Classification of Mobile Price Using Machine Learning

Nisha Sunariya[1], Avinash Singh[1], Mehtab Alam[1,*] and Vibha Gaur[1]

[1] Department of Computer Science, Acharya Narender Dev College, University of Delhi, New Delhi-110019, India

Abstract

It's critical to comprehend predicted and forecasted prices to develop a successful consumer strategy. The market performance of a product depends on proper pricing. The goal of this research is to determine a pricing range for mobile phones based on specifications including storage, display, battery life, RAM, camera, and more. It would assist consumers in making informed decisions when buying a phone that suits their needs and budget. Making the best choices might be difficult with so many resources at hand. A model that offers guidance using important aspects of mobile phones was developed to deal with this problem. To classify and estimate the price range of a mobile phone, this study maneuvers five machine learning (ML) techniques: Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and K-nearest neighbors (KNN). The models are trained to create outcomes that fall into the low, medium, high, or extremely high categories. The data for this paper was obtained from Kaggle.com. The findings are assessed to achieve the highest level of precision while choosing the most desired features of mobile phones. The findings of this research will have practical implications for both consumers and manufacturers. Consumers can make informed decisions based on the identified influential features, considering their preferences and budget constraints. Manufacturers can use the insights to optimize product offerings, emphasizing features that contribute significantly to higher price ranges. This strategic alignment can enhance market competitiveness and consumer satisfaction. This paper also identifies the best option with the most features of mobiles at the lowest price.

Keywords

Support Vector Machine (SVM), Mobile Price, Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and K-nearest neighbors (KNN)

1. Introduction

Pricing is the most beneficial characteristic in business and marketing. A decision regarding pricing regulations has significant effects on management. It establishes the profit margin on products and is one of the first assessments made by many purchasers. Before making a purchase, consumers are indeed concerned about whether they can afford the item and want to verify the price. The success of a product can be affected by a variety of elements, including pricing, product appropriateness, return rates, and profitability [1]. This study takes the first step towards achieving this objective. The main aim of this paper is to identify the most reliable and appropriate ML classification model for the classification of mobile phone prices. The motivation for this paper stems from the challenges faced by individuals unfamiliar with machine learning while purchasing a mobile phone. It's often difficult for them to discern the crucial features influencing the phone's price. Instead of predicting the exact price, the focus is on establishing a price range that reflects the overall pricing level and identifies the dominant features affecting mobile phone prices. This research adds value to discussions on pricing strategies, consumer decision-making, and the application of machine learning in predicting product prices, offering insights applicable across diverse industries.

Machine learning is a pathway to Artificial Intelligence (AI) [2]. The most recent AI technologies, such as classification, regressions, and supervised and unsupervised learning, are accessible through machine learning [3]. Data analysis and visualization can be aided by a variety of ML tools, such as MATLAB, Python, Cygwin, and others. The categorization of data using ML algorithms is very likely to yield accurate results [4].

A mobile phone, sometimes called a cell phone, portable phone, or phone, uses radio frequency links to place and receive calls [5]. Human's daily lives have become completely dependent on our mobile devices, which keep us linked and even come in handy in crises. With smartphones currently outpacing older mobile devices in usage, mobile phones have grown to be one of the most widely used consumer products ever created [6]. New mobile phone models with improved features are released every year. A mobile price class prediction model is essential for making the optimum product choice. Additionally, this model can be used to examine pre-owned cars, generators, gold, food, medicines, residences, and many other items.

Several aspects are important to consider when estimating a mobile device's price. These consist of the device's CPU type, battery life, and capability to set reminders for significant occasions. The device's dimensions and weight are frequently crucial factors for users. Several criteria, such as the amount of internal memory, the quality of the touch screen, the pixel size, and the amount of RAM, affect a mobile device's pricing [7]. This study divides mobile devices into four price ranges based on a variety of features and specifications: low, medium, high, and very high. These price ranges help in consumer decision-making, competitive pricing, and budget planning. Price ranges can be used as indicators of economic conditions.

This paper is divided into multiple sections. Section 2 provides information on the background work related to this work, while Section 3 briefly defines various prediction models used in the study. The methodology approach and the findings of experimental

prediction are presented in Section 4. The conclusion and future directions are provided in Section 5.

