



Daffodil
International
University

Predicting Mobile Price range using Classification techniques

Submitted by

Ahsanul Hoque Sakib

ID:171-35-1838

Department of Software Engineering

Daffodil International University

Supervised by

Asif Khan Shakir

Senior Lecturer

Department of Software Engineering

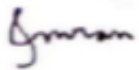
Daffodil International University

This thesis report has been submitted in fulfillment of the requirements for the Degree of
Bachelor of Science in Software Engineering

Approval

This **thesis** titled on “Predicting Mobile Price Range using Classification Techniques”, submitted by Ahsanul Hoque Sakib (ID:171-35-1838) to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and as to its style and contents.

BOARD OF EXAMINERS



Chairman

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 1

K. M. Imtiaz-Ud-Din
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 2

Md Fahad Bin Zamal
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



External Examiner

Professor Dr. Md. Nasim Akhtar
Professor
Department of Computer Science and Engineering
Dhaka University of Engineering and Technology, Gazipur

Declaration

I hereby declare that thesis title “Predicting Mobile Price Range using Classification Techniques” is an original thesis done by me under the supervision of Asif Khan Shakir, Senior Lecturer, Department of Software Engineering, Daffodil International University, towards the partial fulfillment of requirement for the award of degree of Bachelor of Science in Software Engineering during the period of 2017-2021. I also state that this thesis has not been submitted any other place.




Ahsanul Hoque Sakib

ID: 171-35-1838

Department of Software Engineering
Daffodil International University

Certified by:



Asif Khan Shakir

Senior Lecturer

Department of Software Engineering
Daffodil International University

Acknowledgement

To become a successful person, we have to be a hardworking and practical man. That's why, I always try my level best to give my best. For completing my graduation, I think thesis is the best solution for practicing my skill. Because thesis is a bridge between theoretical and practical working. With this willing I join this project.

At first, I would like to thank to the ALMIGHTY ALLAH who always guided me to work on the right path of the life. I must be thankful to my parents and my family for giving me the opportunity and always be to myself.

I am exceptionally obligated to the respectful teachers and specially to my supervisor “Asif Khan Shakir” for giving me necessary information for completing my thesis.

I am grateful to my Department staff members, Lab technicians and non-teaching staff members for their extreme help throughout my project.

And lastly, I would like to express my love to my batch mate for their co-operation and consolation which help me to finish my project.

Table of Content

Approval	i
Declaration	ii
Acknowledgement	iii
Table of Content	iv
List of Equation	v
List of Tables	v
List of Figure	vi
Abstract	vii
Chapter 1	1
Introduction	1
Objective	2
Chapter 2	3
Literature Review	3
Chapter 3	4
Methodology	4
Data Collecting	4
Data Cleaning	5
Attribute Analysis	7
Correlation	17
Selecting Features	17
Algorithms Applied	18
Chapter 4	25
Performance Evaluation	25
Accuracy	25
Precision	25
Recall/Sensitivity	25
ROC Curve	26
Decision Tree	26
Random Forest	26
Chapter 5	27

Result and Discussion	27
Chapter 6	29
Conclusion and Future Work	29
Findings.....	29
Limitation	29
Future Work.....	29
References.....	30
Appendix.....	32

List of Equation

Equation 1: Entropy	18
Equation 2: Root Mean Square	22
Equation 3: Gini Impurity	22
Equation 4: Accuracy	25
Equation 5: Precision	25
Equation 6: Recall.....	25

List of Tables

Table 1: Dataset Distinct Values.....	5
Table 2: Performance Comparison between models.....	27
Table 3: ROC Curve Comparison.....	28

List of Figure

Figure 1: Working Flow.....	4
Figure 2: Dataset	4
Figure 3:Histogram of Attributes	6
Figure 4: Bluetooth	7
Figure 5: Dual Sim.....	7
Figure 6: Four G	8
Figure 7: Three G.....	8
Figure 8: Touch Screen.....	9
Figure 9: Wi-Fi	9
Figure 10: Battery Count	10
Figure 11:Clock Speed.....	10
Figure 12: Front Camera.....	11
Figure 13: Internal Memory.....	11
Figure 14: Mobile Depth.....	12
Figure 15: Mobile Weight.....	12
Figure 16: Cores.....	13
Figure 17: Primary Camera.....	13
Figure 18: Pixel Height.....	14
Figure 19: Pixel Width.....	14
Figure 20: Ram	15
Figure 21: Screen Height	15
Figure 22: Screen Width	16
Figure 23: Talk-Time	16
Figure 24: Correlation Heatmap	17
Figure 25: Best Features Score	17
Figure 26: Decision Tree Levels.....	18
Figure 27: Test and Train Accuracy Using max_depth	19
Figure 28: Confusion Matrix	19
Figure 29: Feature Importance.....	20
Figure 30: Decision Tree Graph.....	20
Figure 31: Pruned Decision Tree(entropy,depth=18)	21
Figure 32: Actual vs Predicted values.....	21
Figure 33: Random Forest.....	21
Figure 34: Tree of Random Forest.....	22
Figure 35: Confusion Matrix Random Forest	23
Figure 36: Actual vs Predictive Values.....	23
Figure 37 : Tuned Random Forest Accuracy	24
Figure 38: ROC Curve Decision Tree.....	26
Figure 39: ROC Curve Random Forest.....	26

Abstract

Machine learning based classification techniques helps to solve the problem related to decision making. In many areas of price prediction are used like housing price prediction, stock price prediction different classification algorithm used. Some of them are used artificial neural network. In this study, three different classification techniques used for prediction the mobile price range. The first one is Naïve Bayes second one is Decision Tree and third one is Random Forest machine learning algorithm. The accuracy got by first two techniques respectively 83% and 84%. As the accuracy of Naïve Bayes is lower than decision tree so Naïve Bayes is not considered. So, for improving the accuracy of Decision Tree, the parameter has been pruned and later Random Forest has been used. It gives 90% accuracy for this dataset. And also, performance evaluation is performed for Decision Tree and Random forest like Precision, Recall.

Keyword: Machine Learning, Decision Trees, Random Forest,

Chapter 1

Introduction

Mobile technology is a technology where users go, this technology also goes. This portable technology consists of two-way communication, computing and networking technology (*Mobile Technology / IBM*, n.d.). The number of mobile users in the world in 2019 is about 3.2 billion and increasingly in 2020 is about 3.5 billion (29+ *Smartphone Usage Statistics: Around the World in 2020*, n.d.). Different commercial activities, university courses, entertainment, communications are also done by a phone. Different organizational tasks, meetings also maintained and held virtually in this pandemic situation. As well as the use of mobile phones is increasing day by day and the prices also vary by their different configurations. There are many companies and many types of quality products in the market.

