Dear Intern

Project report is an inherent component of your internship. We are enclosing a reference table of content for the project report. Depending on the internship project (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections from the table of content mentioned below. You can also add additional sections. The key objective of this report is for you to systemically document the project work done.

|  |  |
| --- | --- |
| Internship Project Title | Rank features of a smartphone-build a python application to classify and rank dataset |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | ICT Academy of Kerala |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 11/11/22 | 10/12/22 | 36.5 | Data Science and Analytics | python |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Risk Vs Reward
* Reflections on the Internship
* Recommendations
* Outcome / Conclusion
* Enhancement Scope
* Link to code and executable file
* Research questions and responses

**Acknowledgements**

First I would like to thank ICT Academy of Kerala, for giving me the opportunity to do an internship from TATA Consultancy Services.

I also would like to thank to the organization for giving an opportunity to do an internship of one month.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of the industry mentor Mr. Debashis Roy.

**Objective**

The objective of this project is to build a python application that ranks the features of a smartphone based on the requests received from various users.

**Introduction**

Smart phone is a mobile phone which offers advanced technologies with functionality similar as a personal computer. While offering a standardized platform for application developers a smart phone performs as complete operating system software. Secondly, there are also very advanced features in smart phones such as internet, instant messenger and e-mail and also built-in keyboard are very typical. Because of these reasons we can say a smart phone a miniature computer with the similarities of a simple phone.

Analyst house Gartner gives the definition of a smart phone as: “A large-screen, data-centric, handheld device designed to offer complete phone functions whilst simultaneously functioning as a personal digital assistant (PDA).” (Analyst House Gartner: 2009)

With the growing speed of technological advancement, Smart phones are now an essential part of our daily life routine. When we go for our convenience we also look for those devices which contain multiple features such as office work, mobility, networking and entertainment. As the world is getting advance our needs become sophisticated. Where we need quality, effectiveness and performance we also ask for these all in one single pocket device so we can take that to anywhere with us.

The main aim of the project is to build a python application to classify and rank dataset i.e. to rank features of a smartphone.  The dataset is obtained from the digital discussion room of TCSiON learning platform. This dataset contains information on battery power, clock speed, dual sim, four g, int memory, touch screen, Wi-Fi, price range, ram, talk time etc.

**Internship Activities**

**Milestone #1**: Create a dataset, clean the dataset and also sanitize it and classify the dataset into hardware and software request.

**Day 1 to Day 5**: Watch the vedios in TCS learning hub . Understand the topic , go through different reference link that are pinned in the learning hub page. Downloaded the dataset from digital discussion room and started working on it. Call the dataset into python environment , checked for the basic details ie shape, column names, data types ,etc. Next done the pre-processing ie checked for the null values ,outliers . There was no null values present in the test and train data. There was some outliers in two columns but it is not considered .

Done the Exploratory data analysis. Identified the target column(price\_range) and plot graph for it. Plotted univariate graphs for other features in the dataset. Take cor-relation heatmap and analyse the relation between different features and drawn bivariate and multivariate graphs for different features. Next classify the data into hardware and software and then concate the data.

**Milestone #2:** Merge the classified data and rank the final dataset.

**Day 6 to Day 15:** merged the classified data. Done feature engineering, identified the most important features, conduct feature scaling and modelling, split the data and applied various classification models on it (Logistic model, Decision tree, KNN, Random forest , Gradient Boost, SVM). Call the test data into python environment and predict the output using the highest accuracy model and saved it as a csv file.

**Day 16 to Day 29:** ranked the dataset according to price range and sorted it from smallest to the largest. Prepared the final project report.

Methodology

**Source of data:** The dataset was downloaded from the digital discussion room of TCS iON Remote Internship 125.

**Columns in the dataset**: battery\_power: Total energy a battery can store in one time measured in mAh

blue: Has bluetooth or not

clock\_speed: speed at which microprocessor executes instructions

dual\_sim: Has dual sim support or not

fc: Front Camera mega pixels

four\_g: Has 4G or not

int\_memory: Internal Memory in Gigabytes

m\_dep: Mobile Depth in cm

mobile\_wt: Weight of mobile phone

n\_cores: Number of cores of processor

pc: Primary Camera mega pixels

px\_height: Pixel Resolution Height

px\_width: Pixel Resolution Width

ram: Random Access Memory in Mega Bytes

sc\_h: Screen Height of mobile in cm

sc\_w: Screen Width of mobile in cm

talk\_time: longest time that a single battery charge will last when you are

three\_g: Has 3G or not

touch\_screen: Has touch screen or not

wifi: Has wifi or not

price\_range: This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

**Data Analysis:** it is the technique to collect, transform, and organize data to make future predictions, and make informed data-driven decisions. It also helps to find possible solutions for a business problem. There are six steps for Data Analysis. They are:

* Ask or Specify Data Requirements
* Prepare or Collect Data
* Clean and Process
* Analyze
* Share
* Act or Report

Call the data into python environment and take the basic details about the data. Then done the pre-processing, checked for the null values but there was no null values present in the dataset. Checked for the outliers, there was few outliers in the columns ‘fc’ & ‘px\_height’ but is not considered as an outlier. Then done the Exploratory Data Visualization.

**Exploratory Data Analysis:** EDA is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

**Univariate Visualization:** Univariate data visualization plots help us comprehend the enumerative properties as well as a descriptive summary of the particular data variable. These plots help in understanding the location/position of observations in the data variable, its distribution, and dispersion.

**Bivariate Visualization:** Bivariate analysis is an analysis that is performed to determine the relationship between 2 variables. In this analysis, two measurements were made for each observation.

**Multivariate Visualization:** Multivariate data visualization involves visualizing more than one data value in a single renderer.

**Count plot:** Show the counts of observations in each categorical bin using bars. A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for barplot() , so you can compare counts across nested variables.

