estimate the price range. They tested their proposed method on real datasets of car sharing, house rental and real estate sales. The experimental results show that their proposed method can significantly reduce the cost compared to traditional deep forest and other methods.

Nasser et al. [8] used artificial neural networks to predict the mobile phone price range using a dataset called "Mobile Price Classification" obtained from Kaggle and achieved 96.31% accuracy.

Rahman et al. [9] aim to predict cow price ranges using any cow image. They collected cow images from different online e-commerce sites selling cows and used convolutional neural networks (CNN) method to classify cow images. For price range classification, they used linear regression to classify cows into four different price ranges: low, medium, high, and very high. Their results showed that the price range of a cow can be predicted with 70% accuracy.

Güvenç et al. [10] compared KNN and deep neural network (DNN) models in mobile phone price range prediction. They obtained the data set from Kaggle. According to the comparison results, KNN achieved 93% accuracy and DNN achieved 94% accuracy.

Gegic et al. [11] used used car data collected from online platforms and artificial neural networks, SVM and random forest method to predict car price range. They evaluated their model using test data and obtained 87.38% accuracy.

Yücebaş et al. [12] developed a C4.5- CART tree model using real-time datasets of feature prices obtained from the web. This model is capable of both price and price range prediction. They obtained an root-mean-square error (RMSE) of 13.169 for price prediction and 81% Kappa and 88% Precision scores for price range prediction.

Ortu et al. [13] tested four different deep learning algorithms (Multilayer Perceptron (MLP), CNN, Long Short-Term Memory (LSTM) neural network and Attention Long Short-Term Memory (ALSTM) neural network) by taking into account the technical, trading and social indicators of the two main cryptocurrencies Ethereum and Bitcoin in 2017-2020 to predict the prices of cryptocurrencies. For the daily classification task, they found an increase in accuracy from 51%-55% for the constrained model to 67%-84% for the unconstrained model.

B. Observations and Contributions

Most existing price prediction work uses regression-based methods to estimate a concrete value for price. However, for many real-world applications, it is much more practical to predict a price class (or range) [7]. Although there are many studies in the literature for laptop price prediction, only one laptop price range prediction study was encountered in the literature search. The fact that the prices are divided into three different classes in this study does not overlap much with the laptop price range prediction problem in the real world. In addition, very few machine learning methods have been tested in the laptop price range prediction problem. Based on the data we have, a study was carried out using the price ranges used by technology stores in laptop sales and using random forest classifier, histogram-based boosting classifier, extra trees classifier, and catboost classifier machine learning models, which have not been used in previous studies.

Firstly, a dataset previously used for price prediction is made suitable for use for price range prediction through various preprocessing steps. Then, the success of the trained models for laptop price range prediction is tested using machine learning algorithms. This paper consists of four sections. Section 2 describes the dataset, data preprocessing steps and the laptop price range prediction model. Section 3 presents the results of testing the performance of the trained models with performance metrics. Section 4 contains comments on the study and future work.

II. Materials and Method

The flow diagram of the proposed method for the laptop price range is given in Figure 1.

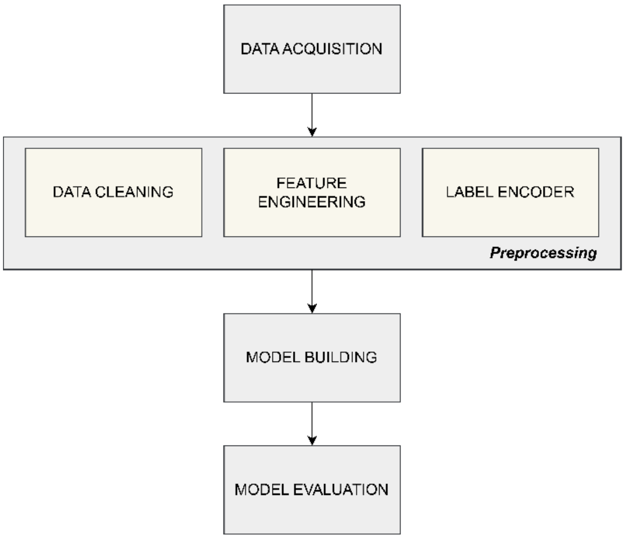

Fig. 1. Flowchart of the methodology used for laptop price range estimation.

