CLASSIFICATION OF MOBILE PHONES USING MACHINE LEARNING: EVALUATING ACCURACY AMONG ALGORITHMS

[1]Dr. T. Shesagiri, [2]M. Srikanth Reddy, [3]Jitree Shruthi, [4]K. Pavan Kumar, [5]R. Jayanth Reddy

[1]Associate Professor, Department of IT, Joginpally B.R Engineering College, Moinabad, Telangana, India

[2]Assistant Professor, Department of IT, Joginpally B.R Engineering College, Moinabad, Telangana, India

[3]Student, Department of IT, Joginpally B.R Engineering College,Moinabad, Telangana, India

[4]Student, Department of IT, Joginpally B.R Engineering College,Moinabad, Telangana, India

[5]Student, Department of IT, Joginpally B.R Engineering College,Moinabad, Telangana, India

Abstract

Pricing plays a crucial role in determining the market performance of mobile phones. This study aims to establish an optimal price range for mobile phones based on key specifications such as capacity, display, quality, storage, battery life, camera performance, and other essential features. By leveraging machine learning techniques, this research classifies mobile phone prices into distinct categories—low, medium, high, and premium—based on specifications, brand, and performance metrics. The machine learning models such as Logistic Regression(LR), Random Forest(RF), Support Vector Machines (SVM), and Decision Tree(DT) were evaluated to determine their effectiveness in price classification. The models were assessed using accuracy, precision, recall, and F1-score as performance metrics. The dataset used for training and evaluation was sourced from Kaggle.

This study offers insights into selecting suitable machine learning algorithms for mobile phone price classification, highlighting the trade-offs between predictive accuracy and computational efficiency. The results emphasize the advantages and limitations of each model in classifying mobile phone prices, providing valuable guidance for pricing strategies in the mobile industry.

Keywords: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest(RF), k-Nearest Neighbor (KNN), Decision Tree (DT).

1. Introduction

The continuous improvements in mobile industries have led to a vast selection of smartphones designed to meet varying user preferences and financial constraints. As the mobile market expands, categorizing devices into appropriate price segments has become essential for manufacturers, retailers, and consumers. Accurate price classification allows manufacturers to develop competitive pricing strategies, helps retailers enhance targeted marketing efforts, and assists consumers in selecting phones that align with their needs and budget.

This study aims to classify mobile phones into predefined price categories based on their specifications, such as storage, display quality, battery capacity, RAM, camera capabilities, and other key features. By leveraging machine learning, this project seeks to facilitate informed purchasing decisions for consumers while also providing insights into pricing strategies for industry stakeholders.

The classification task is addressed using five commonly used machine learning algorithms: Random Forest (RF), Support Vector Machines (SVM), Decision Tree (DT), k-Nearest Neighbours (k-NN), and Logistic Regression (LR). The primary goal is to evaluate the performance of these models.

To ensure reliable results, the dataset undergoes preprocessing steps, including handling missing values, normalizing features, and selecting the most relevant attributes. The data that is processed undergoes for training and testing the models, incorporating cross-validation techniques to enhance robustness.

By systematically analysing the effectiveness of these algorithms, this study provides valuable insights into their applicability for mobile phone price classification. Additionally, the findings contribute to a broader understanding of machine learning techniques in pricing-related classification tasks, offering guidance on selecting the best model for similar datasets for the future.

1.1 Scope of Project

This project aims to categorize mobile phone prices into predefined segments using machine learning techniques. It involves processing a dataset containing mobile specifications, applying data preprocessing methods, and training five widely used classification models. The models are assessed based on performance metrics to determine their effectiveness. The objective is to find suitable model for mobile price classification, offering valuable insights for market analysis and establishing a framework applicable to similar classification tasks in the future.

2. Literature Survey

The use of machine learning for mobile phone price classification has gained attention due to its practical applications in the consumer electronics industry. Earlier studies primarily utilized statistical methods such as linear regression and clustering to analyse patterns in pricing. However, with advancements in machine learning, more sophisticated classification techniques have been developed to better capture complex relationships among mobile phone features.

Researchers have investigated various classification models, to segment mobile phones into different price categories, emphasizing the importance of selecting feature in improving model accuracy. Ensemble techniques like Random Forest(RF) and Gradient Boosting (e.g., XGBoost and LightGBM) are popular and useful for their feature to handle large and high-dimensional datasets effectively.

Recent advancements have also explored deep learning models for mobile price prediction, particularly in analyzing unstructured data such as customer reviews and images. While deep learning approaches offer improved predictive performance, they often require substantial computational resources.

