Implementation of Supervised Algorithm for Classification and Prediction of Smartphone Prices using K-Nearest Neighbour

I. INTRODUCTION

Smartphone industry is booming at an exponential speed. Everyday new mobiles are launched in the market and deciding the price of these mobile phones plays a crucial role. The pricing affects sale and position of the company in market. As new features are added every day in the smartphones, finding the optimum product is challenging from a user's perspective. Supervised learning algorithms can be used to predict the values based on labelled data. These can be used in prediction of mobile prices based on the features available in the smartphone. This prediction can help the manufacturers of smartphones to estimate mobile prices and at the same time competing with the other manufacturers. It can also help the consumers and the buyers to evaluate various mobile phone prices and decide which one to buy. Further it can assist them to assure that they are paying the best price for the mobile they are purchasing. This is extremely beneficial for emerging companies to set the prices in accordance to the market standards. The most essential factor to estimate the cost is the features and the preference of user. For instance, processor plays an important role in deciding the price. Similarly, screen size or camera megapixels play a vital role in differentiating between prices of the smartphones. Youth generally prefer handsets with good processor to play games. On the other hand, older people prefer phones with larger screen sizes. Females opt to choose mobiles with better camera pixels. Hence, the features of a smartphone are key consideration for predicting the prices. In this paper, we aim to visualize the attributes affecting the prices of smartphones and find the correlation among them. Also, we will analyse various models and compare their accuracy values. Finally, we will predict the class values for unlabelled samples when other features are provided with the model with best performance. Organization of this paper is as mentioned below. Section 2 contains the literature survey. Section 3 contains methodology containing the architecture and model development. Experimental setup is described in section 4. Experimental results are summarized in section 5 followed by conclusion in section 6.

II. LITERATURE SURVEY

B. Jeevan et al. [1] provides a method to predict the share prices using Long Short Term Memory (LSTM) and Recurrent Neural Network (RNN). It predicts the stock price of a company which is chosen using one of the methods in collaborative or content based recommendation and computes performance. Tiwari et al. [2] have used Artificial Neural Network (ANN) with back propagation to analyse mobile prices. The authors have focussed on finding relation between features and the price to obtain the results.

Varma et al. [3] are proposing a system to predict house prices in Mumbai. The authors have used forest, boosted and linear regression along with neural network to increase the efficiency of the system. In [4], authors have used DOA architecture to predict the vegetable prices in China. They have combination of neural network along with wavelet transformation forecasting. Prediction of car prices is done in [5] by the authors using quantifying the qualitative data. Also the authors have used knowledge based system for the analysis and rule formation.

Durganjali et al. [6] have boosted weak learners to strong by using Adaboost. Classification algorithms used by the authors are Naïve Bayes, Decision Tree, Logistic Regression and Random Forest to predict the resale price of houses. In [7], authors have tried to predict the price of bitcoin based in data from exchange, Covid-19 data and twitter data. Authors have used four algorithms namely, Adaboost, Decision Tree, SVM and Random Forest.

Vivek et al. [8] describes regression techniques to predict house prices. XGboost, SVM, Decision Tree and Random Forest regression techniques are used by the authors for predicting the house prices. Authors in [9] used gated recurrent network model (GRU) to predict bitcoin prices. Using this method authors could obtain overall accuracy of 94.70%. Liu et al. [10] have used KNN for classifying the prices of commodities on websites. Furthermore, they have used Decision Tree regression technique to predict the prices.

III. METHODOLOGY

Figure 1 shows the architecture for the proposed model. Firstly, the training data is pre-processed. Heat map is generated to find the correlation among the features and the target variable. Then several models are tested for accuracy.