Nowadays, mobile phones are selling and purchasing in a huge number. Within a short timespan new version with new features are launched to market (Balakumar et al., n.d.). There are many features which are important to consider a mobile price like display resolution, ram, camera, processor, mobile thickness.

As far as, the use of mobile technology is increasing day by day and the price of that phone based on their configuration, quality and brand. One goes to market and search for different phones with different configuration of different brands. So, it makes the customer confusing, budget problem and waste of time and so on.

In a work, ANN is used for predicting mobile price range (Khillha & Shawwa, 2020). There 70% data used for training that model and 30% for validation (Khillha & Shawwa, 2020). In another work, KNN and Linear Regression is used comparatively (Balakumar et al., n.d.). There are many examples of price prediction like stock price (Di Persio & Honchar, 2016), energy price (Ebrahimian et al., 2018)(Zehtab-Salmasi et al., 2020) and so on. In another price prediction, ANN MRA was implemented and compared (Lim et al., 2016). ANN is one of the main tools used in machine learning technique (Khillha & Shawwa, 2020). This is a brain-inspired system which is proposed to simulate the way humans learn. Neural Networks involve input and output layers and, in most cases, hidden layers. Hidden layer neurons are connected only to the other neuron but never directly interact with the user program (Khillha & Shawwa, 2020). Machine learning provides the best technique for Artificial Intelligence like classification, regression, supervised learning and unsupervised learning. There are many types of tools for machine learning tasks MATLAB, Python WEKA etc. So, any type of classifier can be used (Balakumar et al., n.d.).

And also, there are different types of feature selection algorithms to select only the best features and minimize the dataset (Balakumar et al., n.d.). So, the complexity of computation reduced.

So, a classification technique is built to predict by price that if a phone is low price or medium or high or very high depends on some major factors or features of a mobile phone.

This classification technique is using decision tree and random forest and comparing their outputs.

Objective

The main objective is to predict the price range of mobile phone. By which future technology and demand price can be predicted and customers confusion will be clear while buying a phone.

Chapter 2

Literature Review

Using machine learning techniques, the prediction of mobile price is an interesting research background because of the newly launched and availability of phones. Muhammad Asim and Zafar Khan (Asim & Khan, 2018) collected dataset from a website(*GSMarena.Com - Mobile Phone Reviews, News, Specifications and More...*, n.d.) and used Decision Tree and Naive Bayes to predict the classification of mobile price which result was respectively 78% and 75% (WrapperAttributeEval Algorithm) and also the number of selected features was respectively 2 & 5. Using InfoGainEval the best accuracy was respectively 75% for 2 features and 71.42% for 6 features.

B.Balakumar¹, P. Raviraj², V. Gowsalya³ (Balakumar et al., n.d.) used Linear Regression and KNN for predicting mobile price and respectively the accuracy was 91.32% and 92.12% using collected dataset(*Kaggle: Your Home for Data Science*, n.d.-a).

Ibrahim M. Nasser, Mohammed Al-Shawwa(Khillha & Shawwa, 2020) used the Artificial Neural Network to predict the price which accuracy was 96.31%. Abu Nasser, et.al (Al-Daour et al., 2020)(Alghoul et al., 2018) built many neural network models for predicting class. They used the ANN model in the prediction of Temperature and Humidity which gives 100% accuracy (Al-Shawwa et al., 2018).

Keval Pipalia¹, Rahul Bhadja² (Pipalia & Bhadja, n.d.) used many different algorithms for performance evaluation in mobile price prediction. They got the accuracy for Logistic regression 81%, KNN 55%, Decision Tree 82%, SVM 84%, Gradient Boosting 90%.

Chapter 3

Methodology

In this section the working flow will describe how the data collected and used for implementation.

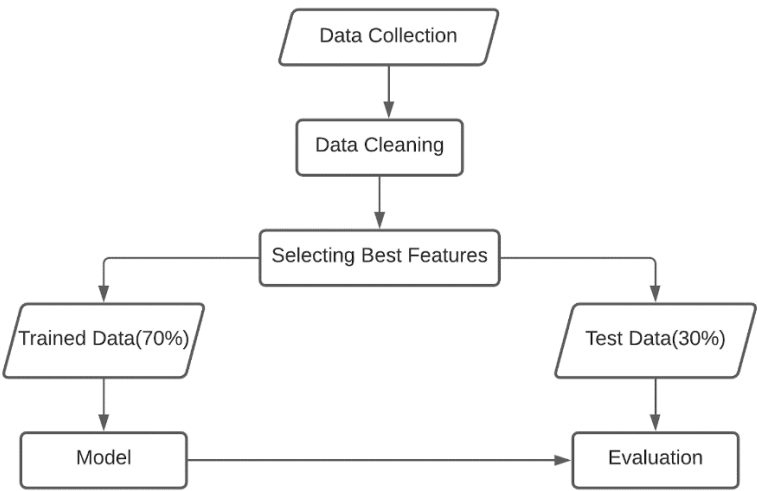


Figure 1: Working Flow

Data Collecting

Features of mobile phone collected from online resource(*Kaggle: Your Home for Data Science*, n.d.-b).

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
842	0	2.2	0	1	0	7	0.6	188	2	2	20	756	2549	9	7	19
1021	1	0.5	1	0	1	53	0.7	136	3	6	905	1988	2631	17	3	7
563	1	0.5	1	2	1	41	0.9	145	5	6	1263	1716	2603	11	2	9
615	1	2.5	0	0	0	10	0.8	131	6	9	1216	1786	2769	16	8	11
1821	1	1.2	0	13	1	44	0.6	141	2	14	1208	1212	1411	8	2	15
...
794	1	0.5	1	0	1	2	0.8	106	6	14	1222	1890	668	13	4	19
1965	1	2.6	1	0	0	39	0.2	187	4	3	915	1965	2032	11	10	16
1911	0	0.9	1	1	1	36	0.7	108	8	3	868	1632	3057	9	1	5
1512	0	0.9	0	4	1	46	0.1	145	5	5	336	670	869	18	10	19
510	1	2.0	1	5	1	45	0.9	168	6	16	483	754	3919	19	4	2

Figure 2: Dataset

Table 1: Dataset Distinct Values

	Ram	Battery	Px height	Px width	Weight	Int_memory	Sc_height	Sc_width	Talktime	Front Camera
Min	256	501	1	500	80	2	5	0	2	0
Max	3998	1998	1960	1998	200	64	19	18	20	19
Mean	2124.213	1238.518	645.108	1251.51	140.24	32.04	12.3	5.76	11.01	4.309
StdDev	1084.732	439.418	443.78	432.199	35.399	18.145	4.21	4.35	5.46	4.34

This table explain the dataset with the maximum, minimum, mean and standard division of each attribute.

