Study on the Performance of Supervised Machine Learning Algorithms in Mobile Price Range Classification

A.S.M. Sabiqul Hassan
Department of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh

Syed Maruful Huq
Department of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh

Mohammad Kamal Hossain Foraji
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Bangladesh

Md. Humayun Kabir
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Bangladesh

ABSTRACT

The prior prediction of the mobile price range based on different features can help potential customers to purchase their target mobile phones. It also helps manufacturers to develop a decision-making model in setting up the price range, e.g., very economical, economical, expensive and very expensive of upcoming mobile phones with different features. This paper explores some machine learning algorithms and their application in classifying of mobile phone price ranges by analyzing a dataset collected from the Kaggle online dataset repository. The dataset was divided into three partitions where a train set consisting of 70% data, validation set and test set each sharing the remaining 30% data equally. Then, different classification algorithms: K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM) and Logistic Regression (LR) were applied to the dataset to develop machine learning models. Finally, SVM model achieved the highest F1 score of 97% among the developed machine learning models. The knowledge extracted from that model can be used as a decision-making tool for predicting the prices of mobile phones and classifying their range in the future.

Keywords

Machine Learning, Supervised Learning, Classification, Mobile Price, Price Prediction, Decision Making, Data Mining

1. INTRODUCTION

In the marketing and business world, price is considered one of the most influential factors for both customers and manufacturers. It determines the acceptance of a product in today’s competitive market. The purchasing decision of a product depends not only on its price but also on its different features to justify the cost. In the mobile phone industry, new models with some advanced features are frequently released which makes price prediction of a product difficult for both customers and manufacturers [1-2].

Traditional pricing methods depend on market analysis and expert judgments. To remain competitive in the market, setting an optimal price, i.e., the minimum cost with the maximum features of a product is essential for the companies. A tool or business model that predicts mobile phone prices based on their various features can help companies set a competitive price and guide customers in decision-making before a purchase [3-4]. By analyzing previous data and identifying important determinants of pricing, machine learning algorithms provide a better solution for price prediction. Mobile phone prices of the developed models can be classified into different categories, e.g., very economical, economical, expensive, and very expensive by applying various machine learning classification algorithms. By reducing dimensionality and computational complexity, feature selection techniques assist in optimizing the performance of the developed machine learning models. As a result, only the most relevant features that influence mobile price prediction are selected [5].

The prediction of mobile prices is a complex challenge in the rapidly changing world of mobile technology. Mobile manufacturers require an effective model to calculate the optimal mobile price based on its various important features, e.g., processor speed, battery capacity, camera quality, display size and memory. On the other hand, customers require a tool that allows them to predict the price of a mobile phone based on their desired features.

The existing research works have explored different classification models developed using machine learning algorithms for mobile phone price prediction and classification [6-9]. However, many studies didn’t follow an integrated approach that balances the performance metrics to provide an extensive evaluation of the developed models. Moreover, previous studies did not compare the classification models to determine the most effective one for mobile phone price categorization. The goal of this research is to address these limitations by utilizing various machine learning techniques and evaluating the performance of the developed models using different evaluation metrics.

This paper is organized as follows. Section 2 provides an overview of mobile price prediction using different classification models. Section 3 describes the approach for classifying mobile phone price ranges using the optimal model among the developed machine learning classification models. In Section 4, the results of the selected classifiers are analyzed. Finally, Section 5 concludes the research with guidelines to future work.

2. RELATED WORKS

The use of machine learning techniques to predict the mobile phone price range has become significantly popular in recent years. To enhance the prediction accuracy, various studies have employed different machine learning algorithms and feature selection methods. Subhiksha et al. [6] developed a classification model to predict mobile phone price ranges using three machine learning algorithms, e.g., LR, RF and SVM. Based on their findings, SVM model achieved the highest accuracy among the developed classification models.