**Pie plot:** A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents.

**Dist plot:** A Distplot or distribution plot, depicts the variation in the data distribution. Seaborn Distplot represents the overall distribution of continuous data variables. The Seaborn module along with the Matplotlib module is used to depict the distplot with different variations in it.

**Correlation heat map:**  Correlation heatmaps are a type of plot that visualize the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship.

**Bar plot:** A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

**Swarm plot:** A swarm plot is a type of scatter plot that is used for representing categoricalvalues. It is very similar to the strip plot, but it avoids the overlapping of points. We can use the seaborn. swarmplot() to create such graphs. It is not advisable to use this type of graph when the sample size is large.

**Reg plot:** This method is used to plot data and a linear regression model fit. There are a number of mutually exclusive options for estimating the regression model.

**Joint plot:** A Jointplot comprises three plots. Out of the three, one plot displays a bivariate graph which shows how the dependent variable(Y) varies with the independent variable(X).

Another plot is placed horizontally at the top of the bivariate graph and it shows the distribution of the independent variable(X).

**Box plot:** In descriptive statistics, a box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.

**Hardware and Software Classification:** train data is classified into hardware and software data.

**Merging:** Data merging is the process of combining two or more similar records into a singleone. Merging is done to add variables to a dataset, append or add cases or observations to a dataset, or remove duplicates and other incorrect information.

**Feature Engineering:** Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model using machine learning or statistical modeling.

**Feature Selection:** Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.

**Feature Scaling:**  Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle

highly varying magnitudes or values or units.

**Modelling:** Data modeling is the process of producing a descriptive diagram of relationships between various types of information that are to be stored in a database. One of the goals of data modeling is to create the most efficient method of storing information while still providing for complete access and reporting.

**Logistic Regression:** Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.

 Logistic regression is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outcome variable.

 Works only when the predicted variable is binary, assumes all predictors are independent of each other and assumes data is free of missing values.

**K-Nearest Neighbours**:Neighbours based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbours of each point.

This algorithm is simple to implement, robust to noisy training data, and effective if training data is large. Need to determine the value of K and the computation cost is high as it needs to compute the distance of each instance to all the training samples.

**Decision Tree:** Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

[Decision Tree](https://analyticsindiamag.com/hands-on-tutorial-how-to-use-decision-tree-regression-to-solve-machinehacks-new-data-science-hackathon/) is simple to understand and visualise, requires little data preparation, and can handle both numerical and categorical data.Decision tree can create complex trees that do not generalise well, and decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

**Random Forest:** it is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases. Slow real time prediction, difficult to implement, and complex algorithm.

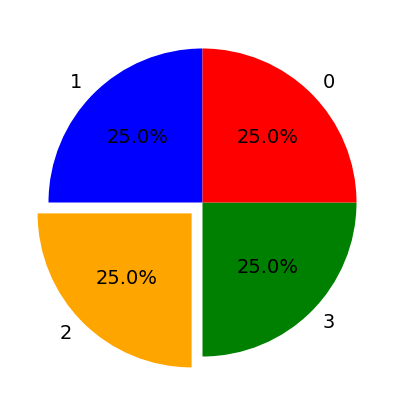
**Support Vector Machine:** it is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

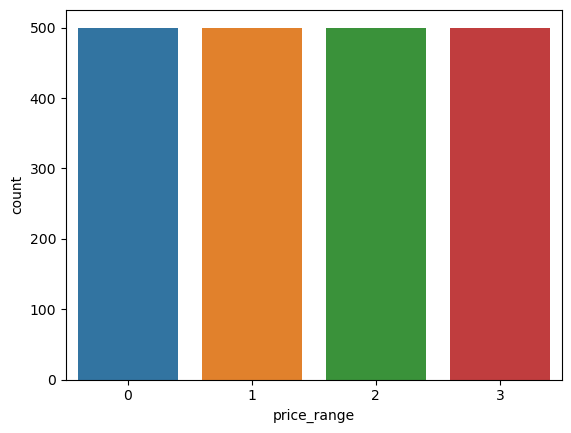
Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient. The algorithm does not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