2. Background Work

This section describes the findings on projecting and estimating the cost of various goods. Sameer Chand-Pudaruth projects the price of used cars in Mauritius. This study discovered that Nave Bayes and DT are ineffective at managing, categorizing, and forecasting numerical values because there were fewer occurrences and incredibly low prediction accuracy was reported [8]. M. Asim and Z. Khan [9] forecasted the price of mobile phones. They strived to have the most accurate predictions while maintaining the lowest cost and highest feature model. Using the DT, 78% accuracy was obtained.

Due to the lack of characteristics and algorithms, extremely low prediction accuracy was observed. Menghan Chen [10] predicted the prices of smartphones with fewer features. Principal Component Analysis (PCA) and Pearson's correlation were used as two feature reduction techniques. Without using any feature reduction approaches, Multi-Layer Perceptron (MLP) had a 92.84% accuracy rate. However, accuracy suffered as features were reduced, falling to 93.22% for the top 15 and 34.06% for the top 5.

K Noor and Sadaqat J [11] predicted the automobile prices using multiple linear regression. They projected prices from independent variables such as the vehicle's model, make, city, version, color, mileage, alloy rims, and power steering. Kuo-

Kun Tseng et al. [12] worked on foretelling e-commerce goods prices using online sentiment analysis. They developed a price prediction algorithm after analyzing news that had an impact on product prices. Aidin Zehtab-Salmasi et al. [13] developed a Multimodal Price Prediction for mobile phone pricing based on its specifications.

Neural networks are more accurate in estimating a house's price, according to Limsombunchai's research [14]. This study offered strong support for prediction superiority without comparing the forecasting abilities of the of hedonic price model and Neural Networks. A smartphone app for stock prediction was created by Abidatul Izzah et al. utilizing enhanced multiple linear regression [15]. Their mobile app's accuracy forecast outperformed the traditional method.

Al-Dhuraibi et al. predicted the price of gold. They predicted whether the price of gold would rise or go down in the future with the help of various ML models. They found that only the K-NN algorithm had an acceptable performance with an accuracy of 60.26% [16]. Mohapatra et. al. predicted the possibility of having breast cancer in a woman using various ML algorithms. They achieved the highest accuracy of 98.7 % using the XGBoost model [17].

While existing studies have explored mobile price prediction using machine learning, a notable gap persists in addressing the needs of non-expert consumers struggling to discern crucial features influencing mobile phone prices. Most research has focused on predicting exact prices, neglecting the practical challenges faced by consumers unfamiliar with machine learning intricacies. Our study uniquely addresses this gap by concentrating on establishing a price range rather than exact figures, providing consumers with a more accessible understanding of pricing levels. Additionally, we aim to identify and highlight the key features influencing mobile prices, offering a user-friendly perspective. By doing so, our

research contributes to making mobile price prediction more transparent and consumer-centric.

3. Prediction Models

This section provides an overview of the prediction models used in this study to predict the pricing of mobile phones depending on their features. A basic explanation of the models is given below.

3.1. Decision Tree

It is a tree-like model where an input is processed through a series of decisions based on features, leading to a predicted output. Decision trees are easy to understand and interpret, making them particularly valuable in various applications [18]. The decision tree makes predictions by asking a series of questions about the input features and eventually reaching a leaf node that provides the predicted outcome. Decision trees find applications in various fields, aiding in classification and regression tasks [19].

3.2. Logistic Regression

Logistic regression, also referred to as the logit model, is a statistical technique used to determine the probability of an event occurring based on a group of independent variables. This method is particularly helpful for determining the correlation between the target variable and one or more other variables. Logistic regression is often employed when dealing with categorical dependent variables. However, the model may be vulnerable to overfitting when numerous predictor variables are present [20].

3.3. K-Nearest Neighbour (KNN)

The KNN algorithm is considered non-parametric due to its lack of assumptions regarding underlying data. Instead, it relies on the similarity between existing and new data to categorize new cases. During the training phase, the algorithm simply stores the available data and classifies new data or cases based on a similarity measure. The classification of data points is based on how their neighbours are classified [21].For each proceedings volume published with CEUR-WS, the titles of its papers should either all use the emphasizing capitalized style or the regular English (or native language) style. Check with the editors of your volume which style you should adopt.

3.4. Random Forest

This classification system uses multiple decision trees on different data sets to improve prediction accuracy. This classifier uses bagging, which involves training many models using distinct subsets of data, as opposed to depending just on a single decision tree [22]. The outcome is determined by combining the results of all the models and using the majority vote approach [23].

