
# Television Price Prediction Based on Features with Machine Learning

Marumoju Dheeraj[1], Manan Pathak[1], G. R. Anil[1(✉)], and Mohamed Sirajudeen Yoosuf[2]

[1] Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India
anilgrcse@gmail.com

[2] School of Computer Science and Engineering, VIT-AP University, Amaravati, India

**Abstract.** Television is both a source of information and a means of communication, and it plays an important role in everyone's life. It broadcasts news, documentaries, sporting events, and other events, among other things. In the market, different models of televisions having different features are available based on the user requirement. This paper tries to develop a model that can offer a client with a fair pricing estimate based on a tradeoff between features and price. A four-step process is devised for this objective, which includes real-time data scraping from an eCommerce website and creation of a model using machine learning algorithms. The algorithms like Multi Linear Regression, SVM (Regressor), Decision Tree Regressor are used for price prediction. Decision Tree Regression was found to be more accurate in predicting television prices in this study.

**Keywords:** Machine Learning · Price Prediction · TV Price · Web Scraping · ECommerce

## 1 Introduction

Today, Television is one of the most significant sources of communication. It presents news, presentations, sports events, and various other things. Television enlightens us about different places and cultures while also providing great entertainment in the form of series, comedies and pantomimes. Television has had a huge influence on our lives since its introduction almost 80 years ago. It is a notable way for a person to spend some leisure time, and it also influences people's opinions on issues of different domains.

A Smart Television combines the features of a television and a smartphone, or combination of a television and a computer. They come with built-in Internet access, letting users surf the web and use other online services.

Because of remarkable technological breakthroughs in television, the future of television is headed to the next level for its viewers. As the television standards evolved from analogue to digital, more channels could be transmitted over the airwaves. Digital transmissions have mostly supplanted analogue broadcasting in recent years. Smart


S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 470, pp. 507–517, 2023.
https://doi.org/10.1007/978-3-031-35078-8_42

TV"s have greater picture quality and can carry more info. Images and Videos can now be displayed in a new high-definition format on television.

When DTV first came out, manufacturers began creating HDTVs. The picture clarity of HDTV is double that of standard definition. These television sets cost around $10,000 to begin with. The price had dropped down to roughly $1,000. With the designing of the 4K and 8K Ultra High Definition (UHD) televisions, which have three to four times the picture quality of a normal HDTV, newer technology has arisen. Plasma televisions are being replaced by smart (LED), Light Emitting Diodes, television.

The internet will play a significant role in television's future. Content creation could eventually be profitable on the internet. Various OTTs (Over the Top) such as Netflix are becoming the standard for watching movies and TV shows. Television commercials will continue to be broadcast in the future. Advertisers understand that everyone uses TVs, thus making this a wonderful way to sell things to future audiences.

- In this paper, we used a regression model to forecast the price of a TV with several features. The prediction is made based on the features. This would be extremely useful for customers in predicting which TV would have the characteristics they desire, as well as for the industry in determining product costs depending on features. This not only assists customers in deciding which TV to buy, but it also assists owners in determining what the TV's appropriate pricing should be for the features it offers. This concept of price prediction will assist customers in making informed decisions when purchasing a TV in the future.
- The methods we used are machine learning, web scraping, data cleaning and prediction models which help the user to get an accurate price of a product based on feature selection. Machine learning (ML) is the learning method of computer algorithms that improve themselves over time by gaining knowledge from observations and using respective data. Machine learning algorithms prepare an exact model based on provided training data to make predictions without having to be externally programmed again, web scraping to acquire all the data from Flipkart (E-Commerce website), and data visualization to visualize the data in a visual style. The best feature selection algorithm and best classifier for the dataset are used to get a conclusion. The primary purpose of this paper is to assess whether a TV with specific features will be economical or expensive.