**Gradient Boosting Classifier:** Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

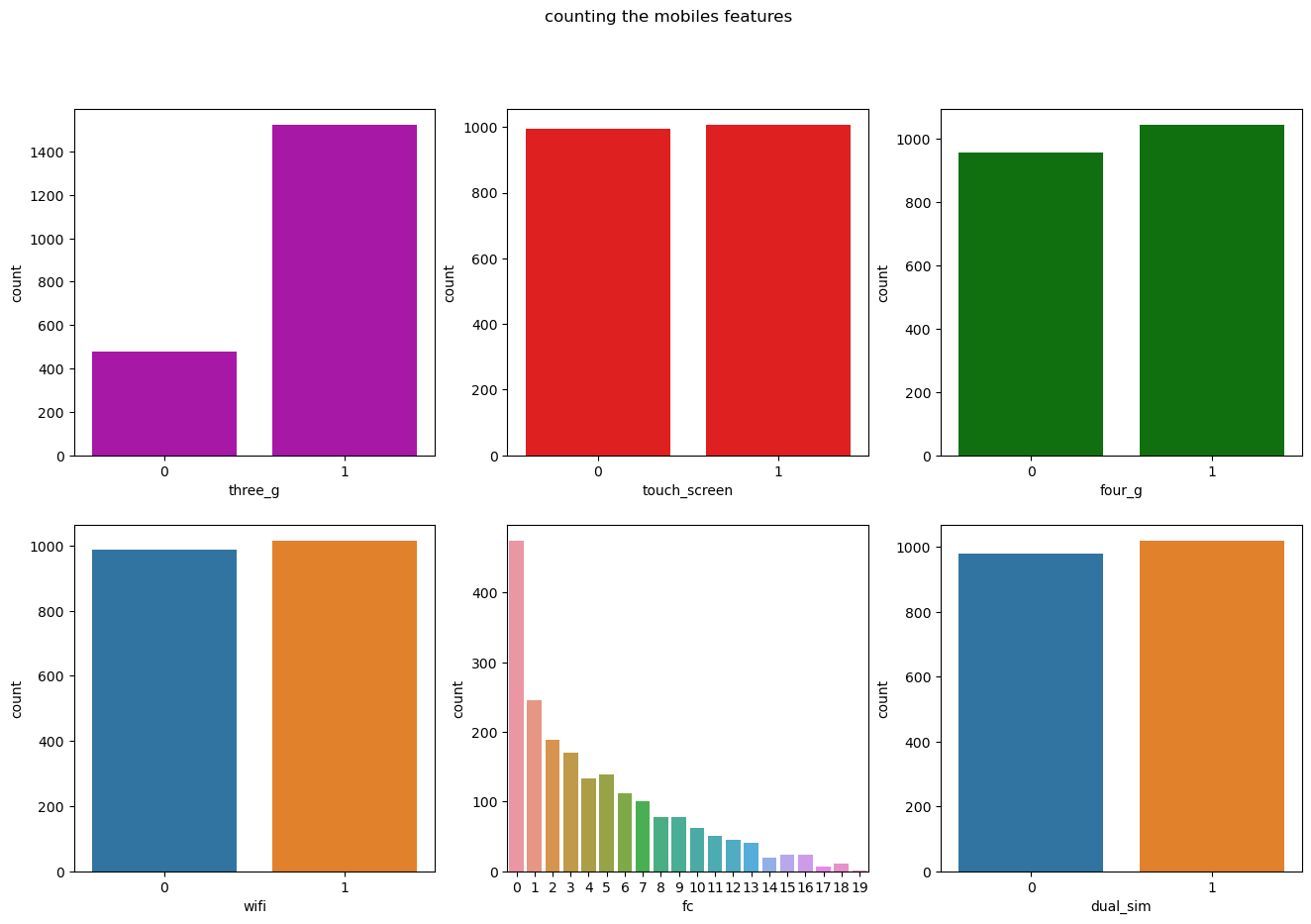
Expolratory data visualization

**Target column:** ‘price\_range’ is the dependent column in the dataset. When plotted a count plot and pie chat , it’s seen that the data is balanced.





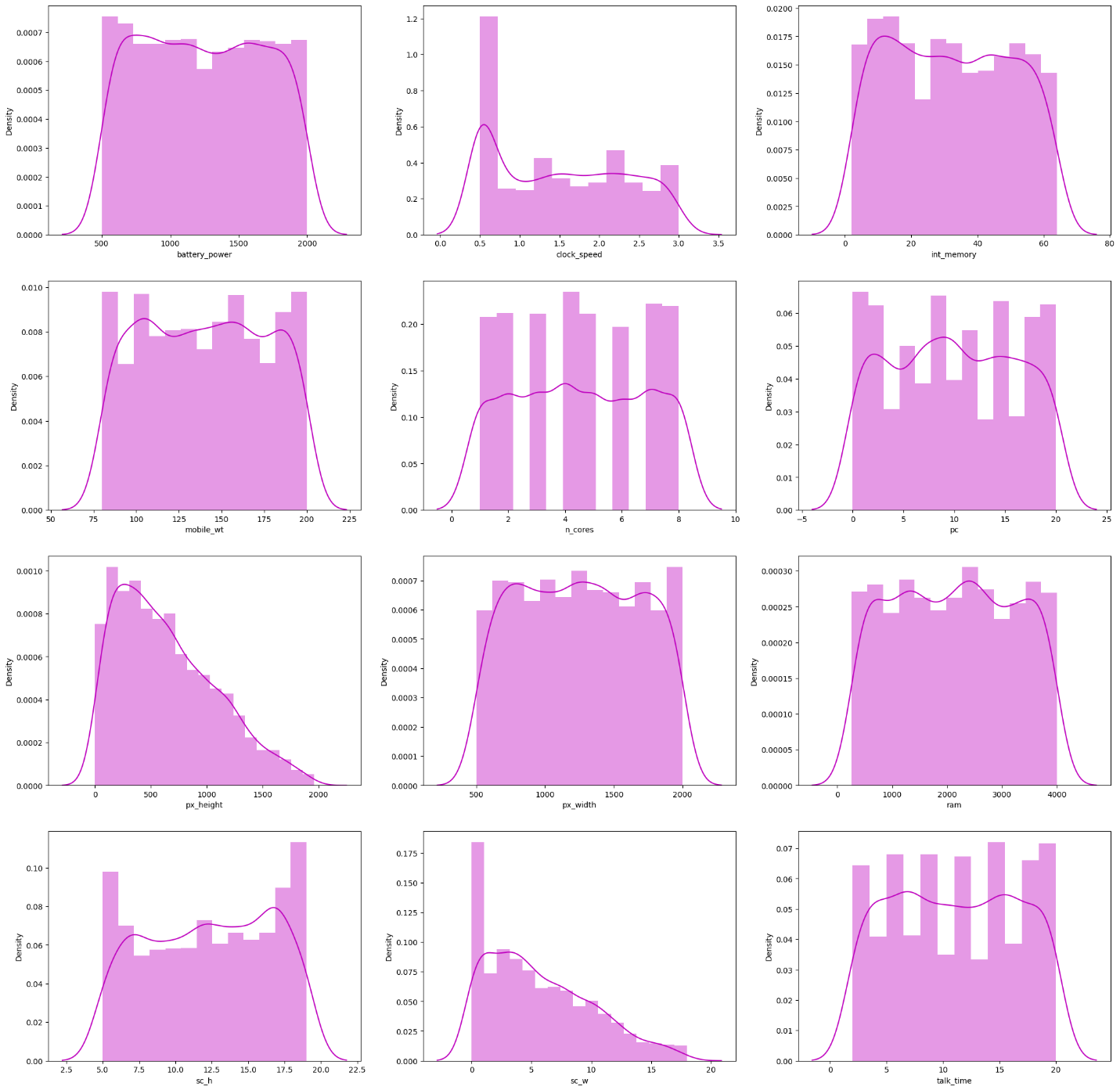
**Univariate visualization:** using countplot , drawn graphs of some features like 3G , 4G, fc, dual sim , wifi , touch screen.