A. Dataset

The dataset used in this study is obtained from Kaggle [14]. The dataset for laptop price prediction contains 13 columns and 1303 rows. The details of the dataset are shown in Table 1.

Table 1. Detailed information about the laptop price dataset.

Feature Name	Data Type	Explanation
Laptop_ID	Numeric	Laptop ID
Company	String	Laptop Manufacturer
Product	String	Brand and Model
TypeName	String	Type (Gaming etc.)
Inches	Numeric	Screen size
ScreenResolution	String	Screen Resolution
Cpu	String	Cpu model, speed
Ram	String	Ram memory (GB)
Memory	String	Memory (HDD etc.)
Gpu	String	Gpu model
OpSys	String	Operating System
Weight	String	Laptop Weight (Kg)
Price_euros	Numeric	Price (euro)

The number of laptops according to the "TypeName" feature is shown in Figure 2.

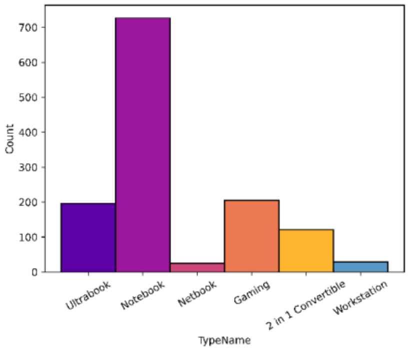

Fig. 2. Number of laptops by "TypeName" feature.

The number of laptops according to the "Company" feature is shown in Figure 3.

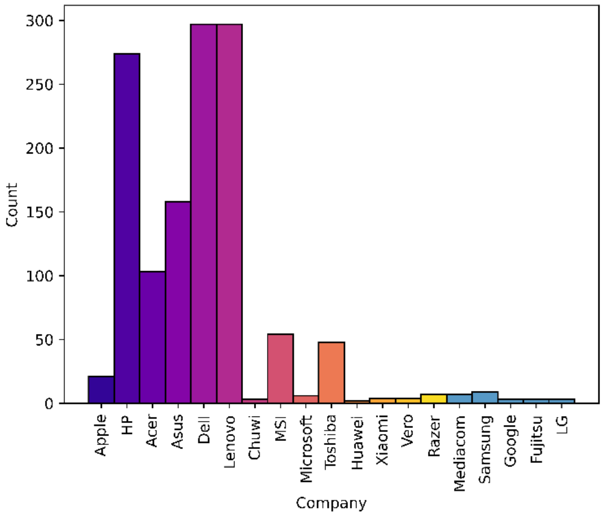

Fig. 3. Number of laptops by "Company" feature.

B. Data Preprocessing

Data preprocessing is one of the most important steps that increases the accuracy and efficiency of the machine learning model, consisting of steps such as removing noise in the data set, checking for empty values, and normalization.

B.A. Data Cleaning

In the data cleaning step, features that could lead to complications and error-prone estimates were removed to optimize the dataset for processing. These are: "laptop_ID, Company, Product".

In the rows of some features, qualifying unit expressions (such as "Kg", "GB") create noise if they do not make a difference within the feature. However, since most machine learning models do not accept textual expressions, they need to be removed. In this direction, the expression "GB" was removed from the columns in the "Ram" feature and the expression "Kg" was removed from the columns in the "Weight" feature.

B.B. Feature Engineering

From some of the features in the dataset, new and meaningful features can be derived to help us solve the problem.

Looking at the "Screen" feature, some laptops have a touch screen. The fact that the laptops have a touch screen directly affects the price. For this, "touchscreen" feature has been created.

From the "Cpu" feature, the "Cpu-speed" feature is derived, which directly affects the price. Each "Cpu" is then grouped according to its brand, model, and family. An example of this grouping is shown in Table 2.

Table 2. An example of grouping the columns in the Cpu feature according to make, model and family

Value	Group Number
Intel Core i3 6006U	0
Intel Core i3 6100U	0
Intel Core i3 7100U	0
Intel Core i5 6260U	1
Intel Core i5 6300U	1
Intel Core i5 6440HQ	1
AMD E-Series E2-9000e	8
AMD E-Series E2-6110	8
AMD E-Series 6110	8

A meaningful grouping was also made for the "Gpu" feature. The "Price_euros" feature was also grouped according to the price ranges in the report titled "Global Laptop Market Outlook" [1] published by Expert Market Research to be meaningful and suitable for the real world. This grouping is shown in Table 3.