3.5. Support Vector Machine

The SVM model is useful for solving both classification and regression problems [24]. It is widely used in tasks involving machine learning classification. The major goal of this strategy is to locate the best decision boundary for a group of points that belong to the same class [25].

4. Experiment

This section outlines the proposed procedure used in the experiment carried out for the investigation. Figure 1 describes and illustrates the essential phases of the procedure.
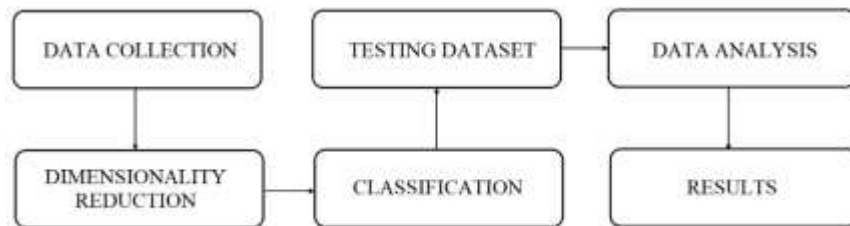

Figure 1. Outline of the Experiment.

4.1. Data Collection

The data set for this paper was obtained from Kaggle.com [26], and it includes details like battery life, CPU speed, weight, RAM, and other factors of mobile phones. The dataset comprised 2000 instances and 24 attributes of mobile phones. A sample of the distinct values from the used data set is shown in Figure 2. The characteristics used in the proposed study are explained below.

id	battery_power	blue	clock_speed	dual_sim	fc	four_g
1	1043	1	1.8	1	14	0
2	841	1	0.5	1	4	1
3	1807	1	2.8	0	1	0
4	1546	0	0.5	1	18	1
5	1434	0	1.4	0	11	1
6	1464	1	2.9	1	5	1
7	1718	0	2.4	0	1	0
8	833	0	2.4	1	0	0
9	1111	1	2.9	1	9	1
10	1520	0	0.5	0	1	0

Figure 2. Overview of data set with well-defined values

    i.   Battery Power: The battery's power output directly affects how long it can be used. A battery with a higher capacity will be able to hold more energy and function for a longer period of time. It is expressed as mAh.

    ii.   Clock Speed: The number of cycles completed by a CPU in a second defines clock speed.

    iii.   Dual Sim: It permits the use of two unique sims in the same device.

iv.   Four_g: It defines the generation of mobile network connectivity.

v.   Internal Memory: The amount of data storage that is available on the phone's drive. It is measured in gigabytes.

vi.   Front Camera (FC): It indicates whether or not the smartphone has a front camera. The resolution of FC is measured in megapixels.

vii.   Bluetooth (blue): It indicates whether or not the mobile phone has the Bluetooth feature.

viii.   Mobile weight: It stands for the weight of cell phones. Nowadays, consumers prefer using lighter phones.

ix.   Mobile depth: It reflects the thickness of a mobile phone in millimeters.

4.2. Dimensionality Reduction

It gets harder to construct a training set and use it efficiently as the number of features/attributes rises. In the initial stage of dimensionality reduction, attributes with missing values were examined and their average or mean was substituted by eliminating or including rows. The dataset initially contained 2000 rows across 21 columns, but after pre-processing, it was discovered that the attribute "mobile depth" cannot have a value lower than 0.6mm, thus it was removed. Additionally, two entries had "pixel height" values of 0, which was unacceptable. To choose the most relevant feature from the initial collection, the feature selection approach was used. This involved removing any unimportant, irrelevant, or distracting information [27]. In addition to model accuracy, understanding the importance of individual features in predicting mobile phone prices is crucial. Feature importance analysis provides insights into which characteristics significantly contribute to the pricing model. This analysis can guide manufacturers and consumers in recognizing key attributes that influence the cost of a mobile device. Therefore, the final dataset was reduced to 1998 rows and 20 columns and divided the data into two parts i.e. training data and test data.
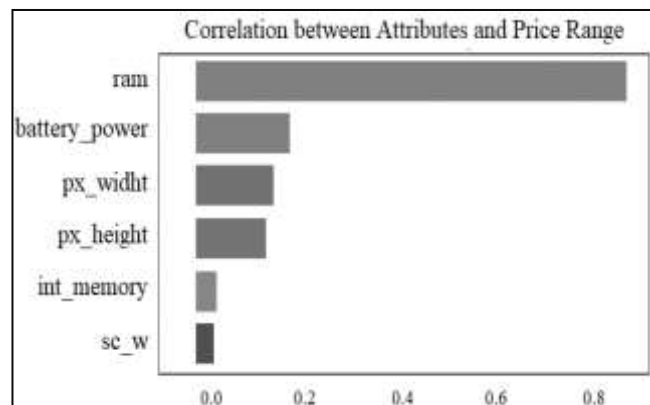