**Dist plot:** drawn the dist plot of some features like "battery\_power","clock\_speed","int\_memory","mobile\_wt","n\_cores","pc","px\_height","px\_width",

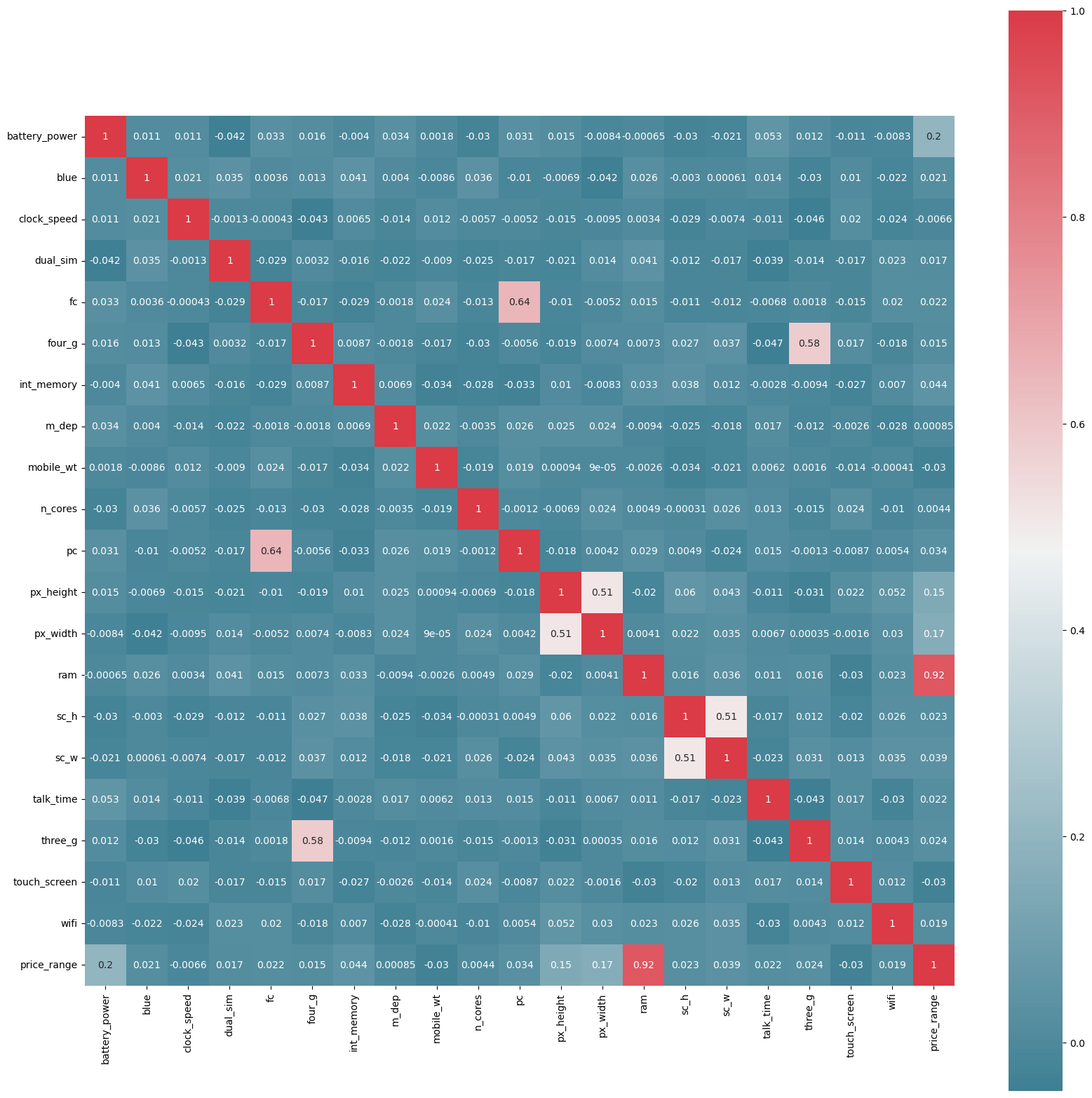
"ram","sc\_h","sc\_w","talk\_time","price\_range".

This shows the distribution of continues variables present in the dataset.



**Correlation heatmap**: from the heat map we can understand that there is an high positive correlation between ram and price range of smartphone. There is a correlation between fc and pc and also a correlation between px\_height and px\_width and between three g and four g.