Kalaivani et al. [7] focused primarily on predicting the mobile phone price ranges using SVM, RFC and LR. They used a Chi-Squared based feature selection method to the dataset to improve classification accuracy. After feature selection, they found that SVM outperformed the other classifiers and achieved an accuracy of 96%. In another study, Asim et al. [8] emphasized the importance of selecting appropriate models for accurate mobile phone price prediction. They found that LR model enhanced with the Elastic-Net parameter outperformed other classification models and achieved an accuracy of 96%.

Zehtab-Salmasi et al. [9] suggested the use of deep learning approaches to predict mobile phone price ranges. In their proposal they included five deep learning approaches where one was unimodal and four were multimodal approaches. Their multimodal methods achieved an F1 Score of 88.3% by considering both graphical and non-graphical features. Additionally, multimodal learning generated more accurate predictions than state-of-the-art techniques.

These studies have made some significant progress in the field of mobile phone price range prediction, but there are certain gaps remain at various steps [6-9]. The application of feature selection methods such as Chi-Squared has not extended to a thorough exploration of advanced feature engineering techniques to capture complex interactions between different features. In terms of algorithm diversity, the main focus for the majority of studies is on the traditional machine learning algorithms. The exploration of ensemble methods and deep learning architectures could potentially capture non-linear relationships more effectively.

Many researchers have used datasets from platforms such as Kaggle, UCI Machine Learning data repository which may not fully represent the current global market or ensure the diversity of mobile phone features. To achieve better performance, datasets collected from the target market are recommended. Only a few studies have integrated these predictive models into real-world applications, such as decision-making tools for customers or manufacturers [6-9]. To develop more robust and practical models for mobile phone price range classification, these gaps should be addressed.

3. MOBILE PHONE PRICE RANGE CLASSIFICATION PROCESS

In this research work, the dataset used in the developed model was collected from the Kaggle online dataset repository [10]. Then, an optimal classification model was developed using five different machine learning algorithms to classify mobile phone price ranges. The results were analyzed for future use in the decision-making process. The classification task was performed on the Google Colab platform using the Python programming language. The workflow of the optimal classification model is described below to provide a clearer understanding.

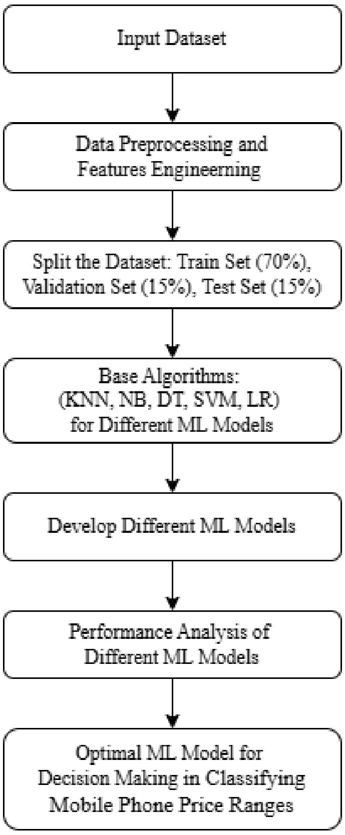
Input Dataset
Data Preprocessing and Features Engineerning
Split the Dataset: Train Set (70%), Validation Set (15%), Test Set (15%)
Base Algorithms: (KNN, NB, DT, SVM, LR) for Different ML Models
Develop Different ML Models
Performance Analysis of Different ML Models
Optimal ML Model for Decision Making in Classifying Mobile Phone Price Ranges

Figure 1: Steps of Mobile Phone Price Ranges Classification

3.1 Description of the Dataset

The selected dataset [10] had 2000 instances with 20 features and only 1 label, i.e., target attribute. The short illustration of the dataset features are given below.