Table 3. Representation of the "price_euros" feature split into ranges

Price Range (Euro)	Group
0 – 499.99	Very Cheap
500 – 999.99	Cheap
1000 – 1499.99	Medium
1500 – 1999.99	Expensive
2000 - …	Very Expensive

The pie chart for the price range is shown in Figure 4.

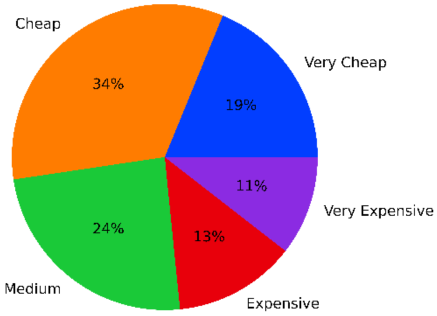

Fig. 4. Percentage pie chart of laptops by "price range" feature

B.C. Label Encoding

Label encoding is the digitization of features. This is because most machine learning models only accept numeric features as input. An example of this process is the "TypeName" feature. The values of the "TypeName" feature before and after label encoding are shown in Table 4.

Table 4. Values of the "TypeName" feature before and after label encoding

Old Value	New Value
Ultrabook	0
Notebook	1
Netbook	2
Gaming	3
2 in 1 Convertible	4
Workstation	5

All changes made to the dataset features during the data preprocessing phase are shown in Table 5.

Table 5. All changes made to the dataset features during the data preprocessing phase

Feature Name	Explanation
Laptop_ID	Feature dropped.
Company	Feature dropped.
Product	Feature dropped.
TypeName	Label encoding applied.
Inches	No changes made.
ScreenResolution	Touchscreen feature was derived, and label encoding was applied to the resolution feature.
Cpu	The cpu speed feature is derived and cpus are grouped by cpu family.
Ram	Textual expressions in the column (such as “GB”) have been removed.
Memory	Memory type and memory sizes are derived.
Gpu	Gpus were grouped in a meaningful way and label encoding was applied.
OpSys	Label encoding applied.

The latest version of the dataset is shown in Table 6.

Table 6. The latest version of the dataset

Feature Name	Data Type	Explanation
TypeName	Numeric	Type (Gaming etc.)
Inches	Numeric	Screen size
ScreenResolution	Numeric	Screen Resolution
Cpu	Numeric	Cpu model
Ram	Numeric	Ram memory
Gpu	Numeric	Gpu model
OpSys	Numeric	Operating System
Weight	Numeric	Laptop Weight (Kg)
TouchScreen	Numeric	Is touchscreen?
CpuSpeed	Numeric	Cpu speed
HDD	Numeric	HDD capacity
SDD	Numeric	SDD capacity
FlashStorage	Numeric	FlashStorage capacity
Hybrid	Numeric	Hybrid capacity
Price-range	Numeric	Price range

C. Laptop Price Range Prediction Model

Four different machine learning methods were used to predict the laptop price range. These are: random forest classifier, histogram-based gradient boosting classifier, extra trees classifier and catboost classifier.

C.A. Random Forest Classifier

The random forest method, which is used to solve problems such as classification and regression, is an ensemble learning method that makes decisions using a large number of decision trees. For classification, it uses the class that the majority of decision trees provide as the decision. For regression tasks, it returns the average or mean prediction of individual trees [15].

C.B. Histogram-based Gradient Boosting Classifier

Gradient boosting using decision trees has the problem that the model is slow. To avoid this problem, continuous input variables can be divided into several hundred unique values. Gradient boosting using this solution is called histogram-based gradient boosting, i.e. Histogram-based Gradient Boosting Classifier is a combination of a gradient boosting machine and a histogram-based algorithm for the construction of the classification tree structure [16].

C.C. Extra Trees Classifier

The extra trees classifier, like the Random Forest method, uses decision trees to make decisions and interprets them by aggregating the results from the decision trees. There are two main features that distinguish it from the random forest method. One is that the random forest algorithm uses bagging to ensure sufficient differentiation between individual decision trees, whereas the extra trees classifier randomly selects the values used in the splitting of a feature and the creation of sub-nodes [17].