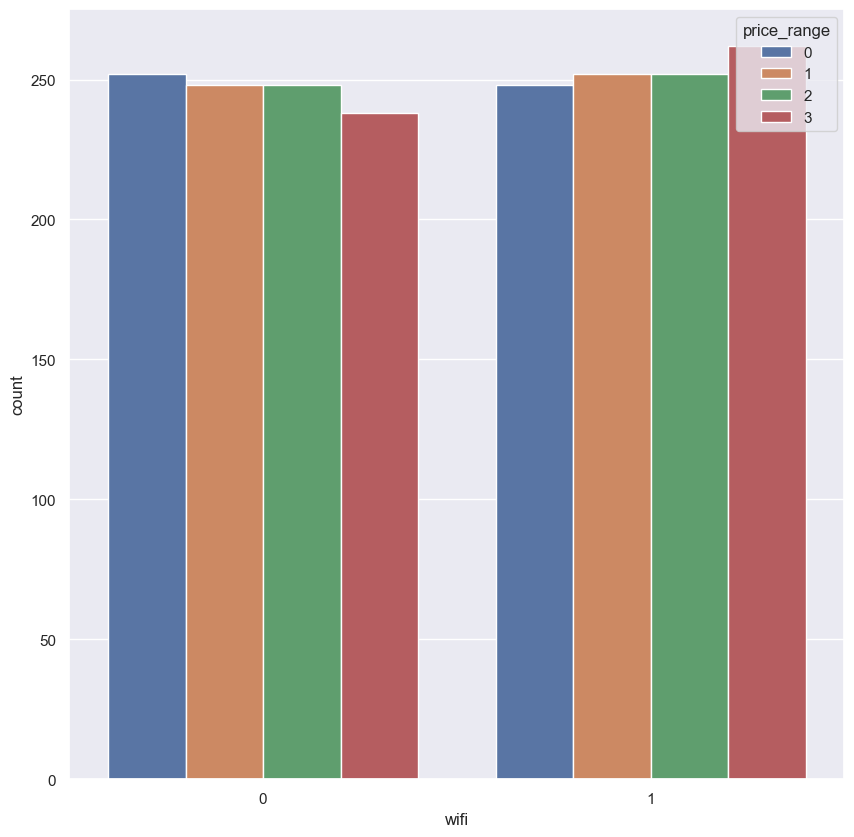
By observing the correlation we have drawn bivariate and multivariate graphs.

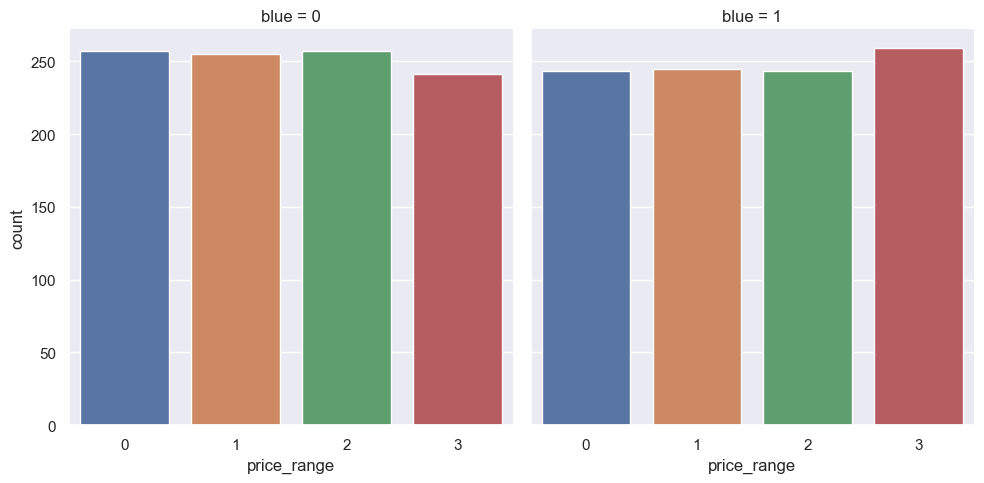


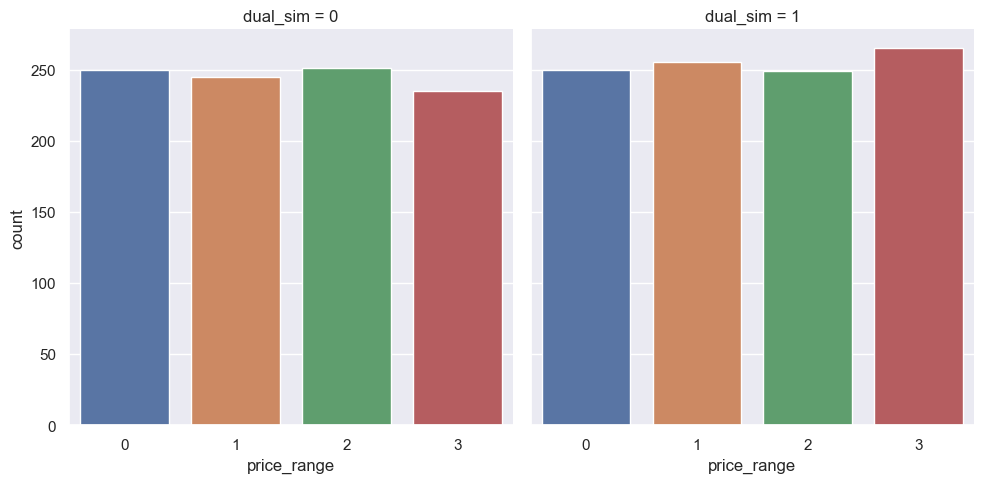
**Bivariate visualization:** since there is an high correlation between ram and price range, we have drawn a swarm plot .The smartphone having high ram capacity has high price range.



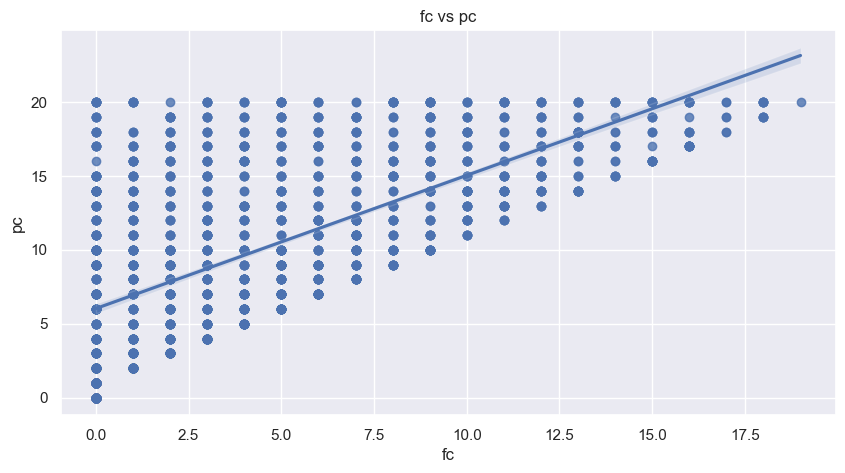
Drawn different bivariate graphs of dependent variable vs independent variable ie price range vs wifi, blue, dual sim.