Data Cleaning

Checking and cleaning if there are any null data. There were blank data in pixel height so this field has been filled with the value. Then check each attribute using histogram. Data cleaning or preprocessing is one of the important parts of research. Handling missing data is one of them. To deal with missing data firstly the data missing on that row to be deleted or secondly put there calculate mean value. Labeling the categorical data is also an important part of data preprocessing.

Histogram of Attributes

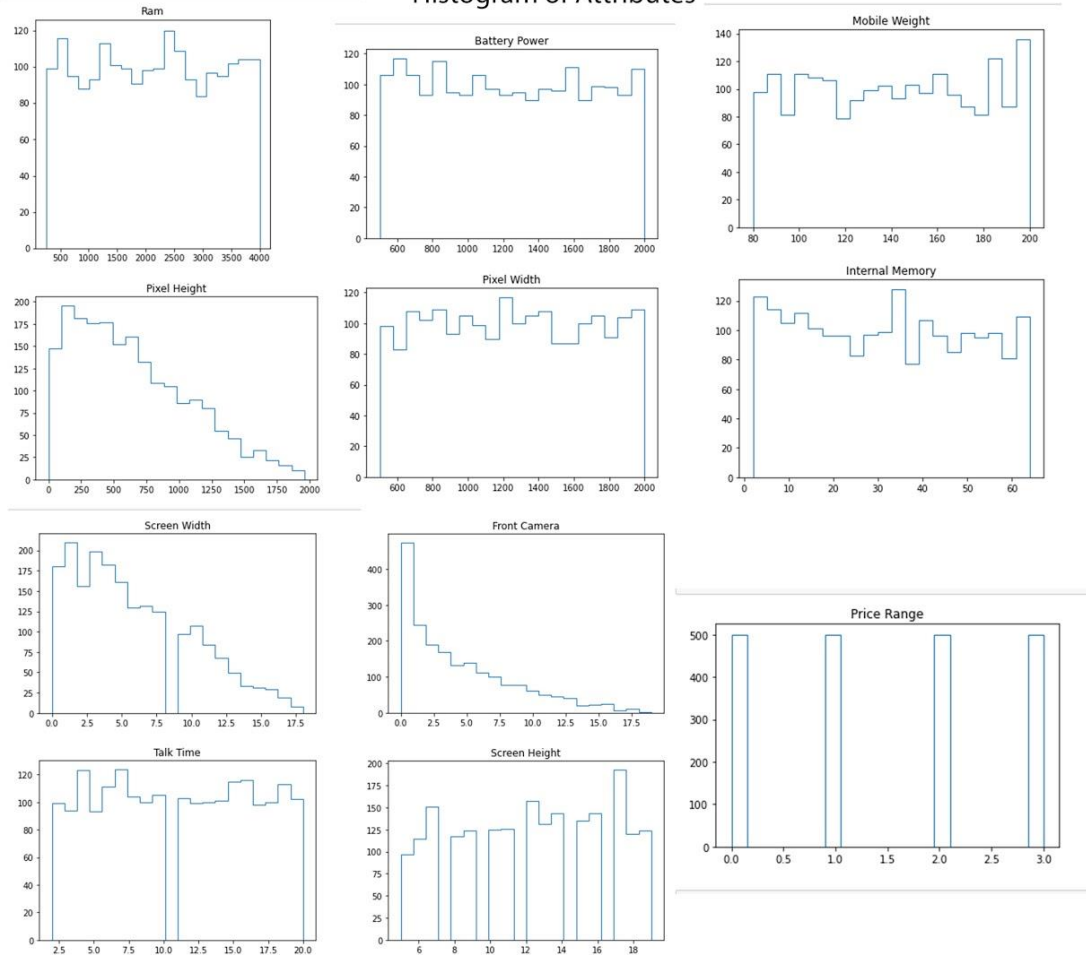


Figure 3: Histogram of Attributes

Attribute Analysis

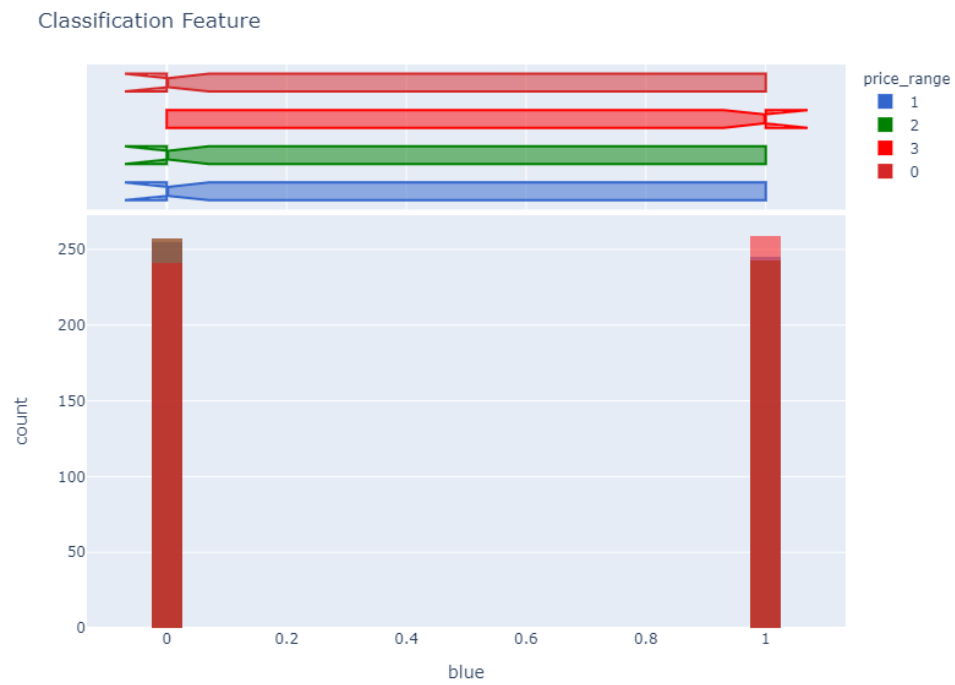


Figure 4: Bluetooth

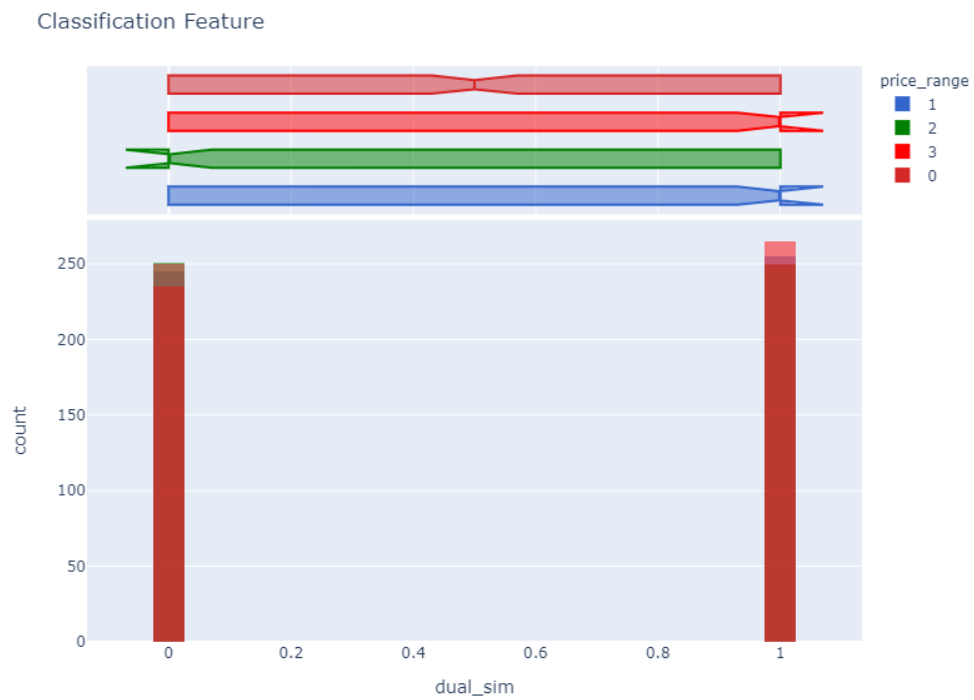


Figure 5: Dual Sim

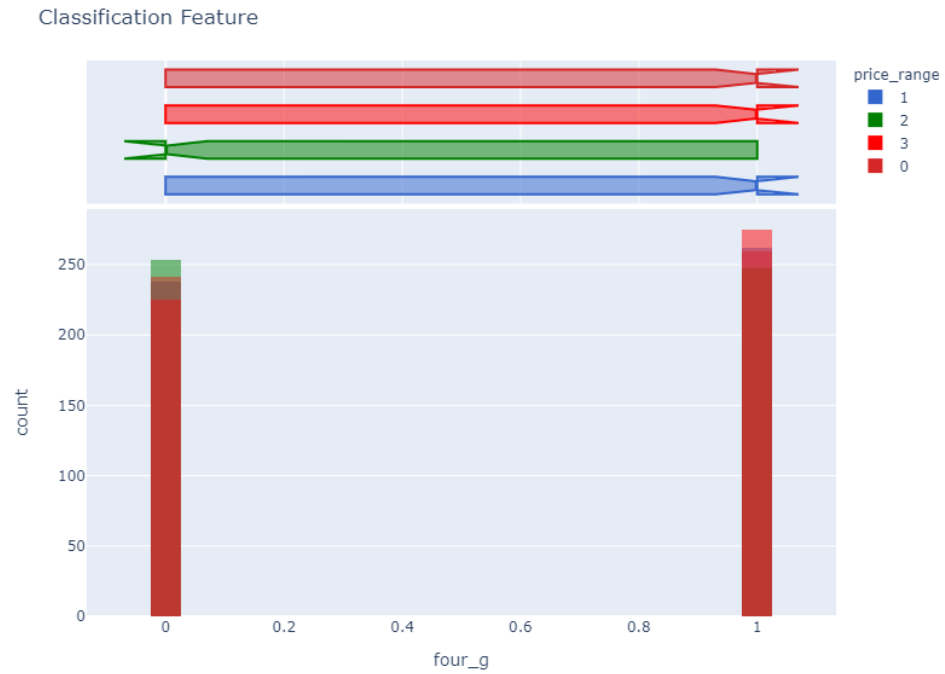


Figure 6: Four G

