DATA PREPROCESSING: CASE STUDY ON MOBILE PRICE CLASSIFICATION DATASET

Ahmad Bulya Hakimi Bin Safawi[1] , Mohamed Tariq Ziyad Bin Mohamed Jahangir[1] and Muhammad Muaz Husaini Bin Rosli[1]

[1] The National University of Malaysia (UKM), Selangor, Malaysia

Abstract. In order to guarantee accurate and trustworthy findings, the data preparation stage of the mobile phone pricing categorization process is an essential step. The process of collecting data entails accumulating pertinent information on mobile phone costs and characteristics from reputable sources. The data collected came from the website kaggle. Data exploration is performed on the data in order to gain a deeper understanding of both the data and its structure. The next step in data cleaning is dealing with missing values, noise, inconsistent data, deleting duplicates, confirming labels, and keeping the formatting consistent. After that, the information goes through a process called feature engineering, which is a vital part of the process of classifying the prices of mobile phones. This process involves choosing and generating relevant features that capture significant phone attributes. The use of feature scaling strategies helps to eliminate bias, going through the data reduction process by placing all features on a scale that is comparable to one another. The accuracy and dependability of mobile phone pricing categorization models are improved when effective data preparation is performed. This is to the benefit of both consumers and companies, as it enables them to make more informed decisions in the highly competitive mobile phone market.

Keywords: Mobile Price, Classification.

1 Introduction

1.1 Definition

A mobile phone, also known as a cell phone or smartphone, is a wireless communication device that is portable. It uses a cellular network to allow users to make and receive calls, text messages, access the internet, and use various applications.

Mobile phones have transformed the way people communicate, providing a convenient and accessible way to stay in touch with others. They have become an essential component of modern society, with billions of people around the world relying on them for personal and professional communication.

The first mobile phones, which were primarily used for voice calls, were introduced in the early 1980s. However, technological advancements quickly

transformed mobile phones into sophisticated devices with a wide range of capabilities. Smartphones today have high-resolution touchscreens, powerful processors, and advanced operating systems that support a wide range of applications.

Beyond communication, these devices provide internet browsing, social media access, email services, GPS navigation, multimedia playback, and an ever-expanding universe of mobile apps. Mobile phones have evolved from productivity tools to entertainment platforms, catering to a wide range of needs and preferences.

1.2 **Mobile Phone Market Price**

The current market outlook for mobile phone prices is that it will continue to rise. The cost of high-end smartphones continues to increase, owing to the introduction of new features and technologies such as 5G connectivity, foldable screens, and advanced camera systems. Furthermore, global supply chain disruptions caused by the COVID-19 pandemic resulted in shortages of critical components, resulting in higher prices for many electronics, including smartphones.

However, there are more affordable options available, particularly in the mid-range segment, which offer many of the same features as high-end devices at a lower cost. Many manufacturers, including Xiaomi, Oppo, and Realme, are focusing on this market segment and releasing new models at competitive prices.

1.3 **Domain Problem**

A mobile company's primary goal is to establish itself as a strong competitor in the mobile phone market by providing high-quality devices at competitive prices. We intend to use data analysis to gain insight into the relationship between various mobile phone features and their selling prices. This will assist us to make well-informed decisions about pricing strategies for the company's mobile phones by understanding this relationship.

The key inquiry of the mobile company revolves around understanding the correlation between the features of a mobile phone and its selling price. To address this inquiry we must consider information such as features of a mobile phone that have a substantial impact on its selling price, do different combinations of features influence the price of a mobile phone and are there any discernible feature thresholds or trends that significantly impact pricing.

By gaining valuable insights into this information, we will be able to make data-driven decisions about the pricing of the company's mobile phones. This will allow us to position the products competitively in the market, taking into account the value offered by various features to potential customers.

1.4 **Dataset**

The dataset that will be incorporated in this paper is procured from kaggle's repository, titled Mobile Price Classification where it contains 21 columns of features and 2000 rows of data. The features includes 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 and price_range. The values range from numeric to categorical. Further information regarding the dataset is discussed in section 3.1 where we conduct data exploration of the dataset.

2 Related Work

Mobile phones have become an integral part of our daily lives, and with the rise in the number of mobile phone models available in the market, it has become increasingly challenging for consumers to select the right device that fits their needs and budget. The variety of features, specifications, and prices of mobile phones make it a complex task for users to compare different models and make an informed purchase decision. As a result, mobile phone price classification has emerged as an important research topic in recent years. Researchers have proposed several approaches to tackle the mobile phone price classification problem. The approaches range from traditional machine learning-based methods to deep learning-based methods, including hybrid approaches that combine machine learning and deep learning techniques. These approaches leverage various features and specifications of mobile phones, such as screen size, camera quality, battery capacity, RAM, storage, processor, and others, to classify mobile phones into different price ranges. In recent years, several studies have been conducted in the field of mobile phone price classification, and the results obtained have shown that these approaches are promising and effective in classifying mobile phones based on their prices. These studies have used different datasets of mobile phones, including publicly available datasets and datasets collected by the researchers themselves, to train and evaluate their models. The studies have also compared the performance of different machine learning and deep learning algorithms and have highlighted the strengths and limitations of each approach. Moreover, some studies have proposed novel approaches, such as transfer learning-based methods, which leverage pre-trained models to classify mobile phones based on their features and specifications. In this section, discuss five previous studies related to mobile phone price classification that have been conducted between the years 2019 and recent times, highlighting their contributions, strengths, and limitations. The review will provide a comprehensive understanding of the current state-of-the-art in mobile phone price classification and will help researchers and practitioners in the field to develop more effective approaches to tackle this problem.

The first study, Nasser et al. (2019) presents a study on the development of an Artificial Neural Network (ANN) model for predicting the price range of a mobile phone. The model was developed using a dataset consisting of mobile phone information, which contained a number of factors that influence the classification of mobile phone price. The ANN model was trained and validated using the dataset, and

the results obtained showed that the ANN model was able to correctly predict the mobile price range with 96.31 accuracy. In order to develop the ANN model, a Multilayer Perceptron Topology was used. It was composed of an input layer, a hidden layer and an output layer. The input layer contained 20 attributes, including battery power, CPU clock speed, has dual sim support or not, Front Camera megapixels, has 4G or not, has Wi-Fi or not, etc. The output variable represented the predicted price range based on those inputs. The learning algorithm was used to determine the importance of the input variables. Once the training process was completed, the network was tested using a test dataset without output variable results. The test data evaluation showed that the ANN model was able to correctly predict the mobile price range with 96.31 accuracy. In conclusion, this study demonstrated the ability of the artificial neural network to predict mobile phone price range. The ANN model was able to correctly predict the mobile phone price range with 96.31 accuracy, thus showing the potential of the ANN model to be used in predictive modelling.

Next, Güvenç and Koçak (2021), examines the classification performance of the KNN and DNN models for the mobile phone price classification problem. The data set used in the study is the "Mobile Price Classification" training and test dataset, which includes the features of mobile phones and gives the prices of the phones classified according to these features. The main purpose of the classification process is to predict what an image is, or to determine what the available data means in a data set with a large number of different categories of data. At the application stage of the study, the DNN model was trained for different possible values of these parameters and values with the best classification performance were given in the study. Categorical cross entropy was used as a loss function in the study. The fl-score score was considered as a priority to decide which model had the best classification performance. At the end of the study, both models were given test data that was not labelled with price data, and the classification estimates made by the models were compared. The comparison of the classification performances of the KNN and DNN models revealed that the DNN model had better classification performance compared to the KNN model. The DNN model had an average accuracy of 98.13%, while the KNN model had an average accuracy of 88.63%. Finally, the estimation of mobile phone price ranges was made by giving unlabeled test data to both models. The results showed that the DNN model was able to estimate the price ranges of the mobile phones more accurately than the KNN model. This high performance showed that DNN classifiers are a strong model, taking into account the results given in the literature in general.