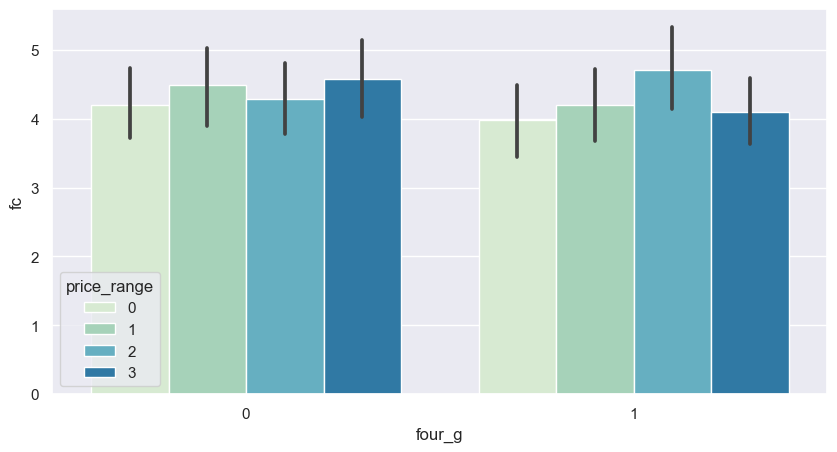


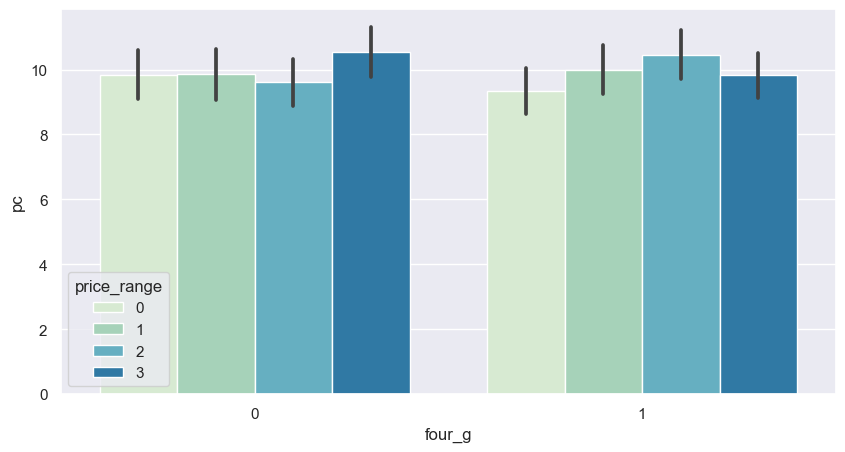


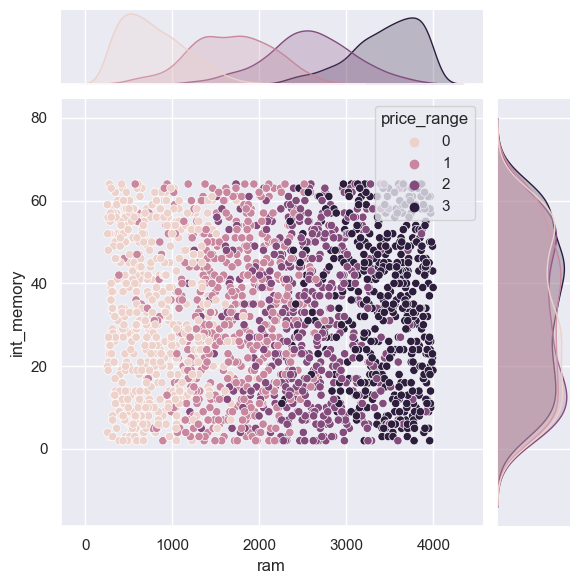
Drawn regression plot between fc and pc. It shows that when fc increases pc also increases.