C.D. Catboost Classifier

The open source Catboost algorithm, developed by Yandex, is a gradient-boosting based machine learning algorithm. Two critical algorithmic advances presented in CatBoost are the implementation of sequential boosting, a permutation-based alternative compared to the classical algorithm, and an innovative algorithm for the processing of categorical features. It works with symmetric trees, which allows it to produce successful results without obtaining very deep trees. It shortens the data preparation process by performing automatic label encoding [18].

III.RESULTS

After data cleaning, feature engineering and label encoding preprocessing steps, the dataset is optimised for machine learning algorithms. Then, the data set was divided into two parts as 70% train and 30% test. Then, the classifiers were trained with the train data set and finally tested on the test data set and the success of the classifiers were calculated with performance metrics. In addition, all classifiers used in the study were used with default parameters.

The work was implemented in the JupyterLab 3.5.0 IDE using python 3.9. Successfully executed on a computer with Intel i7-6700HQ 2.60 GHz CPU and 16 GB RAM.

A histogram plot of the number of price ranges in the training and test dataset is shown in Figure 5.

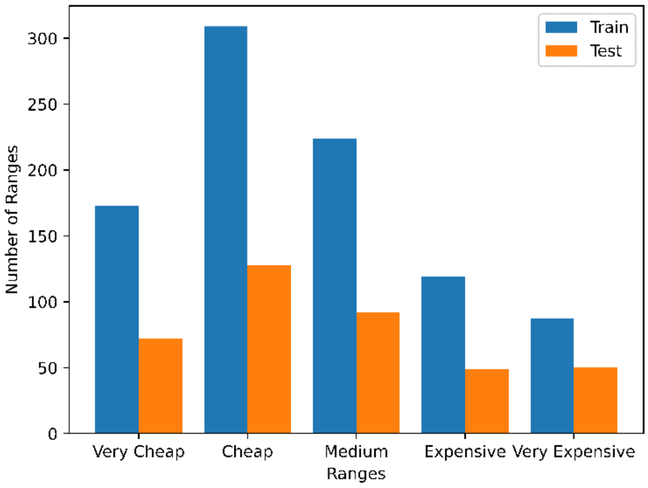

Fig. 5. A histogram plot of the number of price ranges in the training and test dataset

In this study, accuracy, precision, recall, and f1-score metrics are used to evaluate the success of the model. To be used in the formulas below; TP is the number of times the actual price range of the laptop is correctly predicted. TN is the number of predictions that ranges outside the actual price range of the laptop are not the actual price range of the laptop. FP is the number of predictions that ranges outside the actual price range of the laptop are the actual price range of the laptop. FN is the number of predictions that the actual price range of the laptop is outside the actual price range.

Accuracy (ACC) is calculated according to Equation 1. Precision (PR) is calucated according to Equation 2. Recall (RE) is calucated according to Equation 3. F1-score (F1) is calucated according to Equation 4.

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{PR} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{RE} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} = 2 * \frac{PR * RE}{PR + RE} \tag{4}$$

The confusion matrix of the random forest classifier we used in the laptop price range estimation problem is in Figure 6, the confusion matrix of the histogram-based boosting classifier is shown in Figure 7, the confusion matrix of the extra trees classifier in Figure 8, and the confusion matrix of the catboost classifier in Figure 9. The success scores of the classifiers according to the performance metrics derived from these confusion matrices are shown in Table 7.