## 2 Literature Survey

### 2.1 Online Price Prediction

Existing literature on TV pricing prediction is covered in this section. For this prediction, a variety of machine learning methods are used. These works are based on a variety of data sources. Existing datasets from database sites such as Kaggle are fully leveraged possible. Following are their approaches:

Yu-Hsuan Cheng et al. [1] has proposed a model to predict "TV Audience Rating Based on Facebook as source" using machine learning algorithms. To forecast the recent programmer viewership rating by accumulating the word-of-mouse from televised TV shows on Facebook, they used the Back-propagation Network. They presented trend study of audience ratings, which helps to provide the relationship between Nielsen TV

ratings and predicted audience ratings. This model did not mention predicting the TV price.

Scott Seredy et al. [2] has proposed a machine learning model to predict the future TV ratings based on a variety of machine learning algorithms, like linear regression, random decision forests, support vector machines, neural networks and gradient boosting machine (GBM)3 . Each approach here has different pros and cons, where the final method which offers the best combination of scalability and accuracy for the project is the GBM method (especially, the xgboost library that is optimized). This paper aimed to develop a more accurate, more efficient, and more consistent system to improve the accuracy of the ratings.

Sebastian Elf et al. [3] has proposed how supervised machine learning can be used to forecast television ratings for advertisements placements. The supervised machine learning models studied here are Random forests algorithm and support vector Regression. Here it tries to evaluate various machine learning models to check if the task of predicting TV ratings can be self-operated with better accuracy than the manual process. They concluded that the random forest provided them with the best results.

Md. Hafizur Rahman et al. [4] has proposed an approach for "The study's key contribution is the use of LSTM-based machine learning models to predict the pricing". This research shows how Machine Learning Models with Long Short-Term Memory can be used to forecast stock prices, to create two LSTM models for comparison and prediction. These models were built using training data from these companies' stock histories from January 2019 to January 2021. The main aim of this research is to decide which version of the LSTM architecture model provided the most accurate predictions of all.

Vidhi Singrodia et al. [5] has proposed a model on "A Review on Web Scraping and its Applications". The study uncovered a slew of unrelated and unorganized data. The paper discusses a variety of online scraping methods and different features present in web scraping. The paper discussed the benefits and drawbacks of online scraping, as well as the various options for using the data, like Big Data, commercial information and the creation of latest applications and methods, to mention a few.

Rabiyatou Diouf et al. [6] has proposed a model on web Scraping: State of Art and Application fields is the study's fundamental contribution. Scraping's main goal is to collect data from many websites and convert it into easy formats like databases, spreadsheets, and CSV files. Scraping the web, on the other hand, is not only a timeand resource-intensive process, especially when done manually. Several automated systems have been developed as a result of previous research. The main aim of this analysis and study is to examine some of the current Web scraping methodologies, categories, and tools, as well as their application areas.

Carson Kai-Sang Leung et al. [7] has presented a model for stock price prediction using a machine learning model. This paper presents a machine learning technique for business intelligence applications. To recognize the complex inputs as nodes in a graph structure (SSVMS), here an architectural support vector machine is used. SSVM was utilized to anticipate either positive or negative movement in the stock values of partnering firms in the information technology industry using a graph structure. The difficulty of separation oracle helps to determine the complexity of the SSVM cutting plane optimization problem.

Cheng-Ju Liu et al. [8] presented a model on e-commerce platform repurchase customer prediction model based on machine learning. This study used the XGBoost logistic regression technique, which uses a linear model and this is also based on a decision tree. According to the paper, the nonlinear model can exploit these characteristics and produce more precise predictions. This study uses the model fusion approach. The main aim of this research is to avoid easy to fit accuracy of a linear model and overfitting of decision trees. A single model, they said, is more advanced than the XGBoost p/n sample hybrid model.

Kyosuke Morita et al. [9] had proposed a model on Analyzing online price by using machine learning techniques. The paper illustrates a method for analyzing a firm's pricing adjustment behavior. This combines two techniques, which both provide similar findings and complement one another. To find product categories with more flexible pricing choices, this research uses LASSO and elastic net to pick more informative predictors for logistic regression. By adding this knowledge into a machine learning system, they boost the algorithm's prediction potential.