Figure 3. Correlation between attributes and price range

Furthermore, the correlation was chosen because it makes it clear how variables relate to one another, making it straightforward to forecast one variable using data from another. In the realm of mobile phones, it is often observed that there exists a strong correlation between features. By utilizing these features as input data, it is possible to draw inferences

regarding the target variable. It will be easy to forecast one variable using data from another if there is a strong correlation between the variables [28]. The correlation between camera specifications and price range is shown in Figure 3. and emphasizes the importance consumers place on mobile phone cameras. Higher megapixel counts and advanced features contribute to higher pricing, reflecting the growing significance of photography in consumer choices.

In this paper, the highest correlation was found between the following attributes:

1. pc and fc, which represent the primary and front camera in pixels respectively.
2. 3G and 4G, which represent the generation of the mobile phone respectively,
3. px width and px height, which represent the pixel width and height respectively.

4.3. Classification

Classification is a machine learning technique that is frequently applied in the context of supervised learning. It involves leveraging labeled training data to distinguish unique values and classify new observations into specific categories. It classifies unknown items according to what it has learned from the dataset and assigns them to particular classes. A dataset with labels is necessary for classification, and both the labeled training set and the corresponding test set are used for testing. To establish the relationship between actual and expected values, accuracy score, precision, recall, and F1-score were employed in the study and are explained below. Precision, recall, and F1-score provide a more detailed evaluation of model performance, especially in multi-class classification scenarios. Precision measures the accuracy of positive class predictions, recall measures the ability to capture positive instances, and F1-score balances both precision and recall.

Accuracy is a metric that measures the correctness of the model's predictions compared to the actual outcomes. It is often used in classification problems, where the goal is to assign a label to each instance from a set of predefined labels.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

TP - true positives
FN - false negatives
TN - true negatives
FP - false positives.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{Precision + Recall}$$

These metrics are utilized to compare the performance of the machine learning models employed in the study. The SVM model achieves highest precision, recall, and F1-score among the ML algorithms studied in this work reinforcing its effectiveness for prediction.

4.4. Data Analysis

The analysis of the findings is presented in this section.

4.4.1. Distribution of battery power by price range

Figure 4 illustrates how battery life affects a mobile phone's pricing range. The price of the mobile phone is represented by the Y-axis, while battery life is displayed on the X-axis. With the use of the distribution indicated, mobile phones can be divided into low, medium, high, and extremely high categories. Higher battery capacity often leads to higher pricing, aligning with consumer expectations for longer-lasting devices.

4.4.2. Distribution of clock speed by price range

The relationship between a mobile device's price with clock speed is depicted in Figure 4. X-axis shows the price while clock speed is displayed on the Y-axis. It facilitates the classification of the test data set into low, medium, high, and extremely high categories that are provided. Faster processors generally contribute to higher-priced smartphones, catering to consumers seeking high-performance devices. Similarly, the price range of Internal memory (measured in gigabytes) is also computed as it which plays a significant role in pricing. Devices with larger storage capacities are positioned in higher price ranges, addressing the demand for increased data storage.
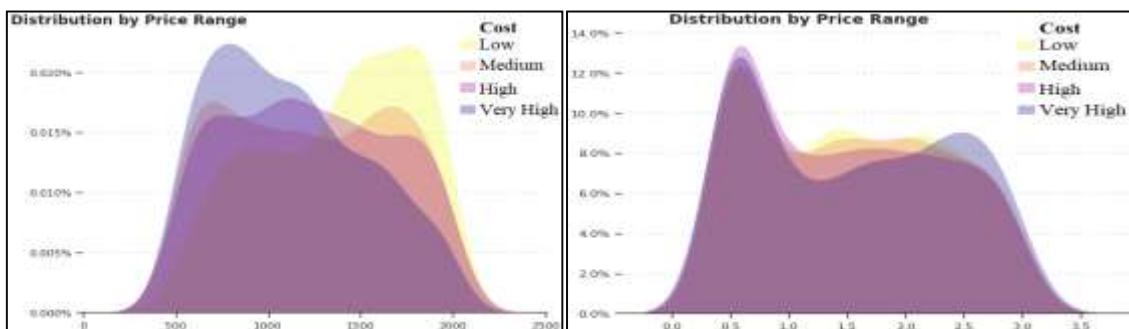