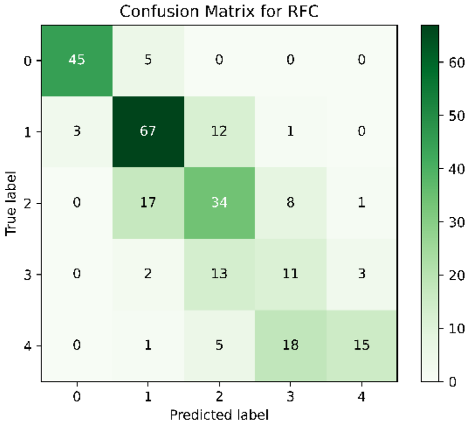

Fig. 6. The confusion matrix of the random forest classifier

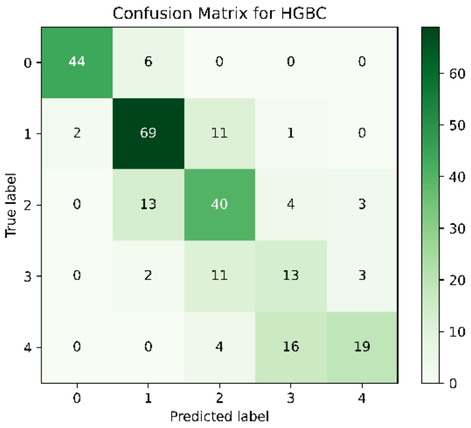

Fig. 5. The confusion matrix of the histogram-based gradient boosting classifier

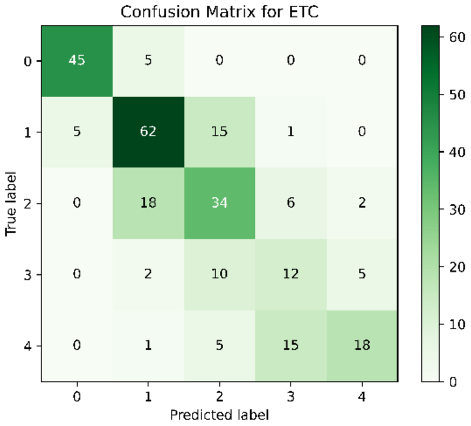

Fig. 8. The confusion matrix of the extra trees classifier

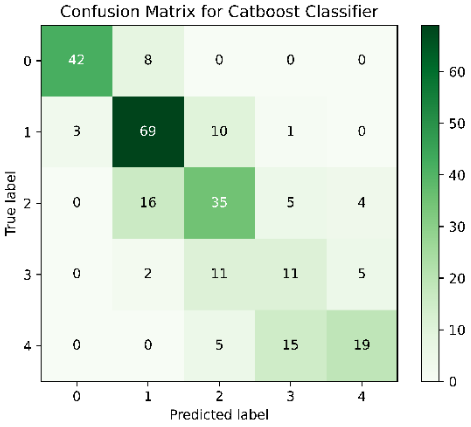

Fig. 9. The confusion matrix of the catboost classifier

Table 7. Performance measures of classifiers

Classification Algorithm	Accuracy	Precision	Recall	F1
Random Forest	0.66	0.68	0.66	0.66
Histogram-based Gradient Boosting	0.70	0.72	0.70	0.70
Extra Trees	0.68	0.69	0.68	0.68
Catboost	0.68	0.69	0.68	0.68

IV. CONCLUSION

Although there is more than one study for laptop price prediction, only one study has been conducted for laptop price range prediction, which is more practical and efficient in real world problems. In this study, the fact that the prices are divided into three different classes does not overlap much with the real-world laptop price range prediction problem. In addition, very few machine learning methods have been tested in the laptop price range prediction problem. For these reasons, there is not much information about the solution of this problem in the literature. To overcome these problems and contribute to the literature, this study has been carried out. Firstly, a dataset, which was previously used for laptop price prediction, was organized to be used in laptop price range prediction. Then, the dataset was visualized and subjected to pre-processing steps such as data cleaning, feature engineering and label encoder. After the preprocessing steps, an efficient, noise-free and suitable for training data set was obtained. After dividing the data set into two parts as training and test, four different machine learning algorithms (random forest classifier, histogram-based boosting classifier, extra trees classifier and catboost classifier) were trained with the training data set. The trained models were tested on the test dataset and their performance was measured with success criteria. The histogram-based boosting classifier gave the best result with 70% accuracy. In the future, this problem can be solved by using more than one data set, by using machine learning methods used in this study, by creating a hybrid model with optimization algorithms and classification algorithms, or by using deep learning methods.

Authors' Contributions

Author 1 has participated in the design of the study, wrote the code, performed the experiments, and drafted the manuscript. Author 2 has contributed to the manuscript and provided the initial idea, and to the design and supervision of the study. All authors have read and approved the final manuscript.

Statement of Conflicts of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The authors declare that this study complies with Research and Publication Ethics

REFERENCES

[1] "Global Laptop Market Report and Forecast 2024-2032," Laptop Market Share, Size, Trends, Growth, Analysis 2024-2032, https://www.expertmarketresearch.com/reports/laptop-market (accessed Oct. 30, 2023).