Jean-Denis ZAFAR et al. [10] has proposed a model to Estimate Hedonic Models And Extensions To Other Predictive Methods based on Web Scraped Laptop Prices. In order to compare costs, the consumer price index should account for the difference in quality between the old and new products when an item goes missing and needs to be replaced. Hedonic regressions, in which product attributes are used as price explanatory factors, are used to calculate the difference. This study looks at how web scraping can be used to get larger amounts of pricing and attribute data, particularly for electronic items.

Ilya Igorevitch Raykhe et al. [11] has proposed a model on Automatic Price Prediction for eBay Online Trading in Real-Time. This paper is about a trading program that may be utilized on eBay. Here the technology forecasts the selling price of an item based on its qualities and features on eBay. The training data here is previous laptop transactions on eBay. The evolutionary algorithms are used by the system to determine feature weights in a k-Nearest Neighbor technique. Hence the time a reseller spends on the trade activities decreases because of the trained model which helps most of the market research to be done automatically.

## 3 Methodology

The suggested method's workflow is divided into four steps, as shown in Figure 1 (see Fig 1). This section goes through each phase in great depth. The first and most significant component is 3.1 Web scraping, a method for scraping data from an E- Commerce website (Flipkart) in real time so that the study may be conducted in real time and the data set assumed or not. The next part (3.2) is about data wrangling, also called as data cleaning, which is referred to as extracting relevant information present in the dataset, and the algorithms will be detailed in the final phase.

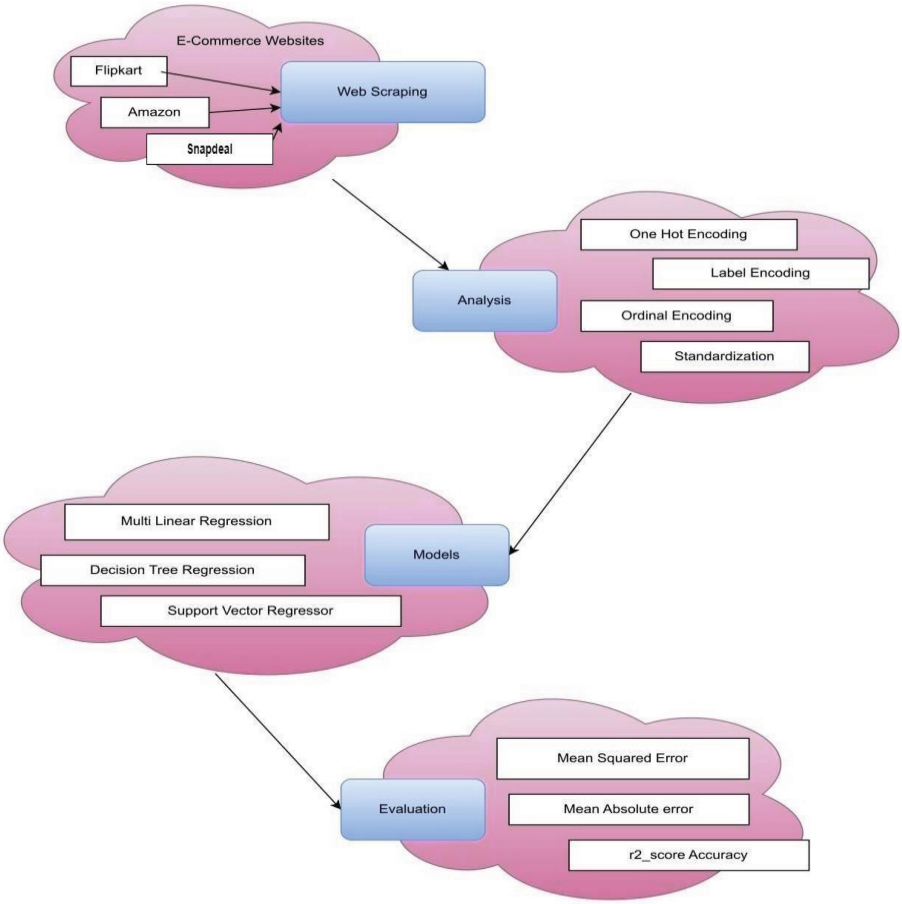


**Fig. 1.** Methodology

### 3.1 Web Scraping

For any Data scientist, engineer, or anyone who works with large amounts of data, scraping data from the web is a useful ability to have. Scraping data from a webpage is known as web scraping. The different phases in web scraping are shown in figure below in Fig. 2.