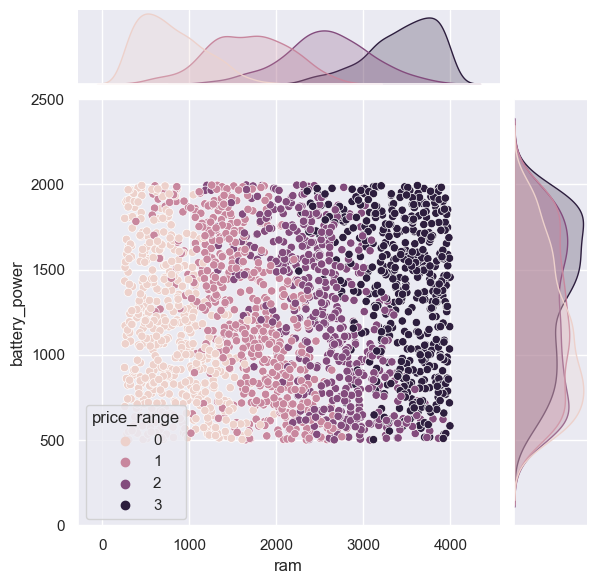


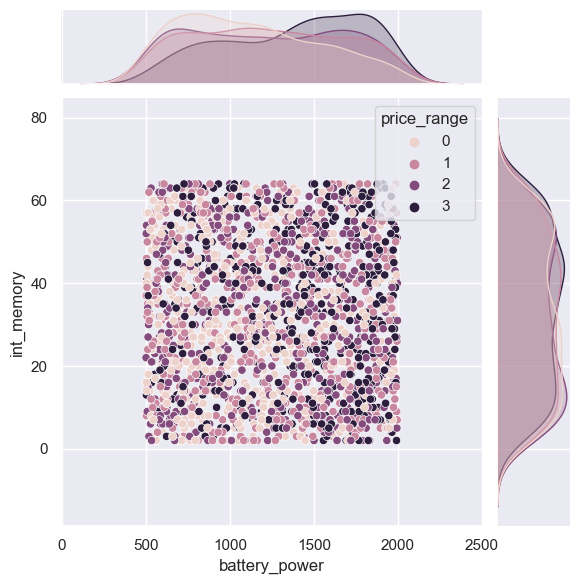
**Multivariate visualization:** drawn different multivariate graph between price range,fc,4G,pc. And between ram , internal memory,price range,battery power,dual sim ,talk time,n cores and clock speed.

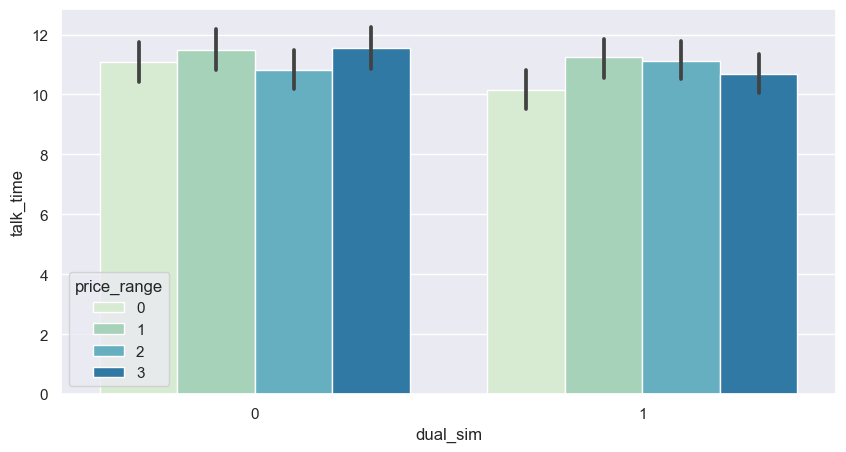


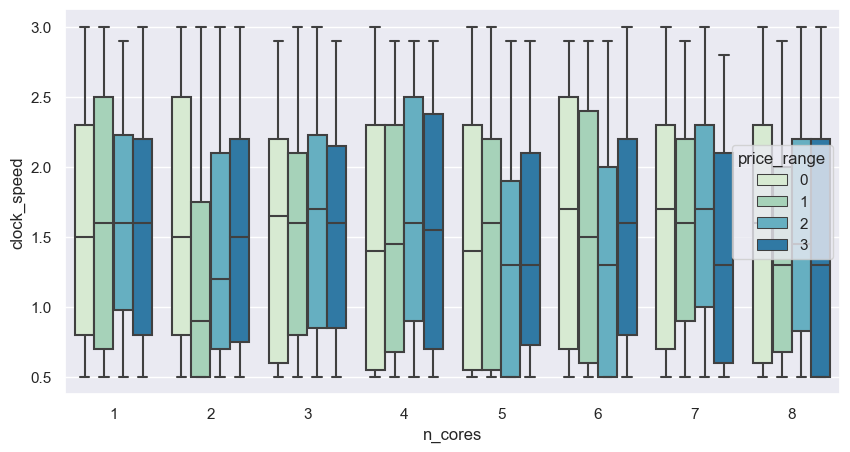












Feature engineering

The classified hardware and software data is merged together. Next done the feature engineering, selection and scaling. Applied SelectKBest method to find out the best twelve features and the best features and it scores are:

Features Score

13 ram 931267.519053

11 px\_height 17363.569536

0 battery\_power 14129.866576

12 px\_width 9810.586750

8 mobile\_wt 95.972863

6 int\_memory 89.839124

15 sc\_w 16.480319

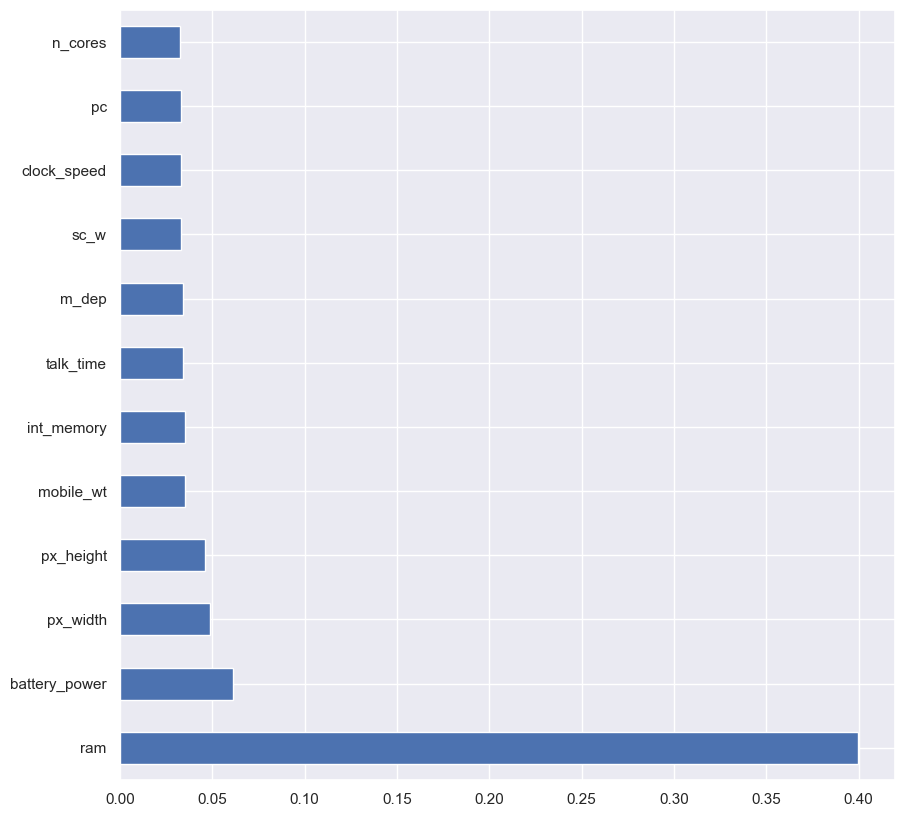
16 talk\_time 13.236400

4 fc 10.135166

14 sc\_h 9.614878

10 pc 9.186054

9 n\_cores 9.097556



Then by using StandardScalar standardise the train data.