Table 1: Description of the Dataset

Feature	Data Examples	Data type
Battery Power	501, 842 etc. in mAh	Numeric
Bluetooth Support	0 i.e. False, 1 i.e. True	Boolean
Clock Speed	0.5, 1.2 etc. in GHz	Numeric
Dual Sim	0 i.e. False, 1 i.e. True	Boolean
Front Camera	3, 7 etc. in MP	Numeric
4G Support	0 i.e. False, 1 i.e. True	Boolean
Internal Memory	7, 53 etc. in GHz	Numeric
Mobile Depth	0.6, 0.9 in cm	Numeric
Mobile Weight	136, 188 etc. in gm	Numeric
No of Processor Core	3, 5 etc.	Numeric
Primary Camera	2, 6 etc. in MP	Numeric
Resolution in Height	20, 1263 etc. in pixels	Numeric
Resolution in Width	756, 1988 etc. in pixels	Numeric
RAM Limit	2549, 2631 etc. in MB	Numeric
Screen Height	9, 11 etc. in cm	Numeric
Screen Width	3, 7 etc. in cm	Numeric
Batter Backup	7, 9 etc. in hours	Numeric
3G Support	0 i.e. False, 1 i.e. True	Boolean
Touch Screen Support	0 i.e. False, 1 i.e. True	Boolean
Wifi Support	0 i.e. False, 1 i.e. True	Boolean
Target Attribute: Phone Price Range	0 i.e. very economical , 1 i.e. economical, 2 i.e. expensive and 3 i.e. very expensive	Numeric

3.2 Dataset Preprocessing and Features Engineering

The collected dataset was preprocessed for different scenarios, e.g., dropping of all null values, conversion of categorical values into numerical ones using the One Hot Encoding method [11]. Then, features engineering and scaling were applied on the dataset to prepare it for better analysis of the developed classification models.

3.3 Split the Dataset

The preprocessed dataset was divided into three partitions where train set consisting 70% data, validation set and test set each sharing the remaining 30% data equally. Then, various machine learning algorithms were implemented to develop an optimal model.

3.4 Applied Machine Learning Algorithms

To develop an optimal model for classifying mobile phone price ranges based on the available features, five machine learning algorithms were selected.

3.4.1 K-Nearest Neighbors (KNN)

It is a basic classification algorithm that finds the result of n=k data points in its target data space and the final class is decided based on the majority of the class [12]. It is used by the researchers for small datasets as it is easier to use effectively.

3.4.2 Decision Tree (DT)

It creates a structure similar to tree that has internal nodes representing a decision based features and leaf nodes representing a class label. For classification problems in structured dataset, it is easier to interpret [13].

3.4.3 Naïve Bayes (NB)

It is a probabilistic algorithm that classifies based on Bayes' theorem. In this algorithm, all features are considered conditionally independent. For text classification and other applications with high-dimensional data, it often performs better in spite of being a simple algorithm [14].

3.4.4 Logistic Regression (LR)

It is a statistical algorithm applied for binary and multi-class classification problems. Using a logistic function, it calculates the probability of a category based on input features. It performs better if the relationships among the features are linear [15].

3.4.5 Support Vector Machine (SVM)

It is a powerful algorithm that estimates an optimal hyper plane to differentiate the data-points of different categories. If the data-points are not linearly separable, it particularly performs better for high dimensional spaces [16].

4. RESULT ANALYSIS

In this paper, the used dataset [10] had 2000 instances with 20 features of mobile phones. It was analyzed to classify the price ranges of mobile phones using different classification models: KNN, NB, DT, LR, and SVM.

4.1 Confusion Matrix

A confusion matrix is mostly used to describe the performance of a machine learning classification model. It explains the breakdown of correct and incorrect predictions among different target classes. It also assists in assessing the model accuracy and effectiveness beyond just a single accuracy metric [17-18].

Table 2: A Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

A confusion matrix contains four metrics: *i) True Positives (TP):* it correctly classifies positive instances, *ii) True Negatives (TN):* it correctly classifies negative instances, *iii) False Positives (FP):* it incorrectly classifies negative instances (Type I error), and *iv) False Negatives (FN):* it incorrectly classifies positive instances (Type II error).