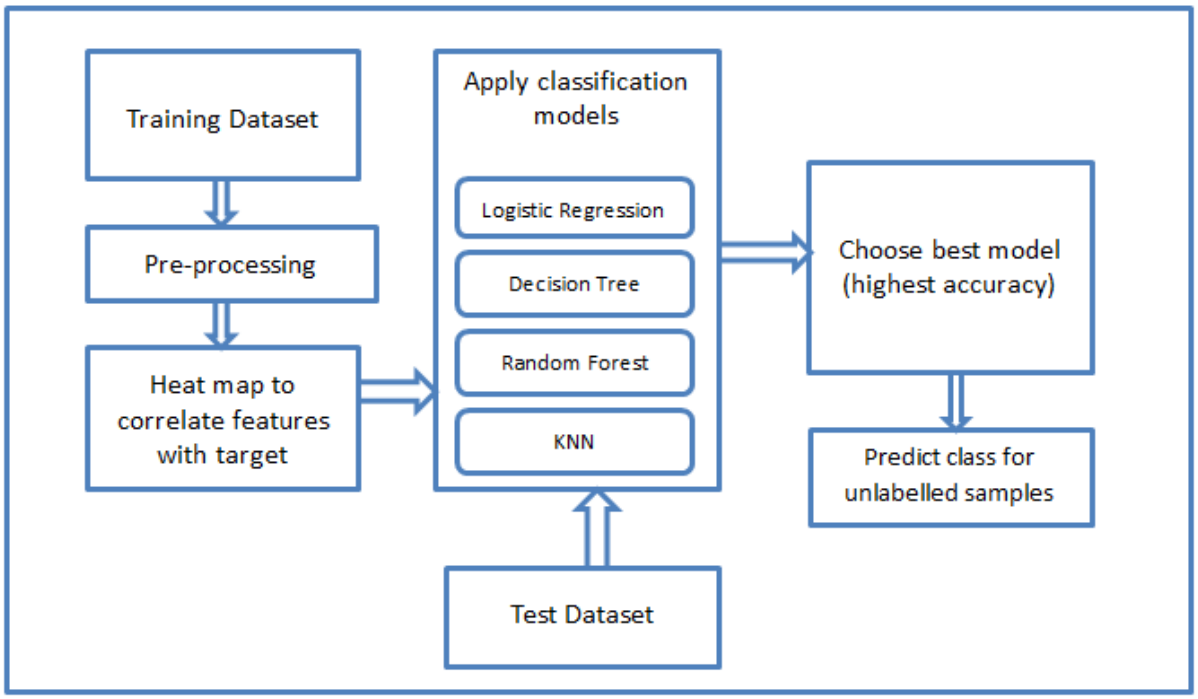

Figure 1: Architecture for proposed model

In this project, we have considered logistic regression, decision tree, random forest and k-nearest neighbour. All models are evaluated with the test dataset for accuracy. Among the four models considered, whichever is giving the best accuracy is chosen for the future predictions. Heat map is generated for all the features as shown in figure 2. The most influential element as seen from the heat map is ram. Most of the variables have little correlation with price_range.