Kalaivani et al. (2021) discusses the development of a machine learning model to predict the price range of a mobile phone. The mobile phone dataset used in the study was obtained from Kaggle platform which contained 21 features related to the key features of the mobile phone. To improve the model performance, the authors applied a feature selection technique to select the most relevant features and used the top 10 features to train the model. The selected features are RAM, pixel height, battery power, pixel width, mobile weight, internal memory, screen width, talk time, front camera and screen height. Three different machine learning algorithms were used to

predict the price range of the mobile phone, namely, Support Vector Machine (SVM), Random Forest Classifier (RFC) and Logistic Regression. The performance of the model was evaluated based on accuracy and Chi-Square statistical test was used to further improve the accuracy of the model. The experimental results showed that SVM provided the best performance with an accuracy of 97%, followed by RFC with an accuracy of 87% and Logistic Regression with an accuracy of 81%. In conclusion, the study showed that the use of machine learning algorithms and feature selection techniques can be used to accurately predict the price range of a mobile phone.

The study conducted by Hu (2022), explores the use of machine learning algorithms to predict the price of mobile phones based on a range of different features. The dataset used for this research was obtained from Kaggle and four relevant features, including ram, battery power, px_width and px_height, were selected. Four machine learning algorithms were then employed to fit the training dataset of mobile price and make a prediction on the price level, with the performance being recorded and appraised by accuracy, precision, recall and F1 score. The first step in the research was data preprocessing, which involved reading the information about the data and employing a data processing method to identify any missing values and different features. The second step was feature selection, which is an important procedure before applying the machine learning algorithms. This process helps to reduce overfitting, improve accuracy and reduce training time by finding the most significant features out of a large feature set. The fourth step was to classify the price level and make predictions using the four chosen machine learning algorithms. Support Vector Machine (SVM), Decision Tree (DT), K Nearest Neighbors (KNN) and Naïve Bayes (NB) classifiers were all used to predict the price level and the results of each algorithm were compared. Overall, SVM performed the best, with an increase in performance when the process of feature selection was applied. The study demonstrates the way to classify and predict the mobile phone price level using four different machine learning algorithms. It also reveals the correlations between the functions of mobile phones and mobile price levels, allowing consumers to make rational consumption decisions in the future. Furthermore, the results could help people to do cost minimization when choosing mobile phones at different price levels.

Kalmaz and Akin (2022), present a study on the estimation of mobile phone prices using machine learning algorithms. The authors conducted experiments to test and evaluate 25 different machine learning algorithms on a dataset gathered from Kaggle. The algorithms were evaluated using four measures: accuracy score, MSE, MAE and R-Squared. The results showed that the SVC algorithm was the most effective for predicting phone prices with an accuracy rate of 0.9470. The authors also analysed the most important features that affect phone prices and concluded that battery power, bluetooth, clock speed, dual sim, front camera megapixels, 4G, internal memory in gigabytes, mobile depth in cm, weight of mobile phone, number of cores of processor, primary camera megapixels, pixel resolution height, pixel resolution width, random access memory in megabytes, screen height of mobile in cm, screen width of mobile in cm, longest time that a single battery charge will last, 3G, touch screen, wifi are critical features that can influence phone prices. In addition, the

authors proposed a Confusion Matrix to visualise and summarise the performance of the classification algorithm used to predict phone prices. Overall, this paper provides a comprehensive analysis of the effectiveness of various machine learning algorithms for predicting phone prices. The results demonstrate that SVC is the most accurate and reliable algorithm for predicting phone prices. The authors also provide useful insights into the most important features that affect phone prices, which can be utilised to further improve the accuracy of the machine learning algorithms.

In conclusion, the task of mobile phone price classification has gained significant attention in recent years due to the rapid growth of the mobile phone market. Various approaches have been proposed to tackle this task, ranging from traditional machine learning techniques to deep learning models. Most of the previous studies have focused on feature engineering and model optimization, while some have also considered the impact of external factors such as brand reputation and customer reviews. Additionally, the availability of large-scale mobile phone datasets has facilitated the development of more accurate and robust models. Despite the promising results achieved by previous approaches, there is still room for improvement in mobile phone price classification. Future research can explore the potential of incorporating more diverse features, such as user behaviour and demographic information, as well as integrating different data sources to enhance the model's performance.

3 Material and Methods

3.1 Data Exploration

3.1.1 Data Quality Report

In order to enhance our understanding of the dataset that we incorporate in this paper, we must first evaluate what each feature represents in the data. As stated previously the data consists of 21 features including the target class which is price_range. The table below describes the data type and details of each of the 21 features within the dataset.

Table 1. Dataset attributes and details.

Attributes	Data Type	Details
battery_power	Numeric	Total energy a battery can store in one time measured in mAh
blue	Binary	Whether or not the smartphone has bluetooth
clock_speed	Numeric	Speed at which microprocessor executes instructions
dual_sim	Binary	Whether or not the smartphone has dual sim support
fc	Numeric	Front Camera megapixels

four_g	Binary	Whether or not the smartphone has 4G
int_memory	Numeric	Internal Memory in Gigabytes
m_dep	Numeric	Mobile Depth in cm
mobile_wt	Numeric	Weight of mobile phone
n_cores	Numeric	Number of cores of processor
pc	Numeric	Primary Camera megapixels
px_height	Numeric	Pixel Resolution Height
px_width	Numeric	Pixel Resolution Width
ram	Numeric	Random Access Memory in MegaBytes
sc_h	Numeric	Screen Height of mobile in cm
sc_w	Numeric	Screen Width of mobile in cm
talk_time	Interval	Longest time that a single battery charge will last when you are talking
three_g	Binary	Whether or not the smartphone has 3G
touch_screen	Binary	Whether or not the smartphone has touch screen
wifi	Binary	Whether or not the smartphone has wifi
price_range	Ordinal	This is the target feature with values of 0 (low cost), 1 (medium cost), 2 (high cost) and 3 (very high cost)

It is important to take note which features fall within the continuous and categorical data type. This is to ensure proper analysis is conducted for each type. For instance a feature that has a continuous data type may benefit from distribution measurements such as mean and median. However, categorical features have no benefit from such measurements and must be treated differently. As such for the purpose of exploring our dataset we have separated the continuous and categorical features of the data. The figure below is a sample of the first five columns of the dataset.

	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	842	0	2.2	0	1	0	7	0.6	188	2	2	20	756	2549	9	7	19	0	0	1	1
1	1021	1	0.5	1	0	1	53	0.7	136	3	6	905	1988	2631	17	3	7	1	1	0	2
2	563	1	0.5	1	2	1	41	0.9	145	5	6	1263	1716	2603	11	2	9	1	1	0	2
3	615	1	2.5	0	0	0	10	0.8	131	6	9	1216	1786	2769	16	8	11	1	0	0	2
4	1821	1	1.2	0	13	1	44	0.6	141	2	14	1208	1212	1411	8	2	15	1	1	0	1

Figure 1. Sample of the dataset

Upon separation it is observed that there are 14 features that fall within the continuous data type. All continuous features are accounted for as the count value for all features is 2000 which corresponds to the initial rows of data, thus there is no missing data. The figure below provides insight on the measures of central tendency for the continuous features.