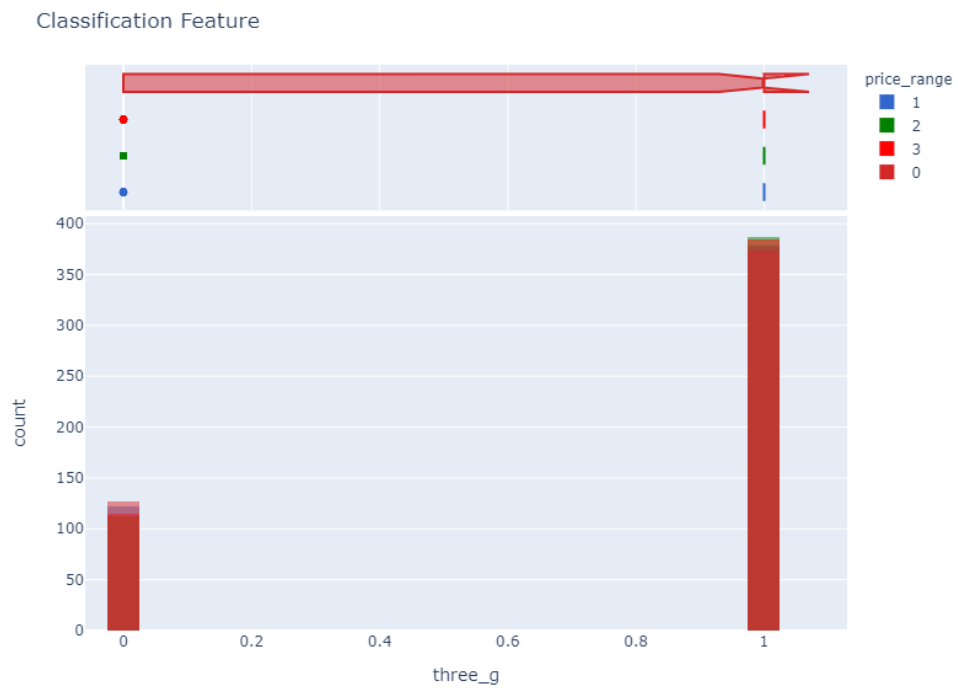


Figure 7: Three G

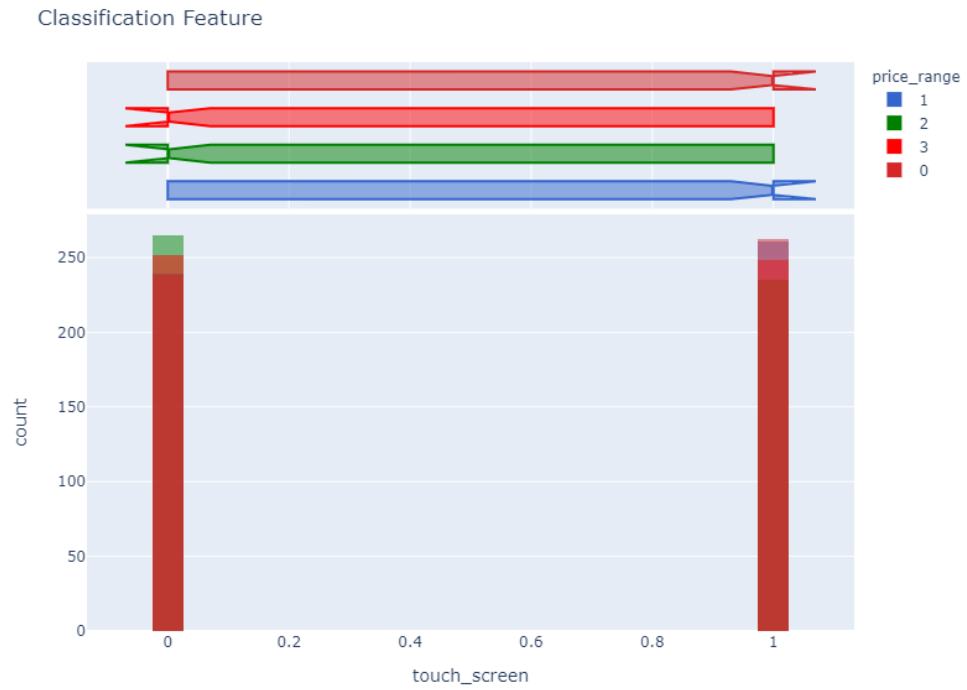


Figure 8: Touch Screen

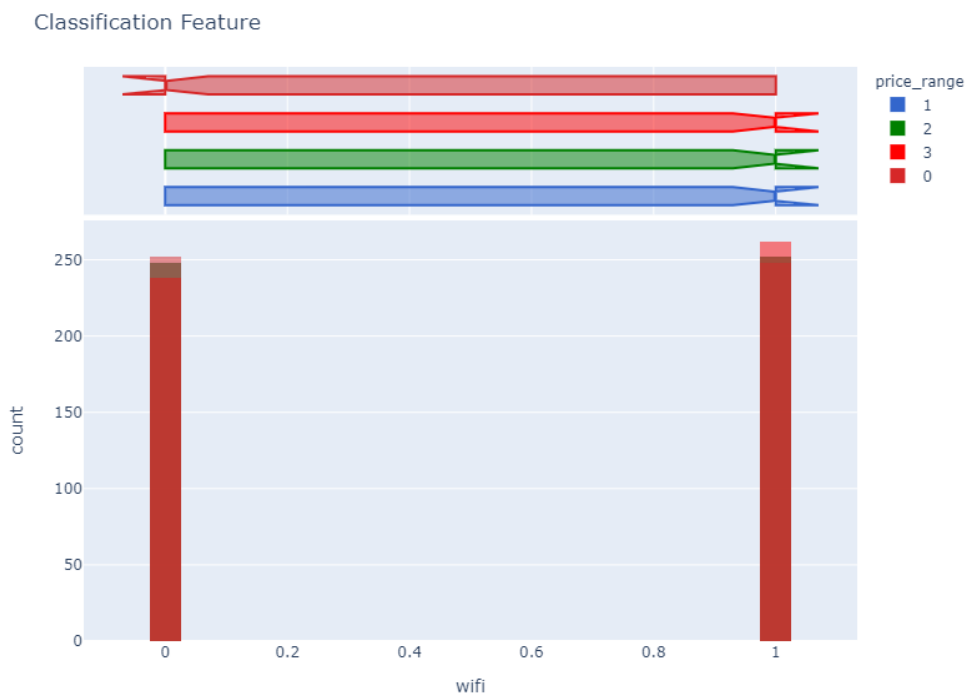


Figure 9: Wi-Fi

From Figure 4 to 9 it is clearly seen that how many data contain in the dataset for each classification label/class. Suppose, there are 250 phones of low price which does not have Wi-Fi and 251 phones of high price which have Wi-Fi.