Figure 4. Matrix for battery power by price range and clock speed and price range

4.5. Results

This research work obtained substantial accuracy rates and supplied pertinent confusion matrices for reference after thoroughly examining numerous machine-learning models. The outcomes have been compared and evaluated with great attention. The distribution of attributes in this article is divided into two types depending on categorical or numerical values. The results 0, 1, 2 and 3 show the range of prices for mobile phones:
where 0 denotes low-range mobile phones,
1 denotes medium-range mobile phones,
2 denotes high-range mobile phones, and
3 denotes very high-range mobile phones, respectively.

The price range for mobile devices is shown in Figure 5. Price prediction falls into the low, medium, high, and very high categories when considering all of the features of mobile devices.
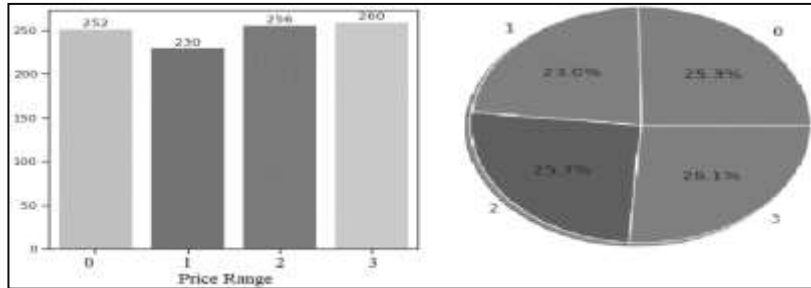
Figure 5. Price range of Mobile Phones

The SVM model works best for classification problems. SVM may incorrectly classify some examples in the training set, but it aims to create a model that is sufficiently generic to provide accurate predictions for new data. Based on how well the model performed on the test dataset, its accuracy was determined. The data needed to be trained in order to construct the RF. This model learns from training data and uses more training data than testing data. In contrast, testing data was utilized to compare the trained model to the predicted dataset. The RF model constructs numerous trees on various sub-samples, choosing the best feature from a random group of features. It uses the average to increase the prediction accuracy and reduce overfitting. The DT produces a set of rules from the given set of labeled data that are further used to classify the data. The accuracy of the SVM model was obtained as 98 percent, whereas the accuracy of the RF model was 88.8 percent. The decision tree's accuracy was 80.5 percent, compared to 82.6 percent for K-NN and 85.5 percent for LR. The accuracy, precision, recall, and F1-score for each of the five strategies are presented in Table 1. The accuracy scores of the five strategies examined in the paper are compared in Figure 6. The graph used to display learning progress is referred to as the learning curve. Learning curves illustrate the performance of models concerning the size of the training dataset. Examining learning curves helps identify underfitting or overfitting issues and provides insights into the model's stability and generalization capabilities.
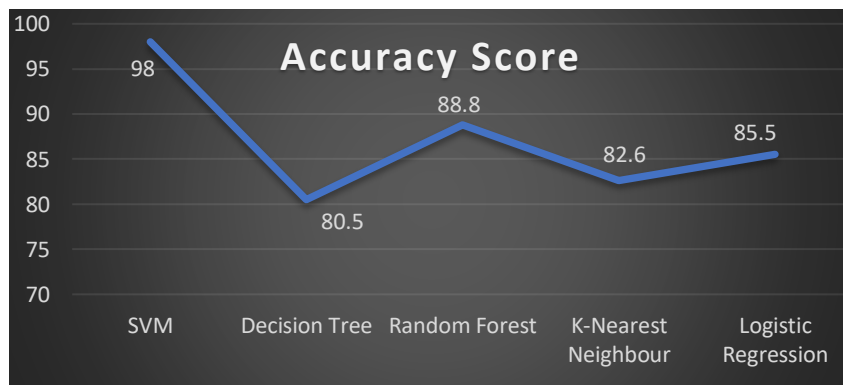

Figure 6. Accuracy graph representation of models

Table 1: Comparison of Accuracy, precision, recall and F1-score of the 5 ML techniques