In this work, Google colab is used to Scrap data from the web. It is necessary to import numerous libraries at first, including NumPy, Pandas, BeautifulSoup, and Requests. The URL of the intended website to be scraped must then be found. We send a request to the targeted website using the requests.get() function, which returns the response object. Using the returned object, we can use content to access the web page's features. HTML-Parser is used to construct the Beautiful Soup object. With this method thus the data is extracted using this soup object and certain built-in functions.

**The Details of the Acquired Data are Described in the Following Sections**

The scraped data contains around 984 rows, which represents the entire data set displayed on the site. The research considered several factors when creating the model, including the rating, display, number of apps supported, Speaker output, number of USB and HDMI ports and Refresh rate. There are a few features that have less collinearity, but they are nevertheless examined for developing models because they must not result in a false output in real time when a person uses the method.

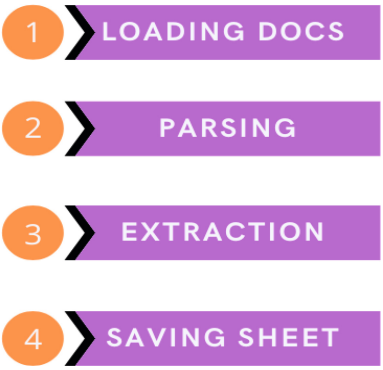


**Fig. 2.** Phases in Web Scraping

A collinearity matrix is a sort of matrix that contains a statistical measure that reflects how closely two or more variables move together.

For a collinearity matrix, as mentioned in Fig. 3, for example when given "Number of apps supported" and "Number of USB's present" then the prediction accuracy is about

0.74. Similarly, when "Number of USB's presen" and "TV ranking" is given the prediction accuracy is very low as 0.08 as it is difficult to predict the price based only on those two factors. Similarly in this way all the accuracy predictions are shown in below collinearity matrix.