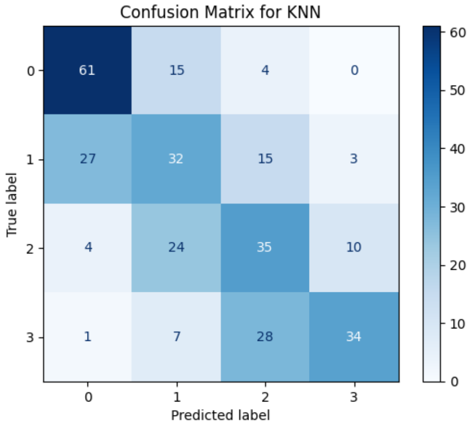

Figure 2: Confusion Matrix for KNN Model

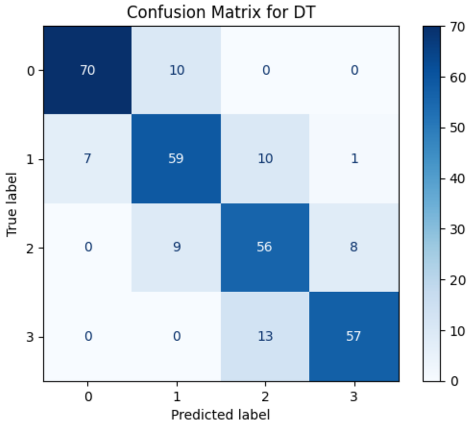

Figure 3: Confusion Matrix for DT Model

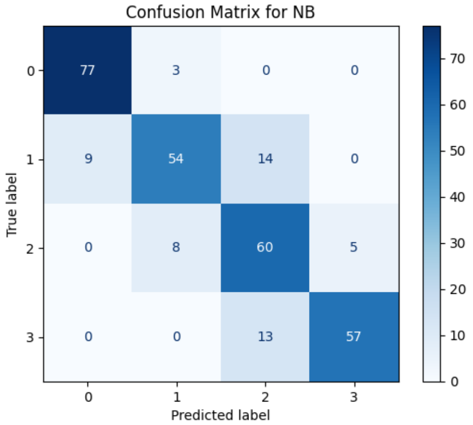

Figure 4: Confusion Matrix for NB Model

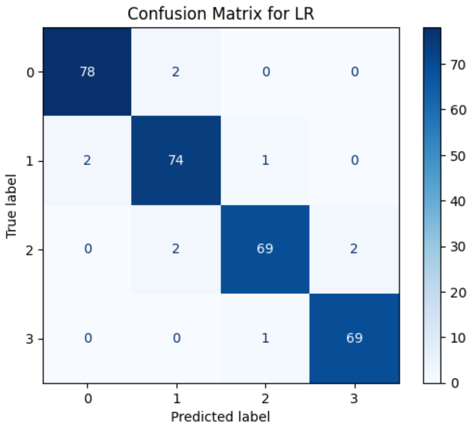

Figure 5: Confusion Matrix for LR Model

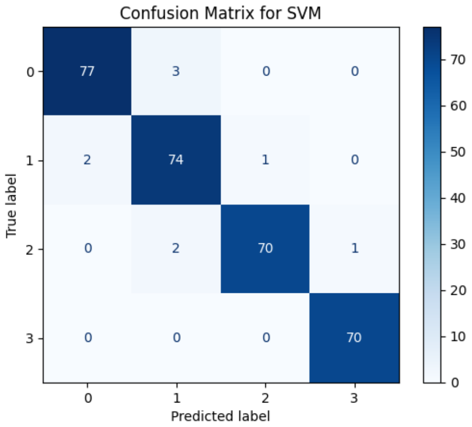

Figure 6: Confusion Matrix for SVM Model

4.2 Performance Evaluation

The use of a confusion matrix is helpful in the diagnosis of performance issues, e.g., class imbalance or misclassification trends which lead to better model optimization. It calculates some performance metrics: Accuracy, Precision, Recall, and F1 Score. These metrics are crucial in the evaluation process of a machine learning model [17, 18].