The growing body of research in this field highlights the ongoing challenge of balancing accuracy, interpretability, and computational efficiency in mobile price classification, ensuring that machine learning models remain both effective and practical for real-world applications.

3. Experiment

3.1 Data Collection and Preprocessing

Data Collection: The dataset used in this paper was sourced from Kaggle, containing detailed specifications of mobile phones, including battery capacity, processor speed, weight, RAM, and other key attributes.

Handling Missing Data: To ensure data quality, missing values were either imputed using appropriate statistical techniques or removed if they were insignificant or could not be accurately estimated.

Feature Selection: Key features such as RAM, battery life, processor type, and screen size were identified as significant factors influencing mobile phone pricing. These features were selected to enhance model performance and reduce unnecessary complexity.

Data Encoding: Categorical variables, such as brand names, are converted into numerical values by using encoding techniques like label encoding or one-hot encoding to make them suitable for machine learning models.

Normalization/Standardization: Normalize numerical features so that they are in the same range (e.g., between 0 and 1).

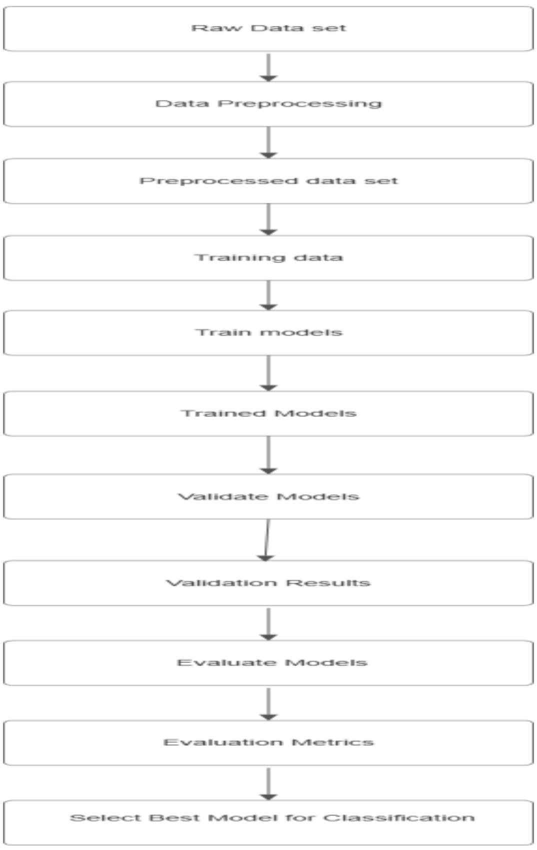


Figure 1: Data Flow Diagram

3.2 Exploratory Data Analysis

Correlation Analysis: Identify correlations between features and the price category using techniques like correlation matrices. Correlation analysis was selected to understand the relationships between different variables, allowing for a clearer assessment of how one feature influences another. This helps in predicting one variable based on the values of others, improving the accuracy and interpretability of the model.

Feature Importance: Rank features by importance using techniques such as feature importance from Random Forest or mutual information.

3.3 Model Training

Build, train and evaluate different models for classification.

Splitting Dataset: To ensure effective model training and evaluation, the dataset is categorized into two parts:

- **Training Part**– where the models are trained using classifiers.
- **Testing Part**– where the model performance is assessed on unseen data.

Choose Models: Train different classification models. The models considered are:

- LR
- DT
- RF
- KNN
- SVM

3.4 Evaluating the Model

Evaluate performance of model using relevant metrics. Classification is used machine learning technique in supervised learning, where a it learns from a labeled dataset to categorize new, unseen data into predefined classes. The effectiveness of a model is assessed using various evaluation metrics, which help determine how well the predicted values align with the actual outcomes.

To measure the model's performance, the following metrics are utilized:
Accuracy: It is the ratio of the number of correct classified instances to the total number of instances. It gives you a global perspective of how well the model is doing. **Precision:** Measures how many of the instances predicted as a particular class are actually correct, making it useful in scenarios where false positives need to be minimized. **Recall:** how well the model is able to effectively retrieve all related examples of a class, with an emphasis on minimizing false negatives. **F1-Score:** Is the harmonic mean of precision and recall and gives a balanced measure which works well especially in case of class imbalance.

Confusion Matrix: Determine the number of cases that were correctly or wrongly identified in each pricing category by analysing the confusion matrix.