	Count	% Miss.	Card.	Mean	Std. Dev.	Min.	1st Qrt.	Median	3rd Qrt.	Max.
battery_power	2000	0.0	1094	1238.51850	439.418206	501.0	851.75	1226.0	1615.25	1998.0
clock_speed	2000	0.0	26	1.52225	0.816004	0.5	0.70	1.5	2.20	3.0
fc	2000	0.0	20	4.30950	4.341444	0.0	1.00	3.0	7.00	19.0
int_memory	2000	0.0	63	32.04650	18.145715	2.0	16.00	32.0	48.00	64.0
m_dep	2000	0.0	10	0.50175	0.288416	0.1	0.20	0.5	0.80	1.0
mobile_wt	2000	0.0	121	140.24900	35.399655	80.0	109.00	141.0	170.00	200.0
n_cores	2000	0.0	8	4.52050	2.287837	1.0	3.00	4.0	7.00	8.0
pc	2000	0.0	21	9.91650	6.064315	0.0	5.00	10.0	15.00	20.0
px_height	2000	0.0	1137	645.10800	443.780811	0.0	282.75	564.0	947.25	1960.0
px_width	2000	0.0	1109	1251.51550	432.199447	500.0	874.75	1247.0	1633.00	1998.0
ram	2000	0.0	1562	2124.21300	1084.732044	256.0	1207.50	2146.5	3064.50	3998.0
sc_h	2000	0.0	15	12.30650	4.213245	5.0	9.00	12.0	16.00	19.0
sc_w	2000	0.0	19	5.76700	4.356398	0.0	2.00	5.0	9.00	18.0
talk_time	2000	0.0	19	11.01100	5.463955	2.0	6.00	11.0	16.00	20.0

Figure 2. Quantitative feature

Consequently for the categorical features there are a total of seven features. All of the features have a cardinality value of two which are binary labels of either yes or no except for the “price_range” feature which consists of four unique values as it is the class label which classifies the price range of low cost, medium cost, high cost and very high cost. As with the continuous features, the categorical features are all accounted for as each one has a count value of 2000. Thus there is no missing data. The figure below provides insight on the qualitative features of the dataset.

	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
blue	2000	0.0	2	0	1010	50.50	1	990	49.50
dual_sim	2000	0.0	2	1	1019	50.95	0	981	49.05
four_g	2000	0.0	2	1	1043	52.15	0	957	47.85
three_g	2000	0.0	2	1	1523	76.15	0	477	23.85
touch_screen	2000	0.0	2	1	1006	50.30	0	994	49.70
wifi	2000	0.0	2	1	1014	50.70	0	986	49.30
price_range	2000	0.0	4	0	500	25.00	1	500	25.00

Figure 3. Qualitative feature

3.1.2 **Visualisation**

From the histogram that will be presented below, the qualitative data only have a cardinality of two, with the exception of the "price_range" variable, which has a cardinality of four due to the fact that it contains four different groups. This can be seen in the figure 4 that is situated below. The quantitative data on the other hand, either have a high cardinality or a high amount of skewness. After this section, the data in question will be handled by the section devoted to data cleaning.

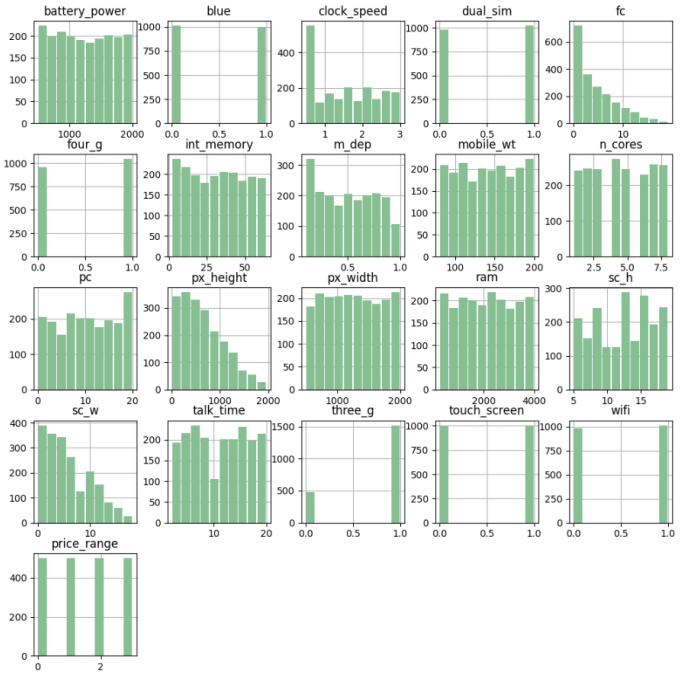

Figure 4. Histogram of features before data cleaning

As can be observed in the figure 5 that is located below, there are just four features that are widely distributed, and these are "battery_power," "px_height," "px_width," and "ram." It has been brought to our attention that the "px_height" feature has an outlier. After giving it a lot of thought and doing some research on some of the most popular phone models, we came to the conclusion that the highest possible value for that function is really within the typical range for mobile phones in this day and age.

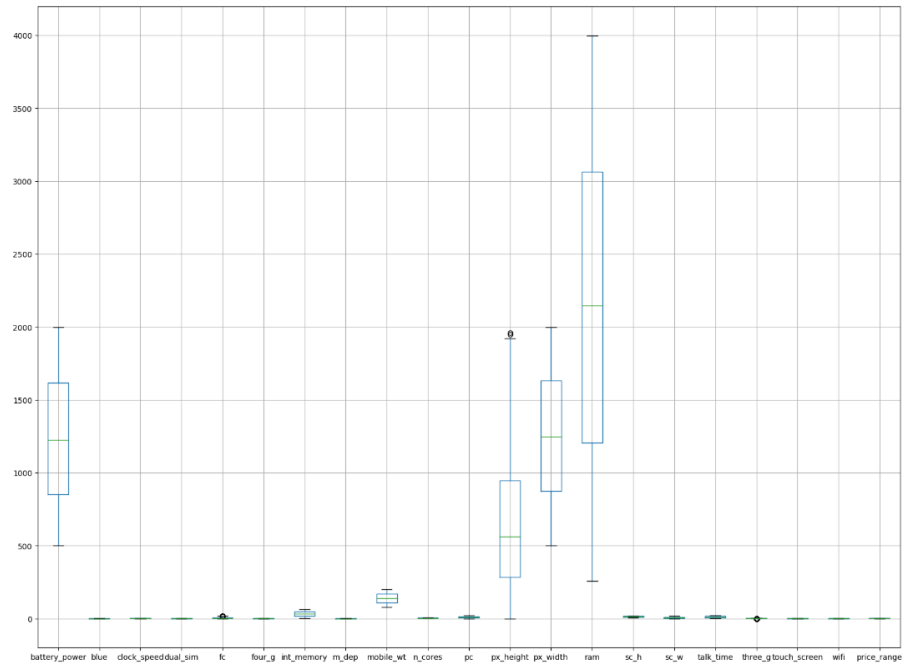

Figure 5. Boxplot of features before data cleaning

There are a few characteristics on the heat map that clearly show an association, and they can be seen in the picture that is found up below, which is referred to as figure 6. Both the "pc" and "fc" features have a strong positive link with one another. Aside from that, "ram" and "price_range" are the ones that have the most association with each other. This can be demonstrated using reason and logic given that the price of a mobile phone would steadily go up as the amount of RAM it carries increases.

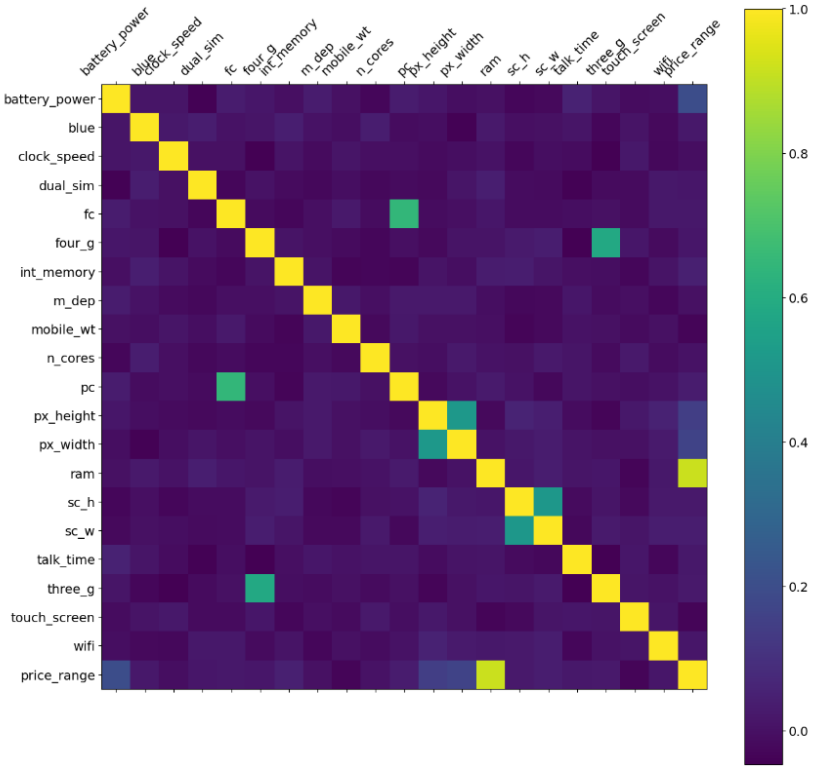

Figure 6. Correlation heatmap of features before data cleaning

The scatter plot matrix below contains the distribution of any two intersected features. This is very helpful to see the big picture on their correlation especially on the positivity or the negativity relationship.