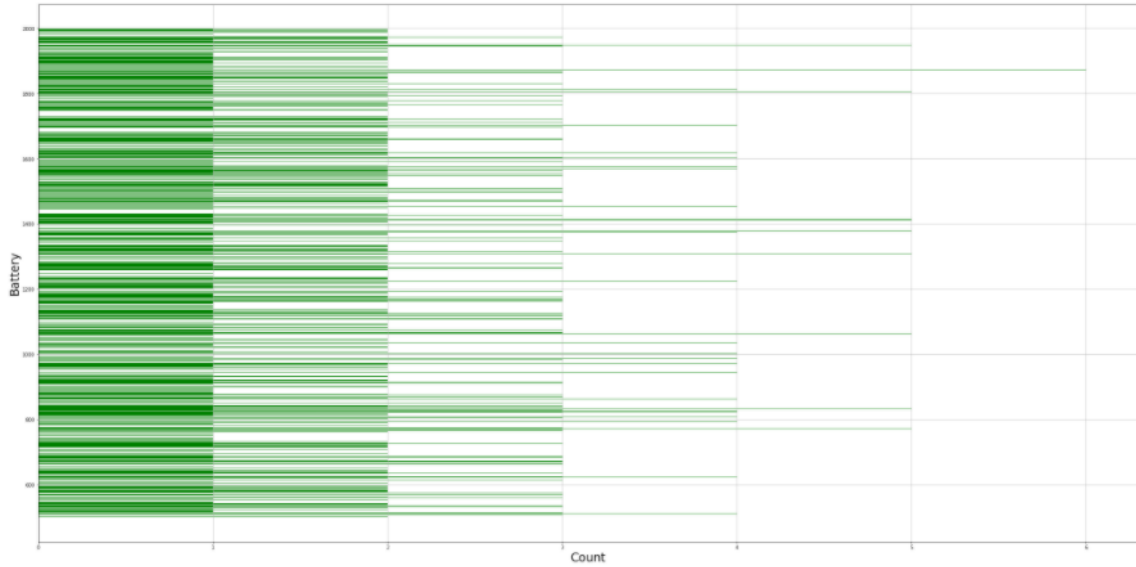


Figure 10: Battery Count

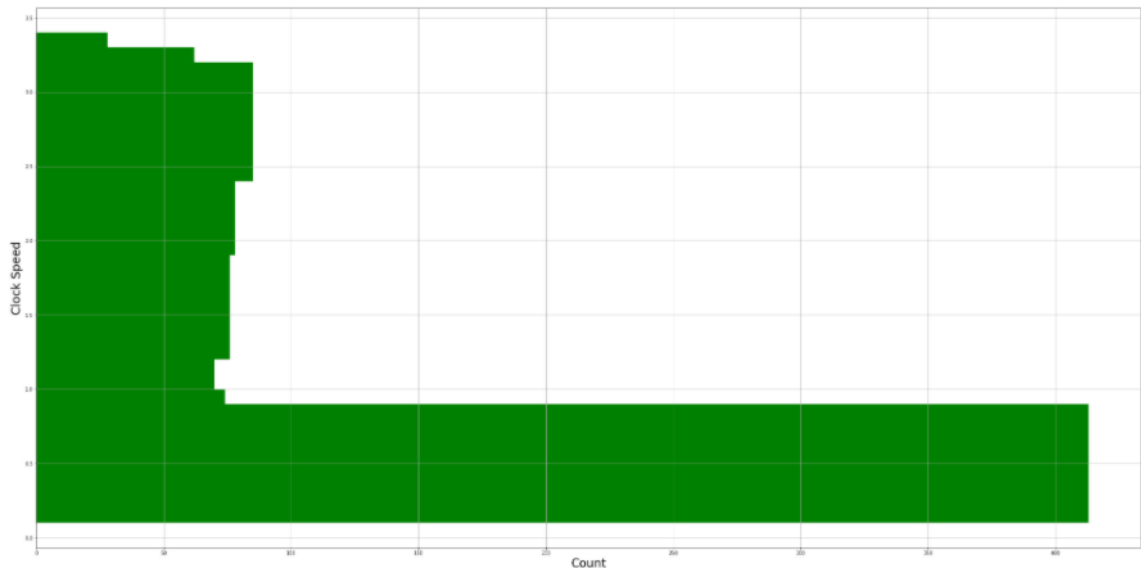


Figure 11: Clock Speed

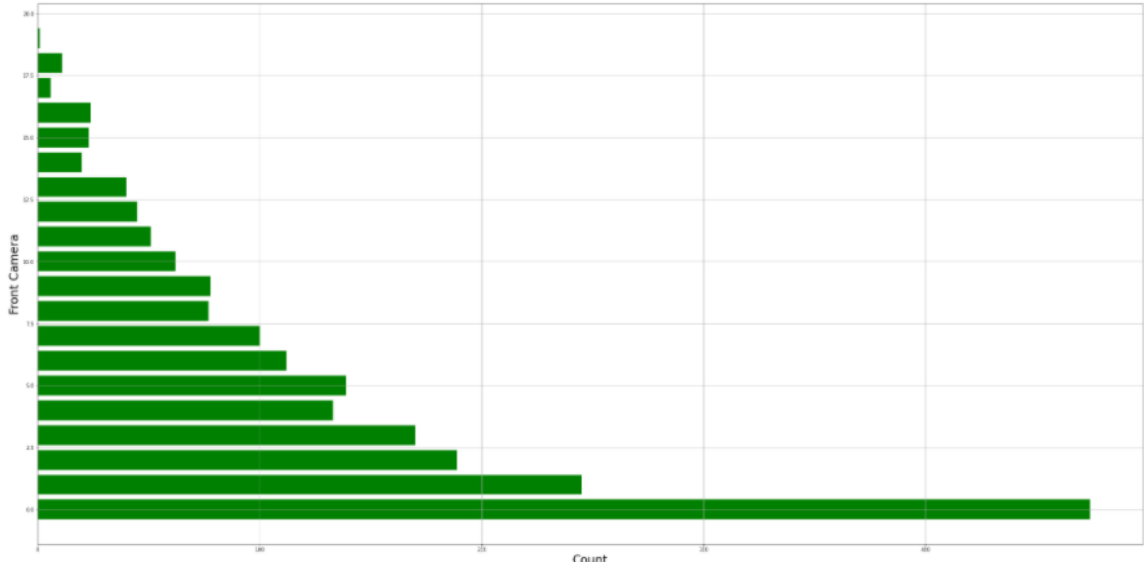


Figure 12: Front Camera

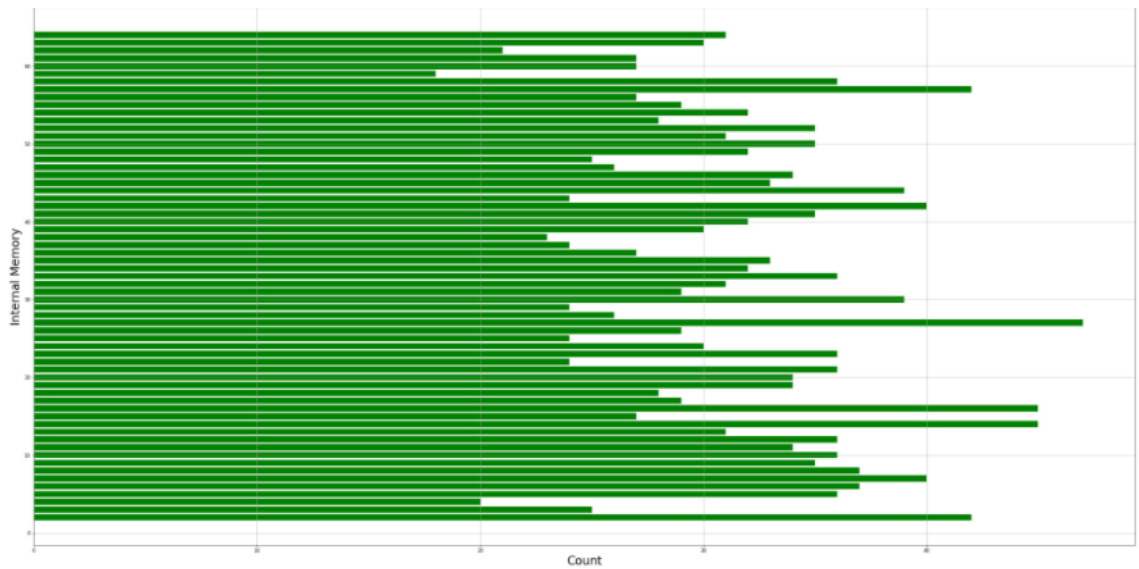