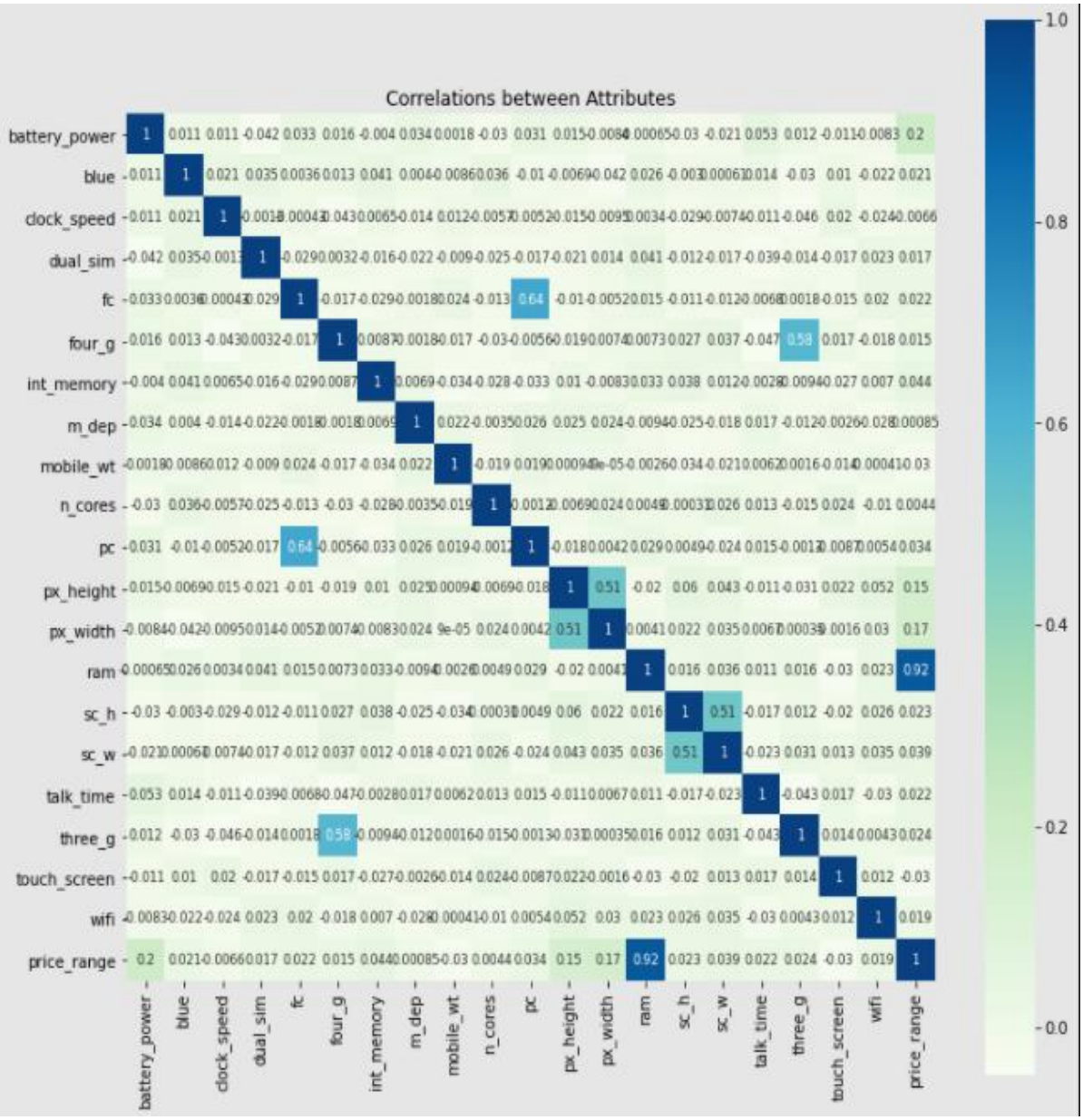

Figure 2: Heat map to measure correlation of feature to target

For this project, we have used k-nearest neighbour (knn) model which is supervised machine learning algorithm used for classification. It basically works on the idea of calculating the distance between the test data point and the input and accordingly predicts the sample.

Euclidean distance is used in knn algorithm for calculating distances. It is calculated as below:

Euclidean Distance:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \quad (1)$$

This equation is simplified as below.

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (2)$$

Distances are also computed using the following formulas:

Manhattan Distance:

$$d(p,q) = \sum_{i=1}^{n} |p_i - q_i| \quad (3)$$

Makowski Distance:

$$d(p,q) = \left(\sum_{i=1}^{n} |p_i - q_i|^c \right)^{\frac{1}{c}} \quad (4)$$

Above mentioned Manhattan and Makowski distance are used along with Euclidean distance in few models to compute the distance.

IV. EXPERIMENTAL SETUP

The results are obtained using the following experimental setup. This project is implemented using python in jupyter notebook. The dataset is obtained from kaggle website mentioned in [11]. The data set contains two files, namely, train.csv and test.csv. Training dataset has 2000 samples with 21 attributes. The class label is the price_range attribute in the dataset has values between 0-4, where 0 represents the lowest price and 4 represent the maximum price. Obtained dataset is pre-processed and has no null or empty values. Model is built using training dataset and k fold cross validation is used for testing the accuracy of the chosen model. Test file has 1000 unlabelled samples and is used for new predictions based on the given dataset once the model is trained.