ML Technique	Accuracy	Precision	Recall	F1-Score

SVM	98	0.981	0.98	0.982
Decision Tree	80.5	0.804	0.805	0.801
Random Forest	88.8	0.899	0.896	0.89
K-Nearest Neighbour	82.6	0.827	0.821	0.828
Logistic Regression	85.5	0.854	0.859	0.852

The learning curves for all five ML techniques are shown in Figure 7. It shows consistent improvement with increasing data size, indicating that the models ben-efit from larger datasets. SVM maintains a consistently high level of performance throughout, indicating its robustness.
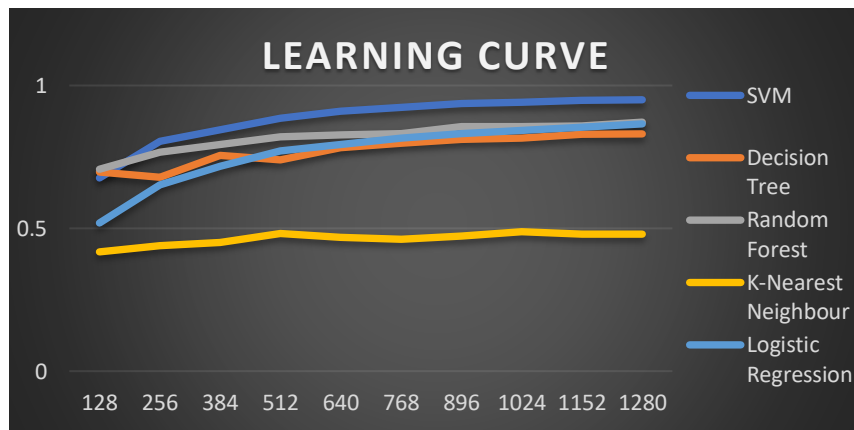

Figure 7. Learning curve for the five ML techniques

5. CONCLUSION

The primary component of any marketing strategy is cost forecasting. Finding the right solution with the best specifications at the lowest price is the best marketing strategy. Products can be evaluated based on the needs, brand, and other aspects. Data mining and analysis are the most effective ways to specify the price range recommendations of premium goods to a customer. The comprehensive analysis of ML models for mobile price classification, along with feature importance and performance metrics, provides valuable insights into the dynamics of mobile phone pricing. In this work, a variety of models were trained using mobile features, and a considerable prediction of the range of mobile prices was made. With a 98 percent accuracy rate, the SVM model was shown to be the most accurate. The proposed work may be used to anticipate costs for many things, including vintage automobiles, healthcare products, homes, etc. A premium product might be suggested by specifying the price range that the customer can afford.

Future work could explore additional features or refine existing ones to improve model performance further. The dataset's size and diversity play a crucial role in model training. Future research may benefit from larger datasets encompassing a wider range of mobile devices, manufacturers, and geographic regions. Addition-ally, more sophisticated AI algorithms may be used to forecast a product's actual pricing. This study can also be

extended with the implementation of a decision matrix and performance scores may be calculated for assigning ranks to the mo-bile devices. A list of mobile devices within the specified price range and with the desired features will assist consumers in making decisions.

References

[1] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain and A. J. Aljaaf, "A System-atic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science," in *Supervised and Unsupervised Learning for Data Science, Springer, Cham*, 2020.

[2] S. K. Das, "Introduction to Mobile Terminals", *Mobile Terminal Receiver Design: LTE and LTE-Advanced*, Wiley Telecom, 2017, pp. 1-8.

[3] I. Sim, "Mobile Devices and Health", *New England Journal of Medicine*, vol. 381, no. 10, pp. 959-968, 2019.

[4] K. Mayuri, "Importance of Pricing," [Online]. Available: https://www.economicsdiscussion.net/marketing-management/pricing/importance-of-pricing/31838.

[5] M. Alam and I. R. Khan, "Application of AI in smart cities," in *Industrial Transfor-mation, Taylor & Francis Group*, 2022, pp. 61-86.

[6] C. Janiesch, P. Zschech and K. Heinrich, "Machine learning and deep learning," *Electron Markets*, vol. 31, pp. 685-695, 2021.

[7] M. Alam, E. R. Khan, A. Alam, F. Siddiqui and S. Tanweer, "The DIABACARE CLOUD: predicting diabetes using machine learning," *Acta Scientiarum Technology*, vol. 46, no. 1, 2023.