Figure 13: Internal Memory

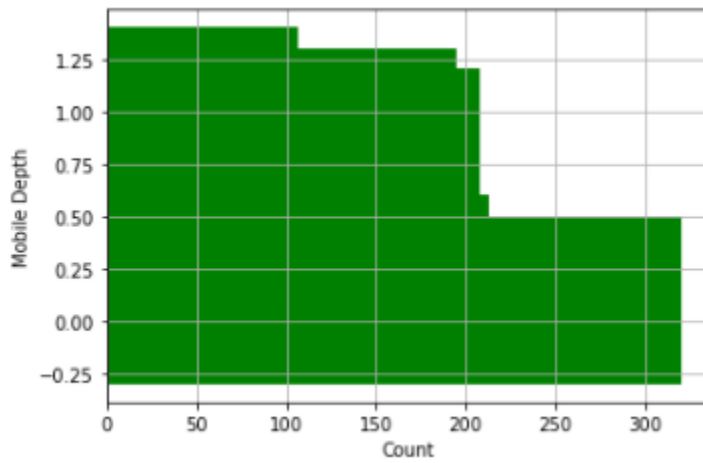


Figure 14: Mobile Depth

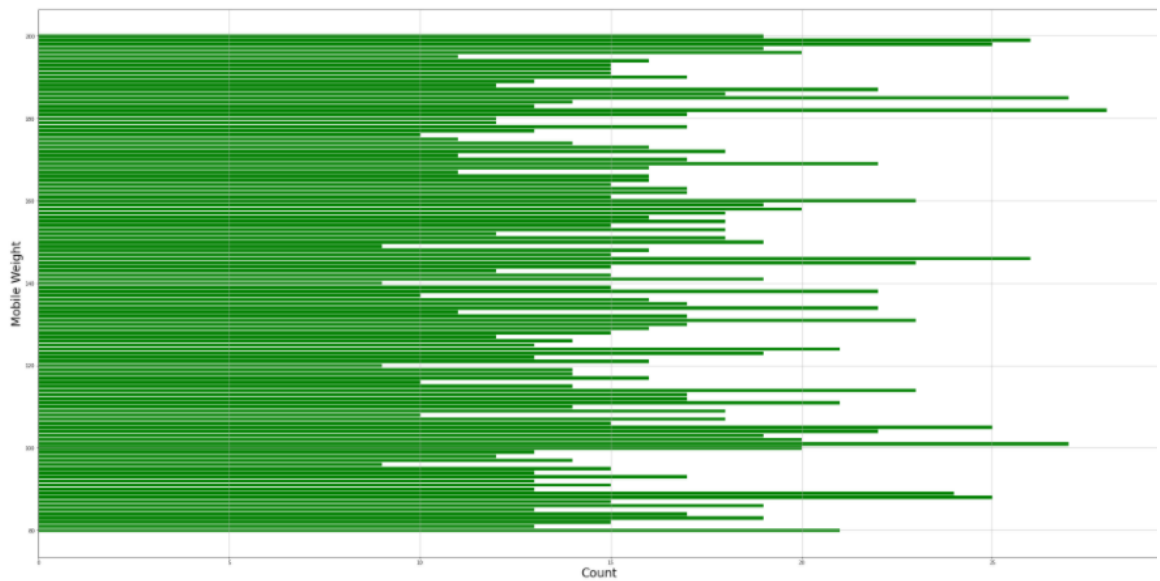


Figure 15: Mobile Weight

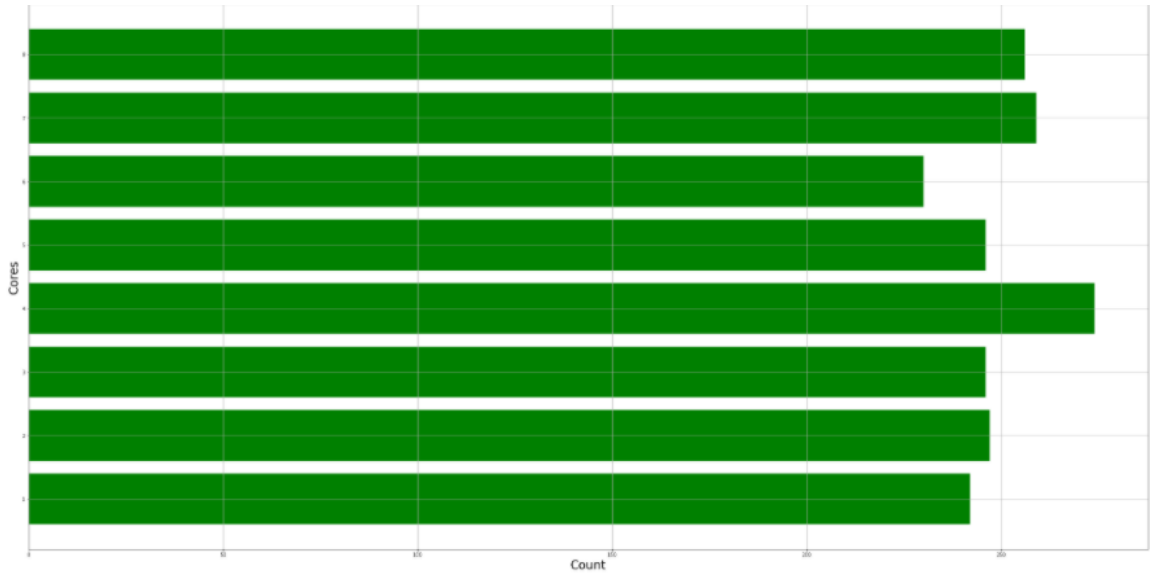