[2] Z. Youhan, "Machine learning can be divided into 3 categorizations: Supervised, unsupervised and reinforcement...," Medium, https://zinayouhan33.medium.com/machine-learning-can-be-divided-into-3-categorizations-supervised-unsupervised-and-reinforcement-9a1b47460f5d (accessed Nov. 6, 2023).

[3] C. L. Reddy, K. B. Reddy, G. R. Anil, S. N. Mohanty and A. Basit, "Laptop Price Prediction Using Real Time Data," *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, Jeddah, Saudi Arabia, 2023, pp. 1-5, doi: 10.1109/ICAISC56366.2023.10085473.

[4] A. D. Siburian *et al.*, "Laptop price prediction with machine learning using regression algorithm," *Jurnal Sistem Informasi dan Ilmu Komputer Prima(JUSIKOM PRIMA)*, vol. 6, no. 1, pp. 87–91, 2022. doi:10.34012/jurnalsisteminformasidanilmukomputer.v6i1.2850.

[5] M. A. Shaik, M. Varshith, S. SriVyshnavi, N. Sanjana and R. Sujith, "Laptop Price Prediction using Machine Learning Algorithms," *2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS)*, Nagpur, India, 2022, pp. 226-231, doi: 10.1109/ICETEMS56252.2022.10093357.

[6] A. A. Syed, Y. Heryadi, Lukas and A. Wibowo, "A Comparison of Machine Learning Classifiers on Laptop Products Classification Task," *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS)*, Hong Kong, 2021.

[7] C. Ma *et al.*, "Cost-sensitive deep forest for price prediction," *Pattern Recognition*, vol. 107, p. 107499, 2020, doi:10.1016/j.patcog.2020.107499.

[8] I. M. Nasser, M. O. Al-Shawwa and S. S. Abu-Naser, "Developing Artificial Neural Network for Predicting Mobile Phone Price Range," *International Journal of Academic Information Systems Research (IJAISR)*, vol. 3, no. 2, pp. 1-6, 2019.

[9] M. A. Rahman, M. A. Kabir, M. E. Haque, and B. M. Hossain, "Machine learning-based price prediction for cows," *American Journal of Agricultural Science, Engineering, and Technology*, vol. 5, no. 1, pp. 64–69, 2021, doi:10.54536/ajaset.v5i1.63.

[10] E. Güvenç, G. Çetin and H. Koçak, "Comparison of KNN and DNN Classifiers Performance in Predicting Mobile Phone Price Ranges", *Advances in Artificial Intelligence Research*, vol. 1, no. 1, pp. 19-28, Jan. 2021.

[11] E. Gegic, B. Isakovic, D. Keco, Z. Masetic and J. Kevric, "Car price prediction using machine learning techniques", *TEM Journal*, vol. 8, no. 1, 2019, doi:10.18421/TEM81-16.

[12] S. Yücebaş, M. Doğan ve L. Genç, "A C4.5 – Cart Decision Tree Model For Real Estate Price Prediction And The Analysis of the Underlying Features", *Konya Journal of Engineering Sciences*, vol. 10, no. 1, pp. 147-161, 2022, doi:10.36306/konjes.1013833.

[13] M. Ortu, N. Uras, C. Conversano, S. Bartolucci, and G. Destefanis, "On technical trading and social media indicators for cryptocurrency price classification through Deep Learning," *Expert Systems with Applications*, vol. 198, p. 116804, 2022. doi:10.1016/j.eswa.2022.116804.

[14] M. Varlı, "Laptop price", Kaggle, https://www.kaggle.com/datasets/muhammetvarl/laptop-price (accessed Oct. 30, 2023).

[15] Tin Kam Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

[16] H. Nhat-Duc and T. Van-Duc, "Comparison of histogram-based gradient boosting classification machine, random forest, and deep convolutional neural network for Pavement Raveling Severity Classification," *Automation in Construction*, vol. 148, p. 104767, 2023. doi:10.1016/j.autcon.2023.104767

[17] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006. doi:10.1007/s10994-006-6226-1

[18] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *arXiv (Cornell University)*, Jun. 2017, doi: 10.48550/arxiv.1706.09516.