**PREDICTION OF MOBILE PRICES**

**PROJECT REPORT**

**COMP 7118 - DATA MINING**

**INSTRUCTOR**: **Dr. Xiaofei Zhang**

**Pooja Sree Gurrala**, **U00885267, pgurrala@memphis.edu**

**Department of Computer Science, University of Memphis, Memphis, TN**

**ABSTRACT**

Mobile phones have become an important part of communication, storing important information, entertainment etc. in our daily lives. As there are many different varieties of mobile phones available on the market, customers always face difficulty determining which mobile devices satisfies their requirements and their budget. This study aims to predict the price of mobile phones according to the specifications using Logistic Regression, K-Nearest Neighbor, Random Forest, and Decision Tree techniques. The performance of the models is evaluated using metrics like Accuracy, Confusion Matrix, Precision, Recall, F1-score etc.

**1)INTRODUCTION**

**1.1 THE DATASET**

For this project, Mobile Price Classification dataset was taken from an open-source platform Kaggle. This dataset consists of data organized into rows and columns. It consists of 2000 samples. It has 21 attributes in which 19 columns have “integer” data type and remaining two have data type as “float”.

**URL**: [Mobile Price dataset](https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification)

**Attributes:**

* Battery power – total energy a battery can store. Measured in mAh.
* Blue – mobile has Bluetooth option or not.
* Clock speed – The speed at which the microprocessor executes the instructions.
* Dual sim – Mobile has dual sim option or not.
* Fc – It is the mega pixels of front camera.
* four\_g – mobile is 4G or not.
* Int memory – The internal memory in gigabytes.
* m\_dep – Mobile depth in centimeters
* mobile\_wt – The weight of mobile phone
* n cores – The number of cores of processor
* pc – Thr mega pixels of primary camera.
* px\_height – Pixel resolution height
* px\_width – Pixel width resolution
* Ram – Random access memory in Megabytes (MB)
* sc\_h – Screen height of mobile in centimeters.
* sc\_w – Screen width of mobile in centimeters
* Talk time – The longest time a battery can last.
* Three\_g – Mobile is 3G or not.
* Touch screen – Mobile has touch screen or not.
* Wi-Fi – Mobile has Wi-Fi option or not.
* Price range – the price range of a mobile phone divided into four classes namely 0,1,2,3.

**1.2 PROBLEM STATEMENT**

Mobile Price Prediction is a real-world application in the mobile phone industry, where the manufacturers, retailers and consumers can benefit from accurate price range predictions.

The task can be framed as a “**supervised multi class classification problem**”, where the target variable is “price range”.

**1.3 CHALLENGES**

The prediction of mobile phones is challenging, due to the complexity of analysis and the number of features to consider that can affect prices.

To predict the prices, it is necessary to examine these features, so feature selection must be done.

**2) SOLUTION**

The solution is to design a Machine Learning algorithm which can analyze various factors and their interactions over time to identify patterns and trends.

**2.1 SUPERVISED LEARNING**

It is a type of Machine learning technique, where the labeled data will have input and output pairs. The objective is to find a mapping from inputs to outputs based on the patterns contained in the labeled data, so that it can correctly make predictions or classification on testing data.

**2.2 ALGORITHMS**

* **Logistic Regression**

It is a type of supervised learning where the objective is to predict the classes or categories based on the input features. The properties of the input are used to model the likelihood that an input belongs to a specific class.

Graphical user interface, text, application

Description automatically generated

Solver: The optimization procedure that is used to identify the coefficients of the logistic regression model that best fits the training data is specified by this parameter.

Max\_iter: Maximum number of iterations.

* **K-Nearest Neighbor**

It is mainly used for classification and regression tasks. It is an instance-based learning approach which is commonly known as lazy learning in which the algorithm stores the input features and their matching labels during training. For predicting unseen data, it finds k-closest neighbors using Manhattan distance or Euclidean distance and predicts the output on how similar the input instances are to the training data.

Graphical user interface, text, application

Description automatically generated

n\_neighbours: Number of neighbors to be considered for a datapoint.

* **Random Forest**

Random Forest constructs several decision trees by randomly choosing subset of characteristics for each tree and training them on various subsets of data via replacement. As a result of each tree being trained on a slightly different subset of features and data, this adds randomness and diversity. For classification - a majority vote, for regression – an average is taken to combine all the individual trees in the predictions to produce the result.

Graphical user interface, text, application

Description automatically generated

max\_depth: Maximum depth for a tree

random\_state: To generate the same number of random numbers, which results in same sequence of decision trees.

* **Decision Tree**

It is a tree-like structure used for classification and regression task, where internal nodes reflect decisions based on feature values, leaf modes represent expected outcomes. The algorithm takes decisions at each internal node to direct the data to expected outcome by recursively dividing the data into subsets depending on the values of a chosen feature at each internal node.