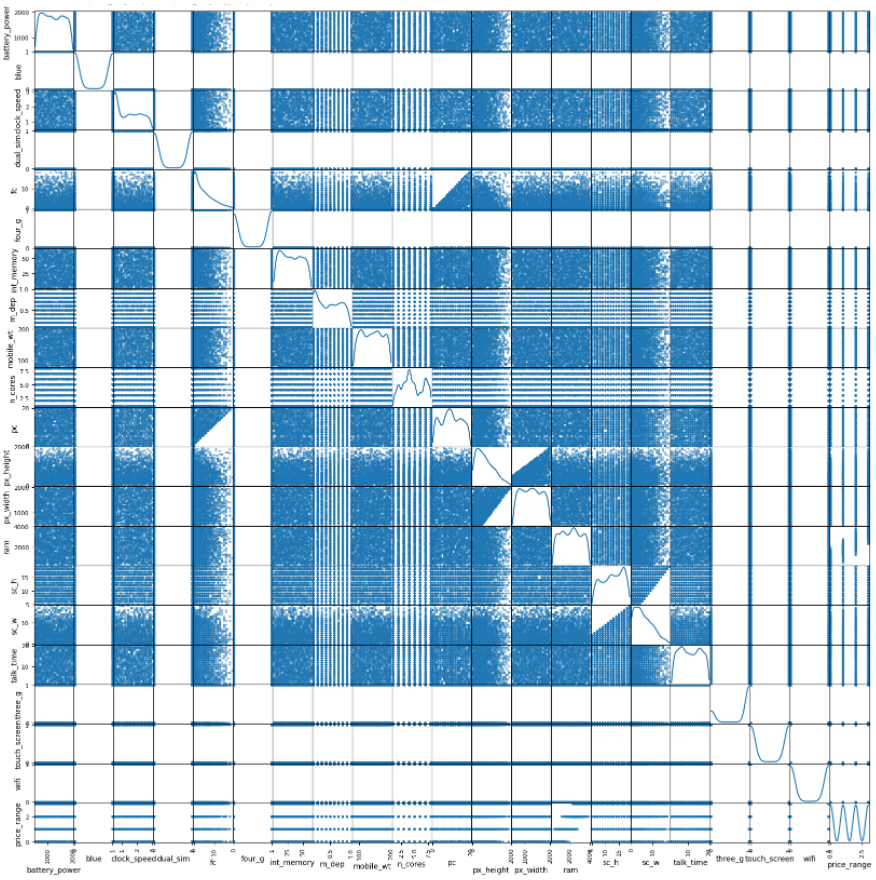

Figure 7. Scatter plot matrix of features before data cleaning

3.2 Data Cleaning

3.2.1 Handling Missing Values

In the context of data, missing values refer to the absence of a particular value or information in a dataset. When data is collected or recorded, it is expected that every variable or feature has a corresponding value for each observation or data point. However, missing values occur when no data or information is available for a particular variable for one or more observations.

Missing values can occur for various reasons, including human error during data collection, malfunctioning equipment, non-response in surveys or questionnaires, data corruption, or intentional omission. For example, a survey participant may choose not to answer a specific question, resulting in a missing value for that particular variable. Missing values can be represented in different ways, depending

on the data format and the data analysis software being used. Common representations of missing values include blank cells, placeholders like "N/A" or "NaN" (Not a Number), or specific codes assigned to indicate missingness, such as "-99" or "-9999".

Handling missing values is an important aspect of data analysis and requires careful consideration. Ignoring missing values or treating them improperly can lead to biassed or inaccurate results. Therefore, appropriate techniques for handling missing values, such as deletion, imputation, or advanced statistical methods, should be applied to ensure robust and valid data analysis.

Fortunately, since the dataset under consideration is complete and devoid of any missing values, there is no need to employ any specific techniques or methodologies to handle the absence of data points. This absence of missing values simplifies the data analysis process, allowing for a more streamlined and straightforward exploration of the dataset and enabling the application of various data mining techniques to extract valuable insights.

3.2.2 **Handling Noise**

Noisy data refers to the irrelevant or erroneous data that may be present in a dataset. Noise data can occur due to various reasons such as measurement errors, data collection errors, data entry errors, or even natural variations in the data. Noise data can negatively impact the accuracy and effectiveness of data mining algorithms and models. It can lead to misleading patterns, incorrect predictions, and reduced overall performance of data mining tasks. Noise data can manifest in different forms. It may include outliers, which are data points that significantly deviate from the normal distribution or pattern of the dataset. Outliers can be caused by measurement errors or rare occurrences in the data. Noise can also appear as missing values, inconsistent data formats, duplicate records, or contradictory information. Dealing with noise data is an important preprocessing step in data mining.

Frequency binning can be used in data mining to handle noise in datasets. The process involves dividing the data into frequency intervals or bins, and then applying noise reduction techniques to each bin individually. This helps to reduce the impact of noise on the overall dataset and improve the accuracy of data mining algorithms. In short, it is also known as discretization, which involves dividing a continuous feature into a set of discrete bins or intervals. The decision to apply a binning process to feature with many distinct values depends on the specific context and purpose of our analysis. Binning can simplify the analysis by reducing the number of distinct values to reduce cardinality. When dealing with a large number of distinct values, the dataset can become complex and challenging to work with. By grouping values into bins, there are broader categories that can be created to make it easier to handle and analyze the data. This simplification can make it more manageable to perform calculations, create visualizations, or derive insights from the data. Binning also can enhance the interpretability of the data. Patterns or trends observed within the bins can be highlighted. This makes it more accessible for a non-technical audience to understand and draw conclusions from the data.

Other than that, it is also essential to examine the distribution of the feature values. The distribution provides insights into how the values are spread across the range and can help to make informed decisions about whether binning is necessary or beneficial. Skewness measures the asymmetry of the distribution. A positively skewed distribution has a tail on the right side, while a negatively skewed distribution has a tail on the left side. Binning may be more relevant when dealing with skewed distributions, as it can help highlight patterns or outliers within specific ranges. However, if the distribution is approximately symmetric, binning may not provide significant benefits.

Within our dataset, it has been demonstrated that the vast majority of the quantitative data have more than ten distinct values or cardinality, with the exception of the attribute referred to as "n_cores" which only has eight distinct values. After additional evaluation of its skewness measurements, the attribute did not go through the binning process since its skewed distribution is fairly even and the measurement parameter is closer to 0 than it is to any other value. Aside from that, the binning procedure is applied to all of the other quantitative data. Consider the "battery_power" feature as a good example. When the characteristics are represented only by their individual values, such as 842, 1021, 563, 615, etc., it may be challenging to identify patterns in the data. However, once the feature has been classified into categories (such as low, medium, high, and very high), it is much easier to analyze and understand the data. Therefore, we would be able to make observations about patterns, such as which battery power levels are more likely to cause a mobile phone's price range to increase.

Through the process of binning, the data can be simplified, allowing for a greater focus on the general patterns and correlations that are more crucial to the analysis or decision-making procedure. To ensure that essential insights are not lost during the process of binning, it is necessary to strike a balance between simplifying and maintaining sufficient detail.