4. System Implementation

The study was executed on a Jupyter Notebook version 2.10.0 employing Python programming environment. This relied on widely used libraries such as Pandas, NumPy, Seaborn, and Matplotlib. These libraries are essential tools, providing various functions and methods to handle, analyze, visualize, and model data.
The experiments were conducted using a 475 gigabytes (GB) computer powered by the intel i5 processor. It had 16 GB of installed RAM, with 15.8 GB available.

5. Applying Classifier and obtaining Results

The attributes in this study are classified into two types: categorical and numerical. The target variable, representing mobile phone price categories, is assigned four distinct labels:

- **0:** mobile phones of low range
- **1:** mobile phones of medium range
- **2:** mobile phones of high range
- **3:** mobile phones of premium range

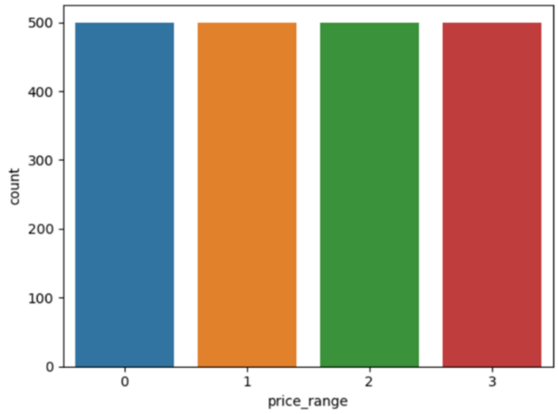


Figure 2: Count Graph of price range

5.1 Applying LR

Logistic regression (logit model) is a method for estimating the probability of an event occurring given the existence of a number of independent variables. It is especially helpful when we need to check the likelihood links between a categorical target variable and one or more predictor variables.
This method is commonly applied in classification tasks where the dependent variable consists of distinct categories. However, when a large number of predictor variables are included, logistic regression may be prone to overfitting, which can affect its generalization to new data.

Results:

```
Accuracy Score: 0.9583333333333334

Confusion Matrix:
 [[144   7   0   0]
 [  3 142   1   0]
 [  0   7 135   6]
 [  0   0   1 154]]

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.95      0.97       151
           1       0.91      0.97      0.94       146
           2       0.99      0.91      0.95       148
           3       0.96      0.99      0.98       155

    accuracy                           0.96       600
   macro avg       0.96      0.96      0.96       600
weighted avg       0.96      0.96      0.96       600
```

Figure 3: Output showing model evaluation when logistic regression classifier is applied

5.2 Applying K-NN

The k-Nearest Neighbors instead of learning an explicit model during training, it stores the dataset and classifies new instances after comparing and finding similarities with existing data points. Classification in k-NN is performed by comparing the majority class with the nearest neighbors of a given data point, using a predefined similarity measure. The algorithm is particularly effective for pattern recognition and classification tasks where relationships between data points are crucial.

Results:

```
Accuracy Score:
 0.5616666666666666

Confusion Matrix: [[115  31   5   0]
 [ 41  71  27   7]
 [ 11  56  61  20]
 [  1   9  55  90]]

Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.76      0.72       151
           1       0.43      0.49      0.45       146
           2       0.41      0.41      0.41       148
           3       0.77      0.58      0.66       155

    accuracy                           0.56       600
   macro avg       0.57      0.56      0.56       600
weighted avg       0.58      0.56      0.56       600
```

Figure 4: Output showing model evaluation when K-Nearest Neighbor classifier is applied

5.3 Applying SVM with Linear

A Support Vector Machine (SVM) employing a linear kernel is approach utilized for linearly separable data. Since a linear kernel assumes a straight-line separation between classes, it is efficient. SVM with a linear kernel is commonly applied in areas such as text classification and sentiment analysis, where features exhibit a linear relationship.

Results:

```
Accuracy Score: 0.905

Confusion Matrix:  [[141  10   0   0]
 [  5 134   7   0]
 [  0  14 132   2]
 [  0   0  19 136]]

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.93      0.95       151
           1       0.85      0.92      0.88       146
           2       0.84      0.89      0.86       148
           3       0.99      0.88      0.93       155

    accuracy                           0.91       600
   macro avg       0.91      0.91      0.91       600
weighted avg       0.91      0.91      0.91       600
```

Figure 5: Output showing model evaluation when SVM with Linear classifier is applied

5.4 Applying SVM with RBF

Support Vector Machine (SVM) with Radial Basis Function kernel is particularly used when data is not linearly separable. The RBF kernel transforms input data into a higher-dimensional space using a Gaussian function, enabling the creation of complex, non-linear decision boundaries. This kernel is influenced by two key parameters:

C(Regularization Parameter): Balances maximizing the margin and minimizing misclassification errors.