4.2.1 Accuracy

It measures the amount of the correctly classified instances among all instances of the given dataset. It is useful only for the balanced dataset but it doesn't provide better result for the imbalanced dataset where one class outranks other classes [17].

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

4.2.2 Precision

It quantifies the number of actual positive among the predicted positive cases. Its main focus is to reduce false positive cases. It is used in critical condition when false positives are costly, e.g., fraud detection or medical diagnosis. The higher precision value indicates fewer irrelevant results are classified as positive [17].

$$Precision = \frac{TP}{TP + FP}$$

4.2.3 Recall

It measures the number of correctly predicted values among actual positive cases. Its main focus is to reduce false negatives. Its importance is noticed when missing positive cases are costly, e.g., cancer diagnosis. The higher recall value helps to correctly identify most of the actual positive cases [17].

$$Recall = \frac{TP}{TP + FN}$$

4.2.4 F1 Score

It is used to maintain a balance between the values of precision and recall. Mostly, it is useful in a scenario when there is an imbalance between false positive and false negatives. Its best use case is the necessity of the balance between precision and recall. It performs better in imbalanced datasets where only accuracy can be misleading. In that case, the performance of a machine learning model can be evaluated using F1 Score instead of Accuracy [17].

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

4.2.5 Performance Comparison

The performance of different classification models are presented in Table-3 using different evaluation metrics for classifying of mobile phone price ranges.

Table 3: Result Statistics of Classification Models

Classification Models	Accuracy	Precision	Recall	F1 Score
KNN	0.540000	0.552866	0.540000	0.539537
DT	0.806667	0.810574	0.806667	0.808076
NB	0.826667	0.834323	0.826667	0.826616
LR	0.966667	0.966744	0.966667	0.966639
SVM	0.970000	0.970291	0.970000	0.970038

Generally, Accuracy and F1 Score both are considered for the result analysis of the machine learning classification models. LR and SVM models provided the highest F1 Score around 97%. Then, DT and NB models provided the second highest F1 Score around 81% and KNN model only provided around 54% F1 Score. The result of Accuracy, Precision, and Recall were also found close to the F1 Score.

As KNN algorithm depends on the distance metric, i.e., Euclidean distance formula, it is highly sensitive to some irrelevant or noisy features [19]. Therefore, KNN model provided the least performance compared to other models for all metrics.

The knowledge extracted from analyzing the developed machine learning models for classifying mobile price ranges can later be used in efficient decision-making [20, 21] for mobile sales promotion. To recommend a group of customers for purchasing mobile phones in different price ranges, the knowledge represented by the developed model on the mentioned dataset can be applied in decision-making.

Figure 7, 8, 9 and 10 visually represent the performance of different classification models developed using selected machine learning algorithms based on the Table-3 statistics.