3.2.3 **Handling Inconsistent Data**

Inconsistent data can significantly impact the accuracy and reliability of data mining analyses. Data mining involves extracting meaningful patterns, relationships, and insights from large datasets, making data quality and consistency crucial. In this essay, we will delve into the concept of inconsistent data in data mining, its causes, and explore various strategies and best practices for handling such data effectively. Inconsistent data refers to conflicting or contradictory information within a dataset. It can manifest in different ways, including errors during data entry, integration of data from multiple sources, incomplete or missing values, redundancy, duplication, and inaccurate measurements. These inconsistencies make it challenging to derive meaningful patterns and relationships from the data, leading to unreliable analysis results.

Examples of Inconsistent Data:

1) Inconsistent spellings: Differently spelled names for the same individual, such as "John Smith" and "Jon Smit."
2) Conflicting values: Discrepancies in product prices, like $1000 and $1200 for the same item.
3) Missing attribute values: Absence of critical information, such as age or address, within certain records.
4) Inconsistent units of measurement: Varied units like kilograms and pounds used for the same attribute.
5) Duplicated records: Multiple entries for the same customer with slight variations in attributes.
6) Contradictory information: Respondents providing conflicting answers to similar survey questions.

Handling Inconsistent Data:

1) Data cleaning: Correct errors, standardize formats, remove duplicates, and address missing values using techniques like imputation or deletion.
2) Data integration: Reconcile differences and inconsistencies when merging data from multiple sources.
3) Data transformation: Convert units, scale values, or apply mathematical operations to normalize the data.
4) Outlier detection: Identify and handle extreme values that introduce inconsistencies.
5) Domain knowledge and expert judgment: Leverage expertise to make informed decisions during data preprocessing.
6) Iterative data exploration: Utilize visualizations, statistical analysis, and iterative techniques to identify patterns and refine preprocessing steps.

Handling inconsistent data is a critical step in data mining to ensure accurate and reliable analysis results. By employing strategies such as data cleaning, integration, transformation, outlier detection, and leveraging domain knowledge, researchers and analysts can address inconsistencies effectively, improving the quality and consistency of data. These best practices enhance the validity and reliability of data mining analyses, leading to more meaningful patterns, relationships, and insights that can drive informed decision-making and foster advancements in various fields.

Fortunately, in this particular scenario, there is no need to handle inconsistent data as the dataset provided is free from such issues. This means that the data is consistent and reliable, without any conflicting or contradictory information. Without the presence of inconsistencies, the data can be directly utilized for data mining and analysis purposes. This allows us to proceed with confidence, focusing on extracting meaningful patterns, relationships, and insights from the dataset, while being assured of the data's quality and consistency. By leveraging this consistent data, valuable insights can be derived, enabling informed decision-making and fostering advancements in this study.

3.3 **Data Reduction**

3.3.1 **Dimensionality Reduction**

The dataset might contain insignificant features that might not help improve the performance of Machine Learning models to be developed and instead undergo a training process for nothing or the worst case could reduce the performance. Dimensionality reduction refers to the selection of only a subset from all features to be trained such that the result still is close as possible to the original one (using all features).

In this project, we implemented attribute subset selection technique using Decision Tree Induction. This technique constructs a Decision Tree prior to the model development to obtain the feature importances of each feature. A threshold of 0.035 has been chosen to filter these features. Features with importance lower than

the threshold will not be used in developing model. After filtering out them, we have 12 features left which are around 60% from all features. The table below shows selected features with their respective importances ascendingly.

Table 2. Feature Importances.

No	Feature	Feature Importance
1	ram	0.3544
2	battery_power	0.0641
3	n_cores	0.0483
4	clock_speed	0.0403
5	px_width	0.0396
6	px_height	0.0391
7	int_memory	0.0367
8	m_dep	0.0365
9	pc	0.0364
10	sc_w	0.0359
11	mobile_wt	0.0357
12	fc	0.0353

3.3.2 **Numerosity Reduction**

Numerosity reduction refers to alternative or smaller forms of data representation of all observations in a dataset. It can be beneficial in situations when the volume of the dataset is very large or when the dataset contains a large amount of irrelevant or duplicated observations. However, since the dataset is considered as a small dataset (contains only 2000 observations), numerosity reduction is not required.

3.3.3 **Data Transformation**

The last step in preprocessing the data is to transform. Data transformation has been done by normalizing the data. Normalization is a technique used to adjust the distance between data points to a common scale. This helps in data analysis and modeling as well as reducing the impact of different scales on the accuracy of machine learning models.

In this project, we use min-max normalization. This is the most basic type of normalization. It gives the minimum value a value of 0, maximum value a value of 1 and the other values between them accordingly. The result of data transformation will be discussed in subtopic 4.2.4.

4 Results of Data Preprocessing

4.1 Histogram of data distribution

The histograms below shows the distributions of data points for each of 21 features before data reduction. We can observe a lot of skewness and noise in them.

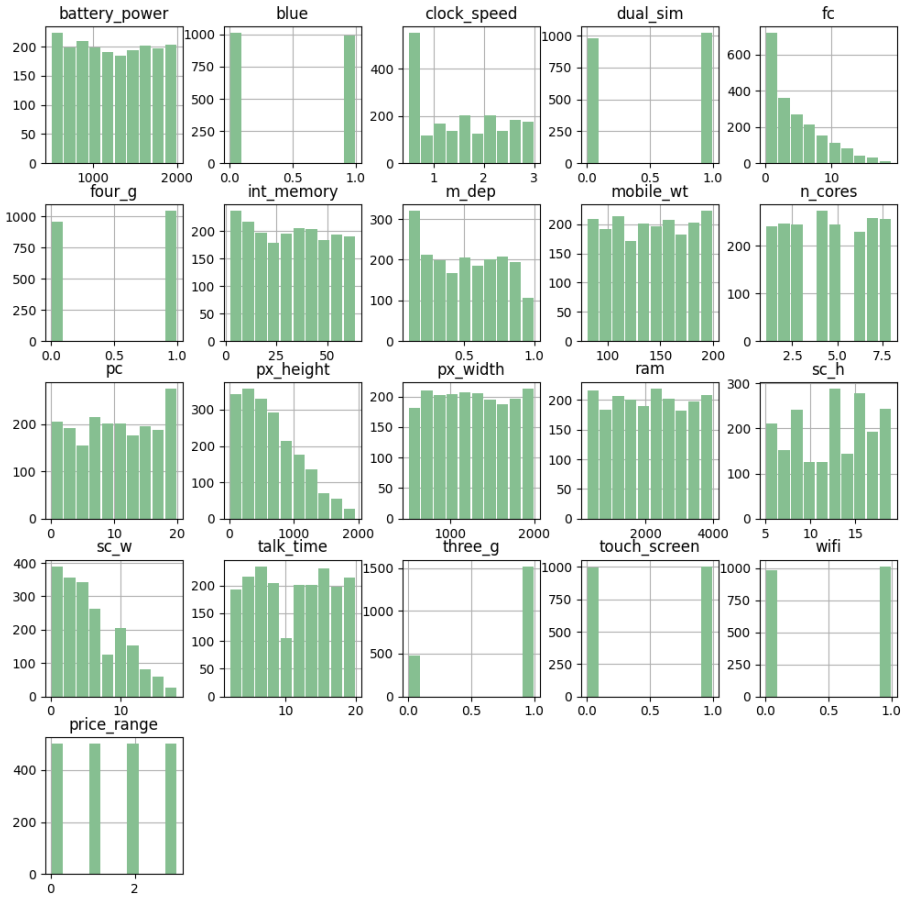

Figure 8. Histogram of features before data reduction

The next 13 histograms shows the distribution of data points after reduction. We can observe that there are now fewer bins with almost equal depth among them for each of the features. Note that the less important features also has been removed.