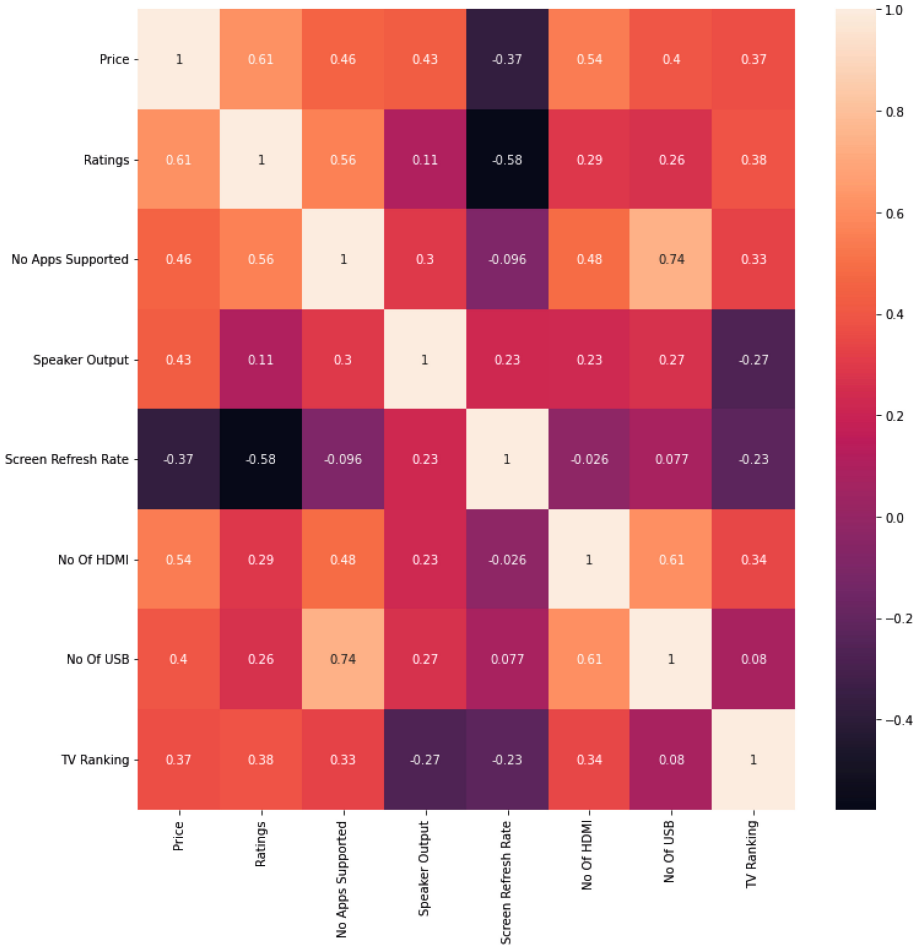


**Fig. 3.** Collinearity matrix

### 3.2 Data Wrangling

After getting the scraped data, it must be cleaned before being used to construct the model. As some TVs have additional characteristics, they must be managed properly. We have a number of options, including utilizing the dropna() function to delete rows with null values.

Structured data includes both categorical and continuous variables. Categorical variables are not recognized or dealt with by most machine learning methods. Machine learning algorithms perform better in terms of accuracy and other performance measures when input is presented to a model as a number rather than a category for training and testing. Majority of machine learning algorithms exclusively work with numerical

data. As a result, categorical characteristics must be represented in a consistent format with the models. Some examples of encoding methods are:

- Label encoding
- One-hot encoding
- Ordinal Encoding

**Label Encoding**
The process of converting labels into a numeric format that machines can read is referred to as "label encoding." The best way to use those labels can subsequently be determined by machine learning algorithms. It is an essential stage in the preprocessing of the structured dataset in supervised learning.

**One Hot Encoding**
The integer encoding is insufficient for category variables with no such ordinal relationship. Further encoding it could lead to poor performance. One-hot encoding can be used to encode the integer representation. The integer encoded variable is removed and a new binary variable is replaced with unique integer value. Because the "color" variable has three categories, three binary variables are required. In the binary variable, the color is represented by a "1" value, whereas the other colors are represented by "0" values.

If the missing data is a numerical variable, use the mean or median value to fill it in. Fill in the missing data with mode if it is a category value. Replace with other number that will not appear in the data. The fillna() function can be used to fill in the null values in a dataset.

With the above work Machine learning models are developed that are discussed in the following section.

## 4 Results (Analysis and Observation)

The following algorithms are considered for price prediction: Multi linear Regression, Decision Tree Regressor, and Support Vector Regressor:

**Mean Squared Error (MSE)**
The average of error squares between estimated and true value, is measured by the (MSE) or Mean Squared Deviation of the estimator. MSE is a risk function that relates to the expected value of the squared error loss. As it can never be negative, integers near zero are favored. The MSE is the estimator's second moment of error, accounting for the volatility and bias of the estimator. The Decision Tree Regressor is picked over the Multilinear Regression, which has an error rate of 0.3061, and the Support Vector Regressor, which has an error rate of 0.1409. The Decision Tree Regressor has an error rate of around 0.0998. The corresponding bar plot regarding mean square error occurred when different algorithms used is shown in figure below (see Fig. 4).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