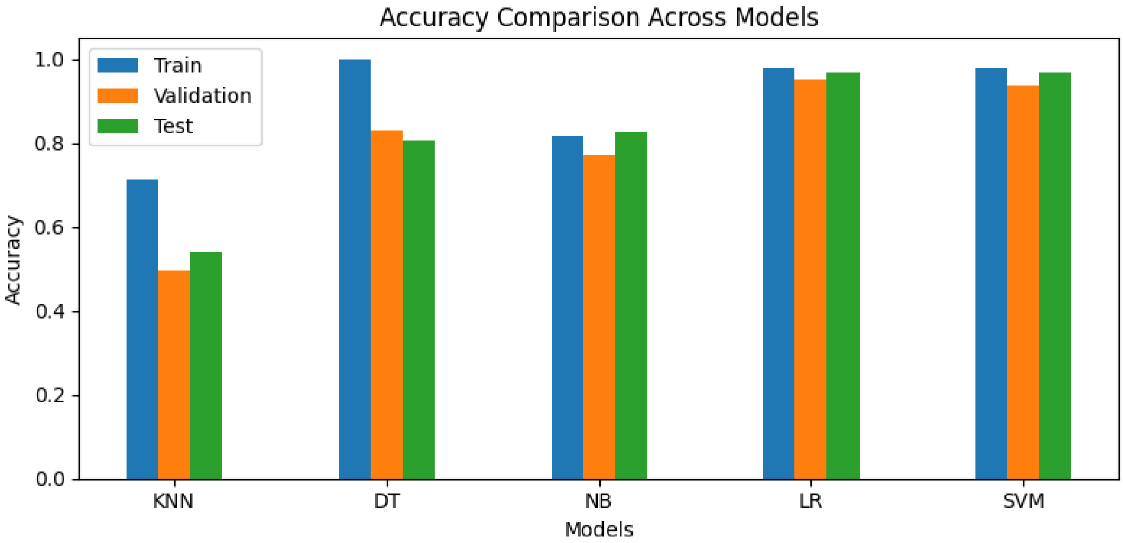

Figure 7: Accuracy Comparison Across Different Classification Models

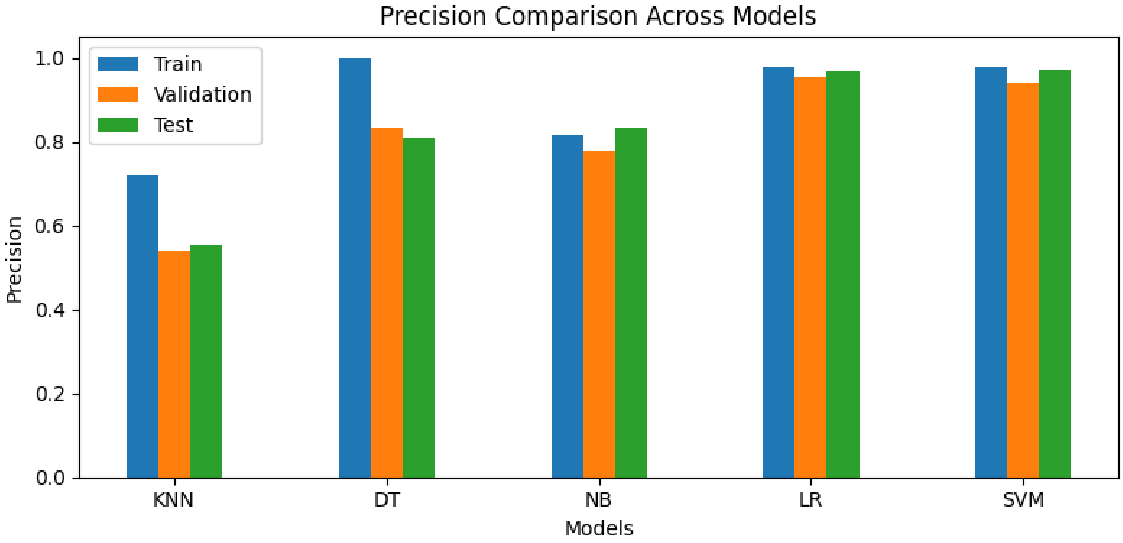

Figure 8: Precision Comparison Across Different Classification Models

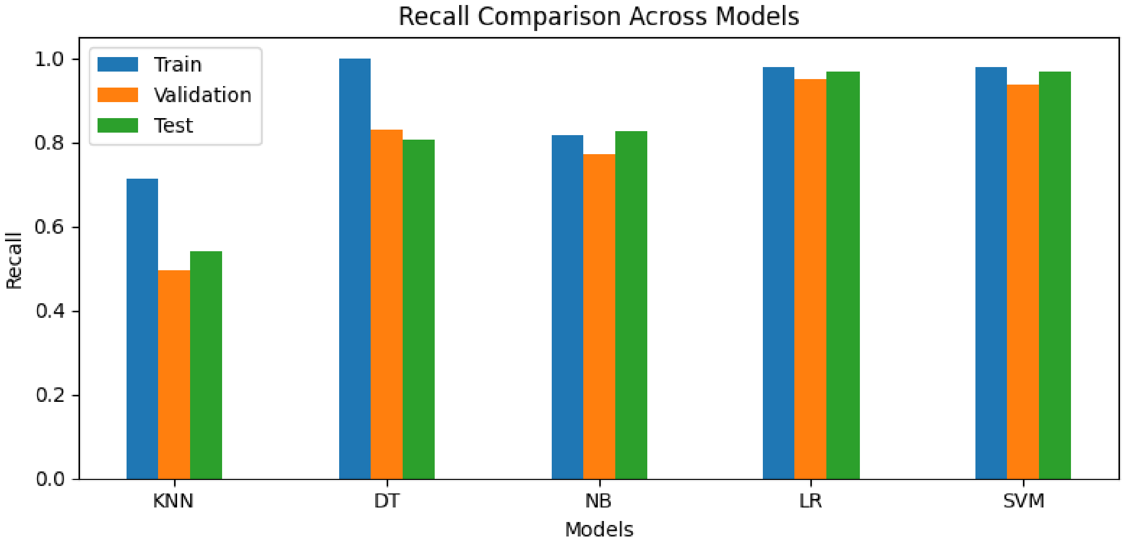

Figure 9: Recall Comparison Across Different Classification Models

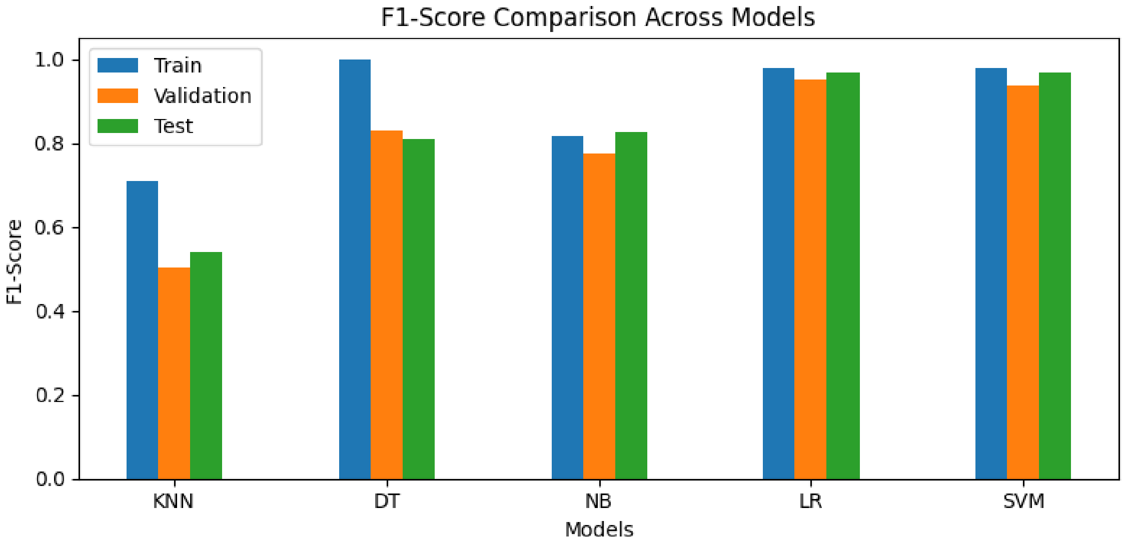

Figure 10: F1 Score Comparison Across Different Classification Models

5. CONCLUSION AND FUTURE WORK

This paper presents an optimal machine learning model for classifying mobile phone price ranges. The dataset used was collected from the Kaggle online dataset repository. After data preprocessing and features engineering, the dataset was divided into three different partitions: train set, validation set and test set. Several machine learning classification algorithms: KNN, NB, DT, SVM and LR were applied to mobile phones price range dataset to train the desired classification models. The performance of the developed models was analyzed using different performance evaluation metrics: Accuracy, Precision, Recall and F1 Score. SVM model provided the highest F1 Score of 96%. In terms of Accuracy and F1 Score, SVM and LR provided the highest performance result of around 97% among the developed classification models. DT and NB models produced the second highest performance within the range of 80-82%. KNN model achieved the least performance of around 54%. The other two performance evaluation metrics: Precision and Recall also generated the result close to F1 Score.