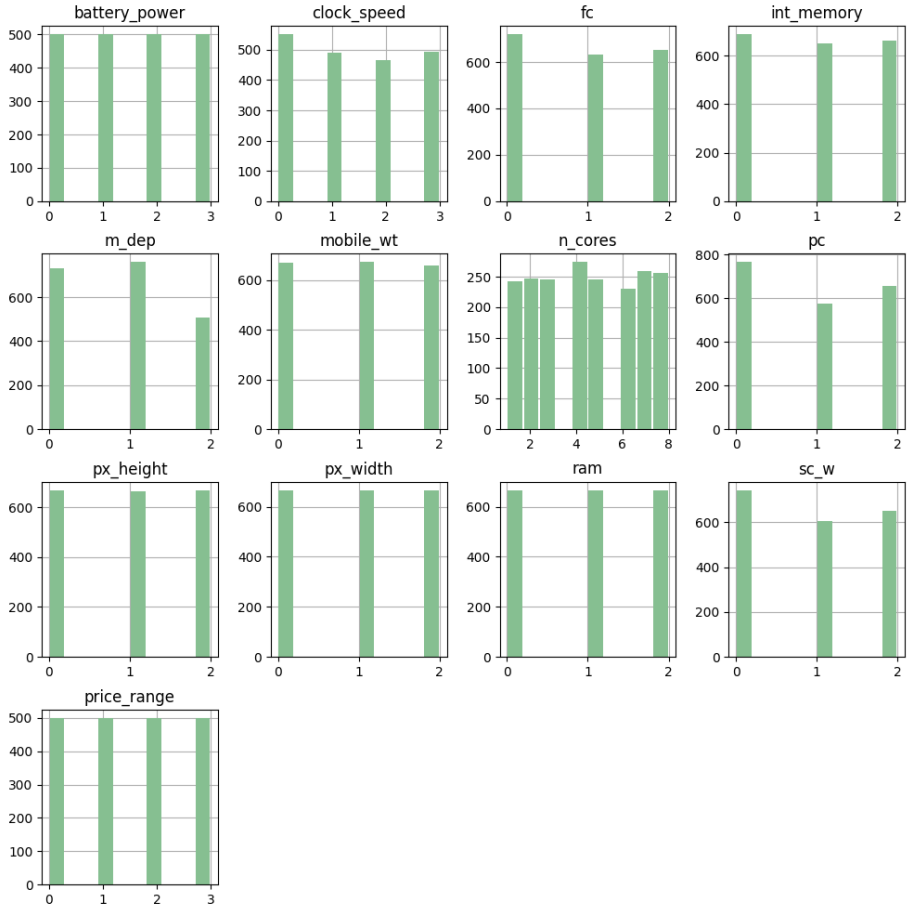

Figure 9. Histogram of features after data reduction

4.2 **Data**

4.2.1 **Raw Data**

In the figure below, it shows the data when it was first obtained from the data source.

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
842	0	2.2	0	1	0	7	0.6	188	2	2	20	756	2549	9	7	19	0	0	1	1
1021	1	0.5	1	0	1	53	0.7	136	3	6	905	1988	2631	17	3	7	1	1	0	2
563	1	0.5	1	2	1	41	0.9	145	5	6	1263	1716	2603	11	2	9	1	1	0	2
615	1	2.5	0	0	0	10	0.8	131	6	9	1216	1786	2769	16	8	11	1	0	0	2
1821	1	1.2	0	13	1	44	0.6	141	2	14	1208	1212	1411	8	2	15	1	1	0	1
1859	0	0.5	1	3	0	22	0.7	164	1	7	1004	1654	1067	17	1	10	1	0	0	1
1821	0	1.7	0	4	1	10	0.8	139	8	10	381	1018	3220	13	8	18	1	0	1	3
1954	0	0.5	1	0	0	24	0.8	187	4	0	512	1149	700	16	3	5	1	1	1	0
1445	1	0.5	0	0	0	53	0.7	174	7	14	386	836	1099	17	1	20	1	0	0	0
509	1	0.6	1	2	1	9	0.1	93	5	15	1137	1224	513	19	10	12	1	0	0	0
769	1	2.9	1	0	0	9	0.1	182	5	1	248	874	3946	5	2	7	0	0	0	3
1520	1	2.2	0	5	1	33	0.5	177	8	18	151	1005	3826	14	9	13	1	1	1	3
1815	0	2.8	0	2	0	33	0.6	159	4	17	607	748	1482	18	0	2	1	0	0	1
803	1	2.1	0	7	0	17	1.0	198	4	11	344	1440	2680	7	1	4	1	0	1	2
1866	0	0.5	0	13	1	52	0.7	185	1	17	356	563	373	14	9	3	1	0	1	0
775	0	1.0	0	3	0	46	0.7	159	2	16	862	1864	568	17	15	11	1	1	1	0
838	0	0.5	0	1	1	13	0.1	196	8	4	984	1850	3554	10	9	19	1	0	1	3
595	0	0.9	1	7	1	23	0.1	121	3	17	441	810	3752	10	2	18	1	1	0	3
1131	1	0.5	1	11	0	49	0.6	101	5	18	658	878	1835	19	13	16	1	1	0	1
682	1	0.5	0	4	0	19	1.0	121	4	11	902	1064	2337	11	1	18	0	1	1	1

Figure 10. Raw data

4.2.2 **Cleaned Data**

The author of the data has actually cleaned the data beforehand. The data at this point should be incomplete. Although it was originally completed, it is still noisy and inconsistent.

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
842	0	2.2	0	1	0	7	0.6	188	2	2	20	756	2549	9	7	19	0	0	1	1
1021	1	0.5	1	0	1	53	0.7	136	3	6	905	1988	2631	17	3	7	1	1	0	2
563	1	0.5	1	2	1	41	0.9	145	5	6	1263	1716	2603	11	2	9	1	1	0	2
615	1	2.5	0	0	0	10	0.8	131	6	9	1216	1786	2769	16	8	11	1	0	0	2
1821	1	1.2	0	13	1	44	0.6	141	2	14	1208	1212	1411	8	2	15	1	1	0	1
1859	0	0.5	1	3	0	22	0.7	164	1	7	1004	1654	1067	17	1	10	1	0	0	1
1821	0	1.7	0	4	1	10	0.8	139	8	10	381	1018	3220	13	8	18	1	0	1	3
1954	0	0.5	1	0	0	24	0.8	187	4	0	512	1149	700	16	3	5	1	1	1	0
1445	1	0.5	0	0	0	53	0.7	174	7	14	386	836	1099	17	1	20	1	0	0	0
509	1	0.6	1	2	1	9	0.1	93	5	15	1137	1224	513	19	10	12	1	0	0	0
769	1	2.9	1	0	0	9	0.1	182	5	1	248	874	3946	5	2	7	0	0	0	3
1520	1	2.2	0	5	1	33	0.5	177	8	18	151	1005	3826	14	9	13	1	1	1	3
1815	0	2.8	0	2	0	33	0.6	159	4	17	607	748	1482	18	0	2	1	0	0	1
803	1	2.1	0	7	0	17	1.0	198	4	11	344	1440	2680	7	1	4	1	0	1	2
1866	0	0.5	0	13	1	52	0.7	185	1	17	356	563	373	14	9	3	1	0	1	0
775	0	1.0	0	3	0	46	0.7	159	2	16	862	1864	568	17	15	11	1	1	1	0
838	0	0.5	0	1	1	13	0.1	196	8	4	984	1850	3554	10	9	19	1	0	1	3
595	0	0.9	1	7	1	23	0.1	121	3	17	441	810	3752	10	2	18	1	1	0	3
1131	1	0.5	1	11	0	49	0.6	101	5	18	658	878	1835	19	13	16	1	1	0	1
682	1	0.5	0	4	0	19	1.0	121	4	11	902	1064	2337	11	1	18	0	1	1	1

Figure 11. Cleaned data

4.2.3 **Discretized Data**

The data points in each feature have been binned into fewer bins to reduce the cardinality. In this discretized data, we can observe the bin category or interval where each data point belongs to.