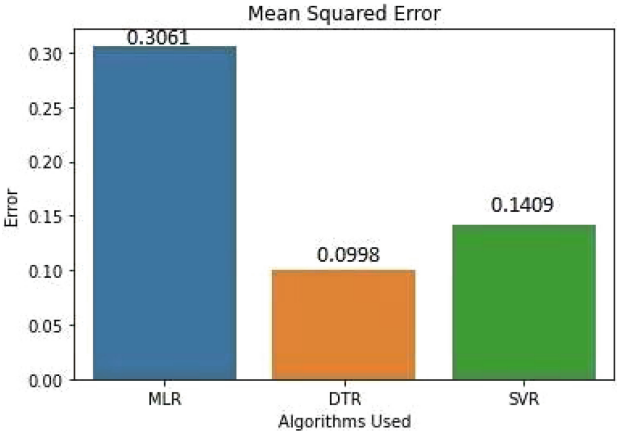


**Fig. 4.** Mean Squared Error

**Mean Absolute Error (MAE)**

Mean Absolute Error is used to calculate the average difference between calculated and real data. It calculates inaccuracy in observations obtained on the same scale. It calculates the discrepancies between real and projected values using a mathematical model. MAE is used to predict accuracy of machine learning models. Here, the mean absolute error occurred when various algorithms used is represented in below figure (see Fig. 5). The Decision Tree Regressor, with an error rate of roughly 0.2128, is chosen above the Multilinear Regression, which has an error rate of 0.4777, and the Support Vector Regressor, which has an error rate of 0.2607.

$$\text{Mean Absolute Error } = (1/n) * \sum |yi - xi|$$

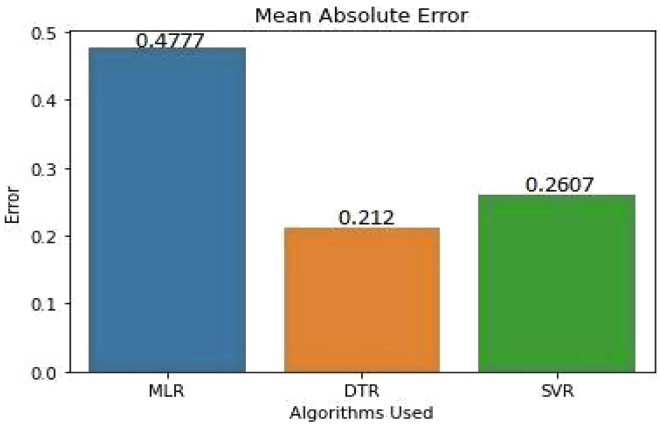


**Fig. 5.** Mean Absolute Error

**R2_score**

The R2 score, also referred to as the coefficient of determination. A linear regression model's performance is evaluated using the R2 score. The amount of variation in the output dependent feature that the input independent variable can predict. The ratio of total deviation of results described by the model is used to determine how well the model reproduces observed data.

$$R2 = 1 - s\_sres/s\_stot$$

where,
s_sres is the sum of the residual errors' squares.
s_stot is the sum of all errors.

Here as a result, the accuracy obtained from Decision Tree regressor is 0.9024, from Multi Linear Regression is about 0.7007 and from Support Vector Regressor about 0.8622. Bar plot representation of the accuracy of the algorithms performed is shown in below figure (see Fig. 6). Thus, Decision Tree Regressor is chosen above from others.

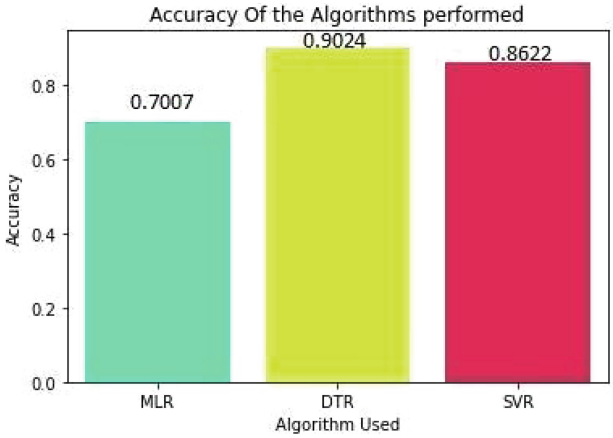


**Fig. 6.** r2_accuracy

## 5 Conclusion and Future Work

The accuracy of the model was tested with Linear Regression, Decision Tree Regression and Support Vector Regressor (SVR). The possibility to utilize a nonlinear kernel is the most exciting part of SVR. This will conclude fitting a curve rather than line in this case, which is known as non-linear regression. This approach employs the kernel trick, with the solution/model being represented in the dual rather than the fundamental. Because Linear Regression performs effectively when there is only one input feature, Decision Tree Regression achieved better result compared to other algorithms.

In future work, prediction and suggesting can be extended with a dedicated web application using real time dynamic scraping. The dataset can be segregated based on brand and experimented with remaining Regression algorithms.