The research work explored different classification models and analyzed their performance. The performance of the developed optimal classification model can be improved by adding more efficient and complex models like Random Forest (RF) and Artificial Neural Networks (ANN). The dataset can be prepared and updated regularly based on the target market and customer data.

6. ACKNOWLEDGEMENTS

The authors are thankful to people who assisted with their valuable opinions and guidance in the process of conducting this research work and writing the paper.

7. REFERENCES

[1] Lashari, S. A., Khan, M. M., Khan, A., Salahuddin, S., and Ata, M. N. (2024). Comparative Evaluation of Machine Learning Models for Mobile Phone Price Prediction: Assessing Accuracy, Robustness, and Generalization Performance. Journal of Informatics and Web Engineering, 3(3), 147-163.

[2] Liang, Q. (2024). Mobile phone price prediction: A comparative study among four models. Applied and Computational Engineering, 48, 212-218.

[3] Chandrashekhara, K. T., Thungamani, M., Gireesh Babu, C. N., and Manjunath, T. N. (2019). Smartphone price prediction in retail industry using machine learning techniques. In Emerging Research in Electronics, Computer Science and Technology: Proceedings of International Conference, ICERECT 2018 (pp. 363-373). Springer Singapore.

[4] Mahoto, N. A., Iftikhar, R., Shaikh, A., Asiri, Y., Alghamdi, A., and Rajab, K. (2021). An Intelligent Business Model for Product Price Prediction Using Machine Learning Approach. Intelligent Automation & Soft Computing, 30(1).

[5] Chen, M. (2023). Mobile Phone Price Prediction with Feature Reduction. Highlights in Science, Engineering and Technology, 34, 155-162.

[6] Subhiksha, S., Thota, S., and Sangeetha, J. (2020). Prediction of phone prices using machine learning techniques. In Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19 (pp. 781-789). Springer Singapore.

[7] Kalaivani, K. S., Priyadharshini, N., Nivedhashri, S., and Nandhini, R. (2021, November). Predicting the price range of mobile phones using machine learning techniques. In AIP Conference Proceedings (Vol. 2387, No. 1). AIP Publishing.

[8] Asim, M., and Khan, Z. (2018). Mobile price class prediction using machine learning techniques. International Journal of Computer Applications, 179(29), 6-11.

[9] Zehtab-Salmasi, A., Feizi-Derakhshi, A. R., Nikzad-Khasmakhi, N., Asgari-Chenaghlu, M., and Nabipour, S. (2023). Multimodal price prediction. Annals of Data Science, 10(3), 619-635.

[10] Mobile Price Range Classification. Dataset url: https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification

[11] Samuels, J. I. (2024). One-hot encoding and two-hot encoding: an introduction. Preprint at, 10.

[12] Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[13] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1, 81-106.

[14] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

[15] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society Series B: Statistical Methodology, 20(2), 215-232.

[16] Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

[17] Tharwat, A. (2021). Classification assessment methods. Applied computing and informatics, 17(1), 168-192.

[18] Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

[19] Bansal, M., Goyal, A., and Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal, 3, 100071.

[20] Kusiak, A. (2002, March). Data mining and decision making. In Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV (Vol. 4730, pp. 155-165). SPIE.

[21] Md. Humayun Kabir. “Study on the Performance of Classification Algorithms for Data Mining". IOSR Journal of Computer Engineering (IOSR-JCE) 21.3 (2019): 23-30.