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
(500.0, 852.0]	No	(1.5, 2.2]	No	(-1.0, 1.0]	No	(1.0, 21.0]	(0.3, 0.7]	(160.0, 200.0]	2	(-1.0, 7.0]	(-1.0, 371.0]	(499.0, 1006.0]	(1470.0, 2711.0]	(4.0, 12.0]	(3.0, 7.0]	(11.0, 20.0]	No	No	Yes	medium cost
(852.0, 1226.0]	Yes	(0.4, 0.7]	Yes	(-1.0, 1.0]	Yes	(42.0, 64.0]	(0.3, 0.7]	(119.0, 160.0]	3	(-1.0, 7.0]	(795.0, 1960.0]	(1488.0, 1998.0]	(1470.0, 2711.0]	(12.0, 19.0]	(-1.0, 3.0]	(1.0, 11.0]	Yes	Yes	No	high cost
(500.0, 852.0]	Yes	(0.4, 0.7]	Yes	(1.0, 5.0]	Yes	(21.0, 42.0]	(0.7, 1.0]	(119.0, 160.0]	5	(-1.0, 7.0]	(795.0, 1960.0]	(1488.0, 1998.0]	(1470.0, 2711.0]	(4.0, 12.0]	(-1.0, 3.0]	(1.0, 11.0]	Yes	Yes	No	high cost
(500.0, 852.0]	Yes	(2.2, 3.0]	No	(-1.0, 1.0]	No	(1.0, 21.0]	(0.7, 1.0]	(119.0, 160.0]	6	(7.0, 13.0]	(795.0, 1960.0]	(1488.0, 1998.0]	(2711.0, 3998.0]	(12.0, 19.0]	(7.0, 18.0]	(1.0, 11.0]	Yes	No	No	high cost
(1615.0, 1998.0]	Yes	(0.7, 1.5]	No	(5.0, 19.0]	Yes	(42.0, 64.0]	(0.3, 0.7]	(119.0, 160.0]	2	(13.0, 20.0]	(795.0, 1960.0]	(1006.0, 1488.0]	(255.0, 1470.0]	(4.0, 12.0]	(-1.0, 3.0]	(11.0, 20.0]	Yes	Yes	No	medium cost
(1615.0, 1998.0]	No	(0.4, 0.7]	Yes	(1.0, 5.0]	No	(21.0, 42.0]	(0.3, 0.7]	(160.0, 200.0]	1	(-1.0, 7.0]	(795.0, 1960.0]	(1488.0, 1998.0]	(255.0, 1470.0]	(12.0, 19.0]	(-1.0, 3.0]	(1.0, 11.0]	Yes	No	No	medium cost
(1615.0, 1998.0]	No	(1.5, 2.2]	No	(1.0, 5.0]	Yes	(1.0, 21.0]	(0.7, 1.0]	(119.0, 160.0]	8	(7.0, 13.0]	(371.0, 795.0]	(1006.0, 1488.0]	(2711.0, 3998.0]	(12.0, 19.0]	(7.0, 18.0]	(11.0, 20.0]	Yes	No	Yes	very high cost
(1615.0, 1998.0]	No	(0.4, 0.7]	Yes	(-1.0, 1.0]	No	(21.0, 42.0]	(0.7, 1.0]	(160.0, 200.0]	4	(-1.0, 7.0]	(371.0, 795.0]	(1006.0, 1488.0]	(255.0, 1470.0]	(12.0, 19.0]	(-1.0, 3.0]	(1.0, 11.0]	Yes	Yes	Yes	low cost
(1226.0, 1615.0]	Yes	(0.4, 0.7]	No	(-1.0, 1.0]	No	(42.0, 64.0]	(0.3, 0.7]	(160.0, 200.0]	7	(13.0, 20.0]	(371.0, 795.0]	(499.0, 1006.0]	(255.0, 1470.0]	(12.0, 19.0]	(-1.0, 3.0]	(11.0, 20.0]	Yes	No	No	low cost
(500.0, 852.0]	Yes	(0.4, 0.7]	Yes	(1.0, 5.0]	Yes	(1.0, 21.0]	(0.0, 0.3]	(79.0, 119.0]	5	(13.0, 20.0]	(795.0, 1960.0]	(1006.0, 1488.0]	(255.0, 1470.0]	(12.0, 19.0]	(7.0, 18.0]	(11.0, 20.0]	Yes	No	No	low cost
(500.0, 852.0]	Yes	(2.2, 3.0]	Yes	(-1.0, 1.0]	No	(1.0, 21.0]	(0.0, 0.3]	(160.0, 200.0]	5	(-1.0, 7.0]	(-1.0, 371.0]	(499.0, 1006.0]	(2711.0, 3998.0]	(4.0, 12.0]	(-1.0, 3.0]	(1.0, 11.0]	No	No	No	very high cost
(1226.0, 1615.0]	Yes	(1.5, 2.2]	No	(1.0, 5.0]	Yes	(21.0, 42.0]	(0.3, 0.7]	(160.0, 200.0]	8	(13.0, 20.0]	(-1.0, 371.0]	(499.0, 1006.0]	(2711.0, 3998.0]	(12.0, 19.0]	(7.0, 18.0]	(11.0, 20.0]	Yes	Yes	Yes	very high cost
(1615.0, 1998.0]	No	(2.2, 3.0]	No	(1.0, 5.0]	No	(21.0, 42.0]	(0.3, 0.7]	(119.0, 160.0]	4	(13.0, 20.0]	(371.0, 795.0]	(499.0, 1006.0]	(1470.0, 2711.0]	(12.0, 19.0]	(-1.0, 3.0]	(1.0, 11.0]	Yes	No	No	medium cost
(500.0, 852.0]	Yes	(1.5, 2.2]	No	(5.0, 19.0]	No	(1.0, 21.0]	(0.7, 1.0]	(160.0, 200.0]	4	(7.0, 13.0]	(-1.0, 371.0]	(1006.0, 1488.0]	(1470.0, 2711.0]	(4.0, 12.0]	(-1.0, 3.0]	(1.0, 11.0]	Yes	No	Yes	high cost
(1615.0, 1998.0]	No	(0.4, 0.7]	No	(5.0, 19.0]	Yes	(42.0, 64.0]	(0.3, 0.7]	(160.0, 200.0]	1	(13.0, 20.0]	(-1.0, 371.0]	(499.0, 1006.0]	(255.0, 1470.0]	(12.0, 19.0]	(7.0, 18.0]	(1.0, 11.0]	Yes	No	Yes	low cost
(500.0, 852.0]	No	(0.7, 1.5]	No	(1.0, 5.0]	No	(42.0, 64.0]	(0.3, 0.7]	(119.0, 160.0]	2	(13.0, 20.0]	(795.0, 1960.0]	(1488.0, 1998.0]	(255.0, 1470.0]	(12.0, 19.0]	(7.0, 18.0]	(1.0, 11.0]	Yes	Yes	Yes	low cost
(500.0, 852.0]	No	(0.4, 0.7]	No	(-1.0, 1.0]	Yes	(1.0, 21.0]	(0.0, 0.3]	(160.0, 200.0]	8	(-1.0, 7.0]	(795.0, 1960.0]	(1488.0, 1998.0]	(2711.0, 3998.0]	(4.0, 12.0]	(7.0, 18.0]	(11.0, 20.0]	Yes	No	Yes	very high cost
(500.0, 852.0]	No	(0.7, 1.5]	Yes	(5.0, 19.0]	Yes	(21.0, 42.0]	(0.0, 0.3]	(119.0, 160.0]	3	(13.0, 20.0]	(371.0, 795.0]	(499.0, 1006.0]	(2711.0, 3998.0]	(4.0, 12.0]	(-1.0, 3.0]	(11.0, 20.0]	Yes	Yes	No	very high cost
(852.0, 1226.0]	Yes	(0.4, 0.7]	Yes	(5.0, 19.0]	No	(42.0, 64.0]	(0.3, 0.7]	(79.0, 119.0]	5	(13.0, 20.0]	(371.0, 795.0]	(499.0, 1006.0]	(1470.0, 2711.0]	(12.0, 19.0]	(7.0, 18.0]	(11.0, 20.0]	Yes	Yes	No	medium cost
(500.0, 852.0]	Yes	(0.4, 0.7]	No	(1.0, 5.0]	No	(1.0, 21.0]	(0.7, 1.0]	(119.0, 160.0]	4	(7.0, 13.0]	(795.0, 1960.0]	(1006.0, 1488.0]	(1470.0, 2711.0]	(4.0, 12.0]	(-1.0, 3.0]	(11.0, 20.0]	No	Yes	Yes	medium cost