Gamma (γ): A higher gamma results in tighter decision boundaries, while a lower gamma produces smoother, more generalized boundaries.

Results:

```
Accuracy Score: 0.87

Confusion Matrix:  [[135  16   0   0]
 [ 14 123   9   0]
 [  0  19 122   7]
 [  0   0  13 142]]

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.89      0.90       151
           1       0.78      0.84      0.81       146
           2       0.85      0.82      0.84       148
           3       0.95      0.92      0.93       155

    accuracy                           0.87       600
   macro avg       0.87      0.87      0.87       600
weighted avg       0.87      0.87      0.87       600
```

Figure 6: Output showing model evaluation when SVM with RBF classifier is applied

5.5 Applying Decision Trees

Decision tree (hierarchical model) that processes input data through a sequence of decision rules based on feature values, ultimately leading to a predicted outcome. It follows a structured approach. They are applied across various domains, helping in decision-making processes by breaking down complex problems into a series of straightforward decisions.

Results:

```
Accuracy Score: 0.8333333333333334

Confusion Matrix:  [[130  21   0   0]
 [ 18 120   8   0]
 [  0  18 111  19]
 [  0   0  16 139]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.86      0.87       151
           1       0.75      0.82      0.79       146
           2       0.82      0.75      0.78       148
           3       0.88      0.90      0.89       155

    accuracy                           0.83       600
   macro avg       0.83      0.83      0.83       600
weighted avg       0.83      0.83      0.83       600
```

Figure 7: Output showing model evaluation when Decision Tree classifier is applied

5.6 Applying RF

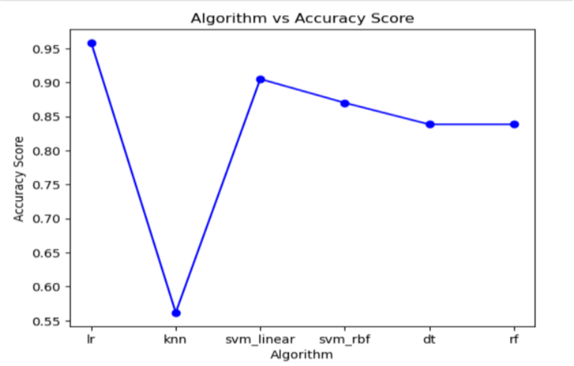
Random Forest enhances prediction accuracy by utilizing multiple decision trees, which are trained on different subsets of data. Instead of relying on a single decision tree, it employs **bagging** (Bootstrap Aggregating), where multiple models are trained on randomly selected portions of the dataset.

Results:

```
Accuracy Score: 0.8333333333333334

Confusion Matrix:  [[130  21   0   0]
 [ 18 120   8   0]
 [  0  18 111  19]
 [  0   0  16 139]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.86      0.87       151
           1       0.75      0.82      0.79       146
           2       0.82      0.75      0.78       148
           3       0.88      0.90      0.89       155

    accuracy                           0.83       600
   macro avg       0.83      0.83      0.83       600
weighted avg       0.83      0.83      0.83       600
```

Figure 8: Output showing model evaluation when Random Forest classifier is applied

5.7 High Accuracy Classifier



Based on the evaluation results, the Logistic Regression model demonstrated the highest accuracy, achieving a 96% accuracy rate.

6. Conclusion

Applying machine learning to mobile price classification demonstrates effectiveness of different algorithms in accurately classifying price categories based on device features. Through a comparative analysis of models like Logistic Regression (LR), Decision Trees, SVM, it is evident that ensemble-based techniques consistently deliver higher accuracy and robustness, especially for complex datasets. Feature importance analysis highlights key factors such as RAM, processor speed, and battery capacity as critical determinants of price classification. While simpler models like Logistic Regression offer interpretability, advanced methods like SVM with RBF kernels and Gradient Boosting provide superior performance in non-linear and high-dimensional scenarios. This paper gives the importance of algorithm selection based on data characteristics and computational constraints, paving the way for scalable and efficient solutions in the mobile phone industry. Future research can focus on incorporating deep learning methods and real-time price prediction models to improve classification accuracy and enhance practical applications.

7. 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 Transformation, 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 Sci ence, 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 Dimensionality Reduction on Classifier Training Time and Quality," 3rd International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP), Bangkok, Thai-land, 2023.