[8] S. Pudaruth, "Predicting the Price of Used Cars using Machine Learning Tech-niques," *International Journal of Information & Computation Technology*, vol. 4, no. 7, 2014.

[9] M. Asim and Z. Khan, "Mobile Price Class prediction using Machine Learning Techniques," *International Journal of Computer Applications*, vol. 179, no. 29, pp. 6-11, 2018.

[10] M. Chen, "Mobile Phone Price Prediction with Feature Reduction," *Highlights in Science, Engineering and Technology*, vol. 34, pp. 155-162, 2022.

[11] K. Noor and S. Jan, "Vehicle price prediction system using machine learning tech-niques," *International Journal of Computer Applications*, vol. 167, no. 9, pp. 27-31, 2017.

[12] K.-K. Tseng, R. F.-Y. Lin, H. Zhou, K. J. Kurniajaya and Q. Li, "Price prediction of e-commerce products through internet sentiment analysis," *Electronic Commerce Research*, vol. 18, no. 1, pp. 65-88, 2017.

[13] A. Zehtab-Salmasi, A.-R. Feizi-Derakhshi, N. Nikzad-Khasmakhi, M. Asgari-Chenaghlu and S. Nabipour, "Multimodal Price prediction," *Annals of Data Science*, vol. 10, no. 3, pp. 619-635, 2021.

[14] V. Limsombunc, C. Gan and M. Lee, "House price prediction: Hedonic price model vs. Artificial Neural Network," *American Journal of Applied Sciences*, vol. 1, no. 3, pp. 193-201, 2004.

[15] A. Izzah, Y. A. Sari, R. Widyastuti and T. A. Cinderatama, "Mobile app for stock prediction using Improved Multiple Linear Regression", *International Conference on Sustaina-ble Information Engineering and Technology (SIET),* Malang, Indonesia, 2017.

[16] W. A. Al-Dhuraibi and J. Ali, "Using classification techniques to predict gold price movement,", *4th International Conference on Computer & Technology Applications,* Istanbul, Turkey, 2018.

[17] S. K. Mohapatra, A. Jain, Anshika and P. Sahu, "Comparative Approaches by using Machine Learning Algorithms in Breast Cancer Prediction," in *2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE),* Greater Noida, 2022.

[18] B. Charbuty and A. Abdulazeez, "Classification based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20-28, 2021.

[19] F.-J. Yang, "An Extended Idea about Decision Trees," in *International Conference on Computational Science and Computational Intelligence (CSCI),* Las Vegas, NV, USA, 2019.

[20] C. G. Raju, V. Amudha and S. G, "Comparison of Linear Regression and Logistic Regression Algorithms for Ground Water Level Detection with Improved Accuracy," in *Eighth International Conference on Science Technology Engineering and Mathematics*, Chennai, In-dia, 2023.

[21] M. Zong, X. Zhu and D. Cheng, "Learning k for kNN Classification," *ACM Transac-tions on Intelligent Systems and Technology*, Volume 8, vol. 8, no. 3, pp. 1-19, 2017.

[22] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal*, vol. 20, no. 1, pp. 3-29, 2020.

[23] A. Sekulic, M. Kilibarda, G. B. M. Heuvelink, M. Nikolic and B. Bajat, "Random Forest Spatial Interpolation," *Remote Sensing*, vol. 12, no. 10, p. 1687, 2020.

[24] S. Y. Chaganti, I. Nanda, K. R. Oandi, T. Prudvith and N. Kumar, "Image Classifica-tion using SVM and CNN", *International Conference on Computer Science, Engineering and Applications*, Gunupur, India, 2020.

[25] J. Cervantes, F. Garcia-Lamont, L. Rodriguez-Mazahua and A. Lopez, "A compre-hensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189-215, 2020.

[26] A. Sharma, "Mobile Price Classification," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification.

[27] A. Yaicharoen, K. Hashikura, M. A. S. Kamal and I. Murakami, "Effects of Dimen-sionality Reduction on Classifier Training Time and Quality," *3rd International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP)*, Bangkok, Thai-land, 2023.

[28] R. Han, Rodriguez-Mayorga and S. Luber, "A Machine Learning Approach for MP2 Correlation Energies and Its Application to Organic Compounds," *Journal of Chemical Theory and Computation*, vol. 17, no. 2, pp. 777-790, 2021.