V. EXPERIMENTAL RESULTS

The accuracy scores of the four models considered in this study is shown in figure 3. Logistic Regression has given accuracy of 0.73. Decision Tree has accuracy of 0.83. Random Forest has accuracy of 0.90. KNN has the highest accuracy of 0.95 for the given dataset.

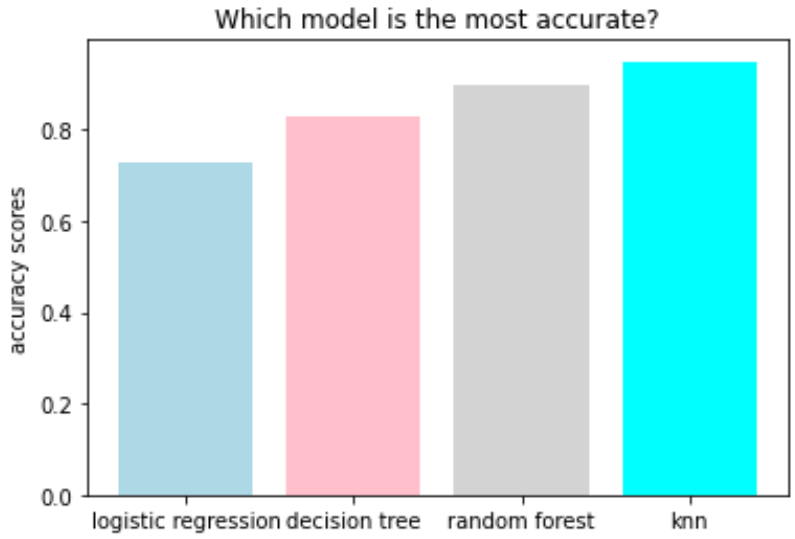

Figure 3: Accuracy of chosen models

Given in figure 4 is the confusion matrix for KNN algorithm. The value of k is chosen by GridSearchCV method in sklearn module in python and is giving the best accuracy when k is chosen as 9. In the figure below Predicted label is on x axis and true label is on y axis.

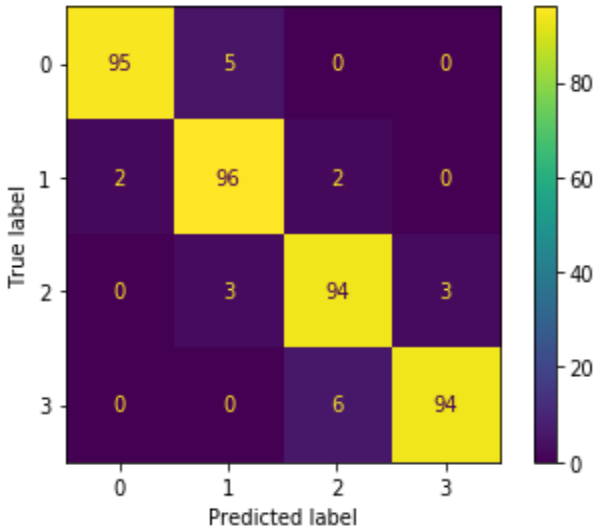

Figure 4: Confusion Matrix for KNN

The table given below shows the metrics for all the four models considered. Precision, Recall, Accuracy, F1 score and support is mentioned in the below table. The values are according to the classes where there are four classes, 0, 1, 2 and 3.

Table 1: Metrics values for all models

Model	Precision				Recall				F1 Score			
Class	0	1	2	3	0	1	2	3	0	1	2	3
Logistic Regression	0.92	0.72	0.57	0.72	0.88	0.64	0.58	0.82	0.90	0.68	0.58	0.77
Decision Tree	0.92	0.79	0.72	0.90	0.89	0.74	0.80	0.88	0.90	0.76	0.76	0.89
Random Forest	0.97	0.85	0.88	0.92	0.91	0.91	0.85	0.94	0.94	0.88	0.86	0.93
K Nearest Neighbour	0.98	0.92	0.92	0.97	0.95	0.96	0.94	0.94	0.96	0.94	0.93	0.95

As KNN has the highest accuracy it can be used to further classify the new samples. Figure 5 below shows the predicted class values for the 5 samples when the other attributes were given to the model. Similarly, this can be used to make future predictions for price of smartphones.