Figure 12. Discretized data

After we have discretized the data as shown above, only then we changed it to its respective value. At this point, we also dropped the unimportant features. The data should now represented as binary (0 or 1) or categorical at nominal level and should look like in the figure below.

battery_power	clock_speed	fc	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_w	price_range
0	2	0	0	1	2	2	0	0	0	1	1	1
1	0	0	2	1	1	3	0	2	2	1	0	2
0	0	1	1	2	1	5	0	2	2	1	0	2
0	3	0	0	2	1	6	1	2	2	2	2	2
3	1	2	2	1	1	2	2	2	1	0	0	1
3	0	1	1	1	2	1	0	2	2	0	0	1
3	2	1	0	2	1	8	1	1	1	2	2	3
3	0	0	1	2	2	4	0	1	1	0	0	0
2	0	0	2	1	2	7	2	1	0	0	0	0
0	0	1	0	0	0	5	2	2	1	0	2	0
0	3	0	0	0	2	5	0	0	0	2	0	3
2	2	1	1	1	2	8	2	0	0	2	2	3
3	3	1	1	1	1	4	2	1	0	1	0	1
0	2	2	0	2	2	4	1	0	1	1	0	2
3	0	2	2	1	2	1	2	0	0	0	2	0
0	1	1	2	1	1	2	2	2	2	0	2	0
0	0	0	0	0	2	8	0	2	2	2	2	3
0	1	2	1	0	1	3	2	1	0	2	0	3
1	0	2	2	1	0	5	2	1	0	1	2	1
0	0	1	0	2	1	4	1	2	1	1	0	1

Figure 13. Discretized data with numerosity reduction

4.2.4 **Transformed Data**

The figure below shows the data after transformation. Note that the target feature should not be scaled.

battery_power	clock_speed	fc	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_w	price_range
0.000000	0.666667	0.0	0.0	0.5	1.0	0.142857	0.0	0.0	0.0	0.5	0.5	1
0.333333	0.000000	0.0	1.0	0.5	0.5	0.285714	0.0	1.0	1.0	0.5	0.0	2
0.000000	0.000000	0.5	0.5	1.0	0.5	0.571429	0.0	1.0	1.0	0.5	0.0	2
0.000000	1.000000	0.0	0.0	1.0	0.5	0.714286	0.5	1.0	1.0	1.0	1.0	2
1.000000	0.333333	1.0	1.0	0.5	0.5	0.142857	1.0	1.0	0.5	0.0	0.0	1
1.000000	0.000000	0.5	0.5	0.5	1.0	0.000000	0.0	1.0	1.0	0.0	0.0	1
1.000000	0.666667	0.5	0.0	1.0	0.5	1.000000	0.5	0.5	0.5	1.0	1.0	3
1.000000	0.000000	0.0	0.5	1.0	1.0	0.428571	0.0	0.5	0.5	0.0	0.0	0
0.666667	0.000000	0.0	1.0	0.5	1.0	0.857143	1.0	0.5	0.0	0.0	0.0	0
0.000000	0.000000	0.5	0.0	0.0	0.0	0.571429	1.0	1.0	0.5	0.0	1.0	0
0.000000	1.000000	0.0	0.0	0.0	1.0	0.571429	0.0	0.0	0.0	1.0	0.0	3
0.666667	0.666667	0.5	0.5	0.5	1.0	1.000000	1.0	0.0	0.0	1.0	1.0	3
1.000000	1.000000	0.5	0.5	0.5	0.5	0.428571	1.0	0.5	0.0	0.5	0.0	1
0.000000	0.666667	1.0	0.0	1.0	1.0	0.428571	0.5	0.0	0.5	0.5	0.0	2
1.000000	0.000000	1.0	1.0	0.5	1.0	0.000000	1.0	0.0	0.0	0.0	1.0	0
0.000000	0.333333	0.5	1.0	0.5	0.5	0.142857	1.0	1.0	1.0	0.0	1.0	0
0.000000	0.000000	0.0	0.0	0.0	1.0	1.000000	0.0	1.0	1.0	1.0	1.0	3
0.000000	0.333333	1.0	0.5	0.0	0.5	0.285714	1.0	0.5	0.0	1.0	0.0	3
0.333333	0.000000	1.0	1.0	0.5	0.0	0.571429	1.0	0.5	0.0	0.5	1.0	1
0.000000	0.000000	0.5	0.0	1.0	0.5	0.428571	0.5	1.0	0.5	0.5	0.0	1

Figure 14. Histogram of features after data reduction

5 Conclusion

In conclusion, the stage of data preparation is an essential component in the process of classifying the costs of mobile phones. It is possible to significantly improve the accuracy and reliability of price classification models by first gathering data that is pertinent and trustworthy, then investigating the data, cleaning the data by addressing issues such as missing values, noise, and inconsistent data, and finally conducting data reduction.

The appropriate classification of prices requires a firm basis, which may be provided by the acquisition of data from credible sources. Dealing with missing values using methods that are appropriate helps to preserve the reliability as well as the comprehensiveness of the dataset. The management of outliers helps to ensure that the results are not distorted, and it also makes the models more reliable. Cleaning the data gets rid of noise such as duplicates, checks the labels, and keeps the formatting consistent, all of which contribute to the overall improvement in the dataset's quality and dependability. Data reduction enables the selection and production of pertinent elements that capture the key qualities of mobile phones, and it does this by reducing the amount of data.

In today's rapidly evolving and very competitive mobile phone industry, consumers and companies alike have the ability to make well-informed judgments by properly preparing data. A more in-depth awareness of the trends in the market, a comparison of the many phone options, and the determination of the most effective pricing strategies are all made possible by accurate price categorization models. In general, the compilation of data is an essential stage that forms the foundation for

effective mobile phone pricing classification. It maintains the correctness, dependability, and utility of the categorization models, which enables stakeholders to make informed decisions and maintain a competitive edge in the mobile phone business, which is always undergoing innovation and change.

Acknowledgements

We would like to give credit to all of the people who assisted us throughout this project directly or indirectly. Without their support and guidance it wouldn't have been possible. We appreciate our lecturer of this course, Prof. Dr. Azuraliza binti Abu Bakar, for her guidance and supervision which has significantly provided a lot of resources and knowledge required in order to complete our project.

Our colleagues were also constantly encouraging us with two-ways discussions to complete these assigned projects together. We are so grateful to them in developing the project, for their willingness and assistance.

Last but not least, we want to extend our gratitude towards Allah The Great Almighty for his blessing on us and giving us health and courage.

References

1. Hu, N.: Classification of Mobile Phone Price Dataset Using Machine Learning Algorithms. In: 2022 3rd International Conference on Pattern Recognition and Machine Learning, pp. 438-443. IEEE, Chengdu, China (2022).
2. Güvenç, E., Çetin, G., Koçak, H.: Comparison of KNN and DNN classifiers performance in predicting mobile phone price ranges. Advances in Artificial Intelligence Research 1(1), 19-28 (2021).
3. Kalaivani, K. S., Priyadharshini, N., Nivedhashri, S., Nandhini, R.: Predicting the price range of mobile phones using machine learning techniques. AIP Conference Proceedings 2387(1), 140010 (2021).
4. Kalmaz, A., Akin, O.: Estimation of Mobile Phone Prices with Machine Learning. In: 2022 International Conference on Engineering and Emerging Technologies, pp. 1-7. IEEE, Kuala Lumpur, Malaysia (2022).
5. Nasser, I. M., Al-Shawwa, M. O., Abu-Naser, S. S.: Developing Artificial Neural Network for Predicting Mobile Phone Price Range. International Journal of Academic Information Systems Research 3(2), 1-6 (2019).
6. The Mobile Economy 2020, https://www.gsma.com/mobileeconomy/, last accessed 2023/5/15.