**from** sklearn.preprocessing **import** StandardScaler

scaler**=** StandardScaler()

x**=**scaler**.**fit\_transform(x)

Then split the data into train and test set by giving 25% to test and remaining to train set. Then by using different classification models checked for the accuracy.

Classification models

**Logistic Regression:** from sklearn library , imported the logistic model and checked for the accuracy. It gives 96.2% of accuracy.

Accuracy is: 0.962

Results from Logistic Regression

|  | **precision** | **recall** | **f1-score** | **support** |
| --- | --- | --- | --- | --- |
| **0** | 0.983871 | 0.976000 | 0.979920 | 125.000000 |
| **1** | 0.936364 | 0.936364 | 0.936364 | 110.000000 |
| **2** | 0.928000 | 0.958678 | 0.943089 | 121.000000 |
| **3** | 0.992908 | 0.972222 | 0.982456 | 144.000000 |
| **accuracy** | 0.962000 | 0.962000 | 0.962000 | 0.962000 |
| **macro avg** | 0.960286 | 0.960816 | 0.960457 | 500.000000 |
| **weighted avg** | 0.962501 | 0.962000 | 0.962155 | 500.000000 |
| **Decision Tree:** from sklearn library imported decision tree classifier and checked for the accuracy. The accuracy score is 84.8%.  Accuracy is: 0.848  Results from Decision Tree   |  | **precision** | **recall** | **f1-score** | **support** | | --- | --- | --- | --- | --- | | **0** | 0.895161 | 0.932773 | 0.913580 | 119.000000 | | **1** | 0.845455 | 0.781513 | 0.812227 | 119.000000 | | **2** | 0.776000 | 0.782258 | 0.779116 | 124.000000 | | **3** | 0.872340 | 0.891304 | 0.881720 | 138.000000 | | **accuracy** | 0.848000 | 0.848000 | 0.848000 | 0.848000 | | **macro avg** | 0.847239 | 0.846962 | 0.846661 | 500.000000 | | **weighted avg** | 0.847481 | 0.848000 | 0.847318 | 500.000000 | |

**KNN:** checked for the k value which gives the highest accuracy and the k value is 14 and the accuracy score is 59%.

accuracy score is : 0.59

Results from K-NN

|  | | | | **recall** | **f1-score** | **support** |
| --- | --- | --- | --- | --- | --- | --- |
| **0** | | | | 0.725806 | 0.725806 | 124.000000 |
| **1** | | | | 0.421429 | 0.472000 | 140.000000 |
| **2** | | | | 0.458647 | 0.472868 | 133.000000 |
| **3** | | | | 0.825243 | 0.696721 | 103.000000 |
| **accuracy** | | | | 0.590000 | 0.590000 | 0.590000 |
| **macro avg** | | | | 0.607781 | 0.591849 | 500.000000 |
| **weighted avg** | | | | 0.590000 | 0.581468 | 500.000000 |
|  |  |  |  |
|  |  |  |  |

Prediction

Since the highest accuracy score is for logistic model, we predict the price range using logistic model and saved it as a csv file.

test\_predict**=**log\_model**.**predict(test\_scaled)

result**=**pd**.**DataFrame(test\_predict)

result

| **0** |
| --- |
| **0** | 2 |
| **1** | 3 |
| **2** | 2 |
| **3** | 3 |
| **4** | 1 |
| **...** | ... |
| **995** | 2 |
| **996** | 1 |
| **997** | 0 |
| **998** | 2 |
| **999** | 2 |

1000 rows × 1 columns

result**.**to\_csv('result.csv')

As it is saved as csv file it automatically get downloaded as an excel file.

Next we have ranked the dataset by price range.

Ranking

The rank() function is used to compute numerical data ranks (1 through n) along axis. By default, equal values are assigned a rank that is the average of the ranks of those values. Index to direct ranking.

By using rank() function ,ranked the train dataset by price range.

Mobile\_Train["rank\_by\_price"]**=**Mobile\_Train["price\_range"]**.**rank()

data**=**Mobile\_Train

data

Then sorted the dataset from smallest to largest price range.

data**.**sort\_values(by**=**["rank\_by\_price"])

Then ranked all the features separately to correct the output, because not all features are good when values are high or low it depends on each and every feature.

Conclusion

RAM is the part of the phone that is used to store the operating system (OS) and where apps and data currently in use are kept. Whereas, phone storage is used to store data such as apps, photos, videos, and files that are necessary for the phone to run. We need at least 4GB. That's enough RAM for web browsing, social media, video streaming, and some popular mobile games. However, while that applies to most smartphone users, the amount of RAM you need depends on the apps you use, and many opt for smartphones with between 6GB-8GB RAM.

Through this analyzation we can conclude that the smartphone with high capacity of RAM costs higher ,than lower ones.

Link to code and executable file

<https://github.com/LINTA-STEPHEN/TCS_iON_RIO_125/blob/main/TCS%20Internship.ipynb>

<https://github.com/LINTA-STEPHEN/TCS_iON_RIO_125>