## References


1. Cheng, Y.-H., Wu, C.-M., Ku, T., Chen, G.-D.: A predicting model of TV audience rating based on the facebook. In: 2013 International Conference on Social Computing, pp. 1034–1037 (2013). https://doi.org/10.1109/SocialCom.2013.167
2. Scott, S.: Using machine learning to predict future tv ratings (2017). https://www.nielsen.com/wp-content/uploads/sites/3/2019/04/using-machine- learning-to-predict-future-tv-ratings.pdf
3. Elf, S.: Comparison of supervised machine learning models for predicting TV-ratings. StockHolm (2020). http://www.diva-portal.org/smash/get/diva2:1451634/FULLTEXT01.pdf
4. Rahman, M.H., Nahid, S.I., Al Fahad, I.H., Nahid, F.M., Khan, M.M.: Price prediction using LSTM based machine learning models. In: 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 0453–0459 (2021). https://doi.org/10.1109/IEMCON53756.2021.9623120
5. Singrodia, V., Mitra, A., Paul, S.: A review on web scraping and its applications. In: 2019 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6 (2019). https://doi.org/10.1109/ICCCI.2019.8821809
6. Diouf, R., Sarr, E.N., Sall, O., Birregah, B., Bousso, M., Mbaye, S.N.: Web scraping: state-of-the-art and areas of application. In: 2019 IEEE International Conference on Big Data (Big Data), pp. 6040–6042 (2019). https://doi.org/10.1109/BigData47090.2019.9005594
7. Leung, C., MacKinnon, R., Wang, Y.: A machine learning approach for stock price prediction. In: ACM International Conference Proceeding Series (2014). https://doi.org/10.1145/2628194.2628211
8. Liu, C.-J., Huang, T.-S., Ho, P.-T., Huang, J.C., Hsieh, C.T.: Machine learning-based e-commerce platform repurchase customer prediction model. PloS One **15**(12), e0243105 (2020). https://doi.org/10.1371/journal.pone.0243105
9. Morita, K.: Analyzing online price by using machine learning techniques (2018). https://doi.org/10.13140/RG.2.2.12552.11522
10. Zafar, J.-D.: Web scraping laptop prices to estimate hedonic models. and extensions to other predictive methods. French (2019). https://eventos.fgv.br/
11. Raykhel, I., Ventura, D.: Real-time automatic price prediction for ebay online trading. In: IAAI (2009)
12. Jabbar, A., Samreen, M.S., Aluvalu, R.: The future of health care: machine learning. Int. J. Eng. Technol. **7**(4.6), 23–25 (2018)
13. Maheswari, U.V., Aluvalu, R., Chennam, K.K.: Chapter 5 application of machine learning algorithms for facial expression analysis. In: Hiran, K.K., Khazanchi, D., Vyas, A.K., Padmanaban, S. (eds.) Machine Learning for Sustainable Development, pp. 77–96. De Gruyter, Berlin (2021). https://doi.org/10.1515/9783110702514-005
14. Anil, G.R., Moiz, S.A.: Personalized dynamic learning plan generator for smart learning environments. Int. J. Recent Technol. Eng. **8**(2), 6175–6180 (2019)
15. Aluvalu, R., Jabbar, M.A., Kantaria, J.: Performance evaluation of clustering algorithms for dynamic VM allocation in cloud computing. In: 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), pp. 1560–1563 (2017). https://doi.org/10.1109/SmartTechCon.2017.8358627
16. Saatvik, A., et al.: VCE Mini Tool Kit – A Smart Approach ForImage Conversion. IRJET (2022). https://www.irjet.net/archives/V9/i2/IRJET-V9I2147.pdf
17. Reddy, K.B., et al.: VMEG Mini Tool Kit – An Intelligent Approach For File Conversion. IJIRT (2022). https://ijirt.org/Article?manuscript=154032
18. Pal'ová, D., Vejačka, M.: The main issues of the education process during the COVID19 Pandemic at the university education. In: 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), pp. 601–606 (2021). https://doi.org/10.23919/MIPRO52101.2021.9596824