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_h	sc_w	talk_time	three_g	touch_screen	wifi	price_range
0	1043	1	1.8	1	14	0	5	0.1	193	3	16	226	1412	3476	12	7	2	0	1	0	3
1	841	1	0.5	1	4	1	61	0.8	191	5	12	746	857	3895	6	0	7	1	0	0	3
2	1807	1	2.8	0	1	0	27	0.9	186	3	4	1270	1366	2396	17	10	10	0	1	1	2
3	1546	0	0.5	1	18	1	25	0.5	96	8	20	295	1752	3893	10	0	7	1	1	0	3
4	1434	0	1.4	0	11	1	49	0.5	108	6	18	749	810	1773	15	8	7	1	0	1	1

Figure 5: Predicted class for 5 samples

VI. CONCLUSION

This study is focussed on utilizing the efficiency of machine learning models in prediction of prices for smartphones. Various experiments were carried out to determine the performance of various supervised machine learning algorithms. The algorithms used are Logistic Regression, Decision Tree, Random Forest and K-Nearest Neighbour. For the given dataset, KNN outperformed the other models considered. With KNN an accuracy of 95% was obtained. Moreover, the confusion matrix was plotted for all the models that were used in the study and the classification report was generated that contains the values for precision, recall, F1 score and support. KNN was further used to predict class for 1000 unlabelled samples given the other features were provided to the model. The chosen algorithm could successfully predict the class for these samples.

VII. REFERENCES

1. B. Jeevan, E. Naresh, B. P. V. kumar and P. Kambli, "Share Price Prediction using Machine Learning Technique," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739647.
2. A. Tiwari, V. Singh, P. Shukla and M. Darbari, "Feature Extraction for Mobile Handset in Coherency with Pricing Factors," 2020 Journal of Mechanics of Continua and Mathematical Sciences, 2020, pp. 297-310, doi: 10.26782/JMCMS.2020.03.00024.
3. A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.
4. G. Zheng, H. Zhang, J. Han, C. Zhuang and L. Xi, "The Research on Agricultural Product Price Forecasting Service Based on Combination Model," 2020 IEEE 13th International Conference on Cloud Computing (CLOUD), 2020, pp. 4-9, doi: 10.1109/CLOUD49709.2020.00009.
5. D. Van Thai, L. Ngoc Son, P. V. Tien, N. Nhat Anh and N. T. Ngoc Anh, "Prediction car prices using quantify qualitative data and knowledge-based system," 2019 11th International Conference on Knowledge and Systems Engineering (KSE), 2019, pp. 1-5, doi: 10.1109/KSE.2019.8919408.
6. P. Durganjali and M. V. Pujitha, "House Resale Price Prediction Using Classification Algorithms," 2019 International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-4, doi: 10.1109/ICSSS.2019.8882842.
7. J. Luo, "Bitcoin price prediction in the time of COVID-19," 2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID), 2020, pp. 243-247, doi: 10.1109/MSIEID52046.2020.00050.
8. V. S. Rana, J. Mondal, A. Sharma and I. Kashyap, "House Price Prediction Using Optimal Regression Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 203-208, doi: 10.1109/ICACCCN51052.2020.9362864.
9. M. Rizwan, S. Narejo and M. Javed, "Bitcoin price prediction using Deep Learning Algorithm," 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), 2019, pp. 1-7, doi: 10.1109/MACS48846.2019.9024772.
10. Y. Liu and Y. Lv, "Commodity Price Evaluation Based on Improved Data Mining Methods," 2020 International Conference on E-Commerce and Internet Technology (ECIT), 2020, pp. 145-148, doi: 10.1109/ECIT50008.2020.00039.
11. Dataset - https://www.kaggle.com/iabhishekofficial/mobile-price-classification