Graphical user interface, text, application

Description automatically generated

random\_state: same random numbers will be generated each time the model is trained, resulting in same decision tree.

**2.3 ASSUMPTIONS**

**LINEARITY**

It represents the relationship between features and target variable to be linear.

In the dataset, RAM has a linear relationship with price.

As RAM increases, price increases.

**2.4 METHODOLOGY**

**2.4.1 DATA PREPROCESSING**

**Feature Selection:**

The dataset has many unnecessary attributes which decreases the effectiveness of machine learning algorithms. So, feature selection is a method where attributes are removed which have less correlation with the target attribute. It must be done to improve the accuracy of the models.

From the dataset, clock speed, n cores, m\_dep are removed.

After feature selection, inconsistencies are removed from the dataset.

Table

Description automatically generated

**2.4.2 MODEL IMPLEMENTATION**

After pre-processing the data, out of the 2000 samples of data, randomly selected 80% of the samples are taken for training the model and the remaining 20% of the samples are taken for testing the model. Using the training data, supervised machine learning algorithms i.e., Linear Regression, K-Nearest Neighbor, Random Forest, Decision Tree is used to fit the model.

**2.4.3 MODEL EVALUATION**

Using 20% of testing data, price range values are predicted by the models. Among the four implemented models KNN is having highest accuracy compared to other models.

**3) EMPIRICAL EXPERIMENTS**

**Graphical user interface, application

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Qr code

Description automatically generated**

**Chart, bar chart

Description automatically generated**

Following are the classification reports and confusion matrix for

1. **Logistic Regression**

**Table

Description automatically generatedChart

Description automatically generated**

1. **K-Nearest Neighbor**

**Table

Description automatically generatedChart

Description automatically generated**

1. **Random Forest**

**Table

Description automatically generatedChart

Description automatically generated**

1. **Decision Tree**

**Table

Description automatically generatedChart, bar chart

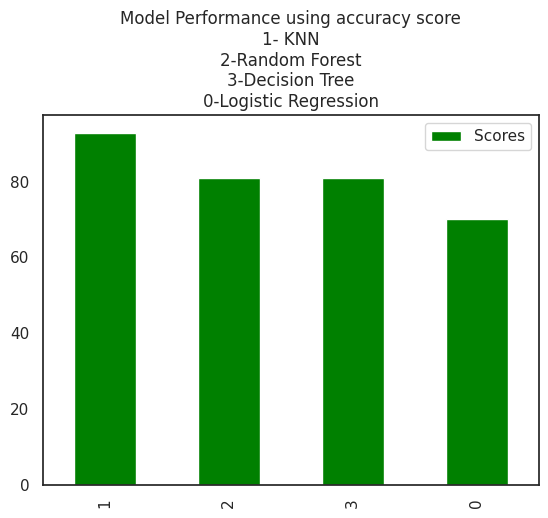
Description automatically generated**

**3.1 OBSERVATIONS**

Below screenshot represents the accuracy scores of the models.

Graphical user interface, text, application, chat or text message

Description automatically generated



**4) INTERESTING CONSIDERATIONS**

**USER-FRIENDLY**

The implemented models are simple. Users can easily understand the model implementations and acquire useful information from them.

**TIME AND SPACE EFFICIENCY**

The models are trained in less time depending on any parameter configurations. Along with this, it is effective in space, as it uses less memory.

**5) LIMITATIONS**

Among the implemented four models, KNN is having highest accuracy. Some more Machine Learning models can be implemented on the dataset to increase accuracy. Each algorithm has its own pros and cons. However, the selection of algorithm depends on specific requirements of the problem.

**6) CONCLUSION**

* Predicting mobile prices is a challenging task as it requires careful data analysis and the use of appropriate ML algorithms.
* It is also important to validate accuracy of models using performance metrics.
* It helps customers to analyze various features of mobiles phones to decide and businesses to understand how the features impact the price range of mobiles and make profit.
* Out of 4 algorithms used i.e., Logistic Regression, K-Nearest Neighbor, Random Forest, Decision Tree, KNN is having highest accuracy compared to other algorithms.
* At last, I can conclude that KNN is the best algorithm for mobile price dataset used in this project.

I have uploaded my project to GitHub, following is the URL,

**7) REFERENCES**

1. Mustafa Cetin, Yunus Koc. "Mobile Phone Price Class Prediction using different Classification Algorithms with Feature Selection and Parameter Optimization" IEEE 2021.
2. Ningyuan Hu. “Classification of Mobile Price Dataset using Machine Learning Algorithms”. IEEE 2022.