Figure 16: Cores

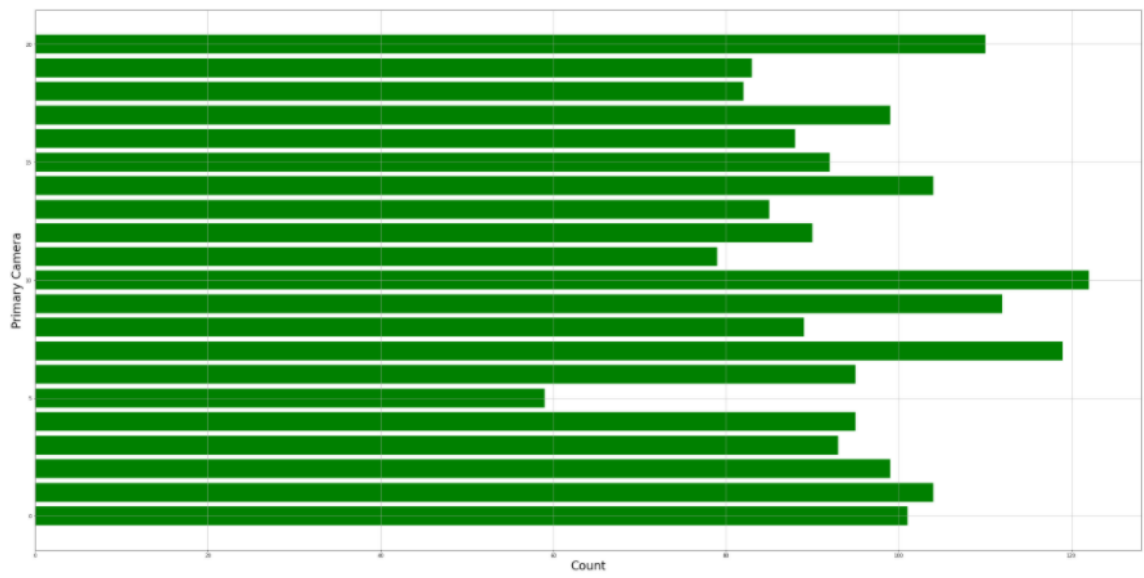


Figure 17: Primary Camera

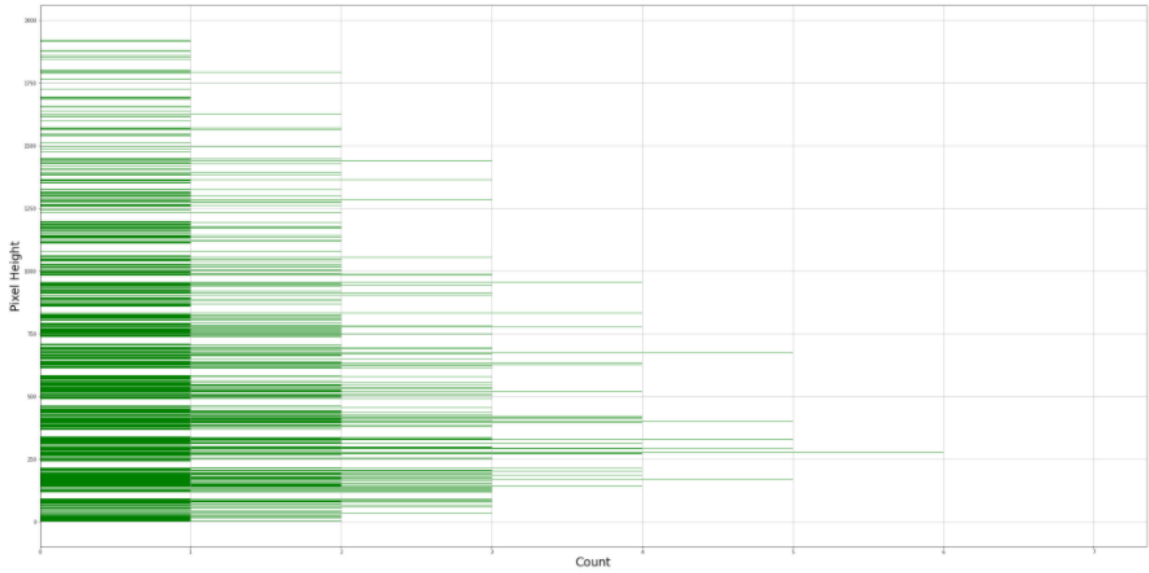


Figure 18: Pixel Height

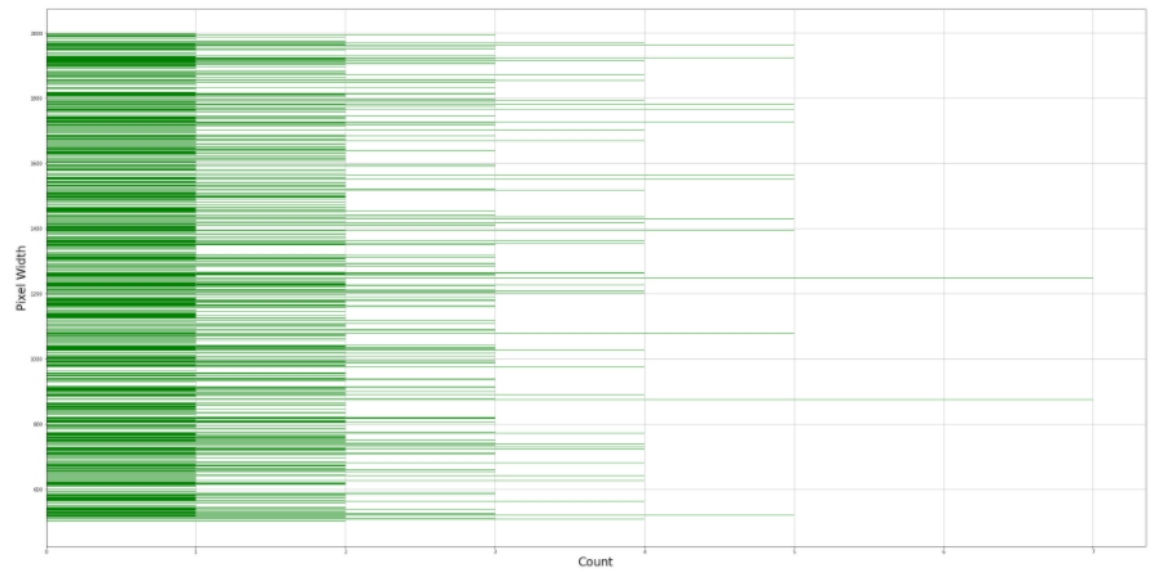


Figure 19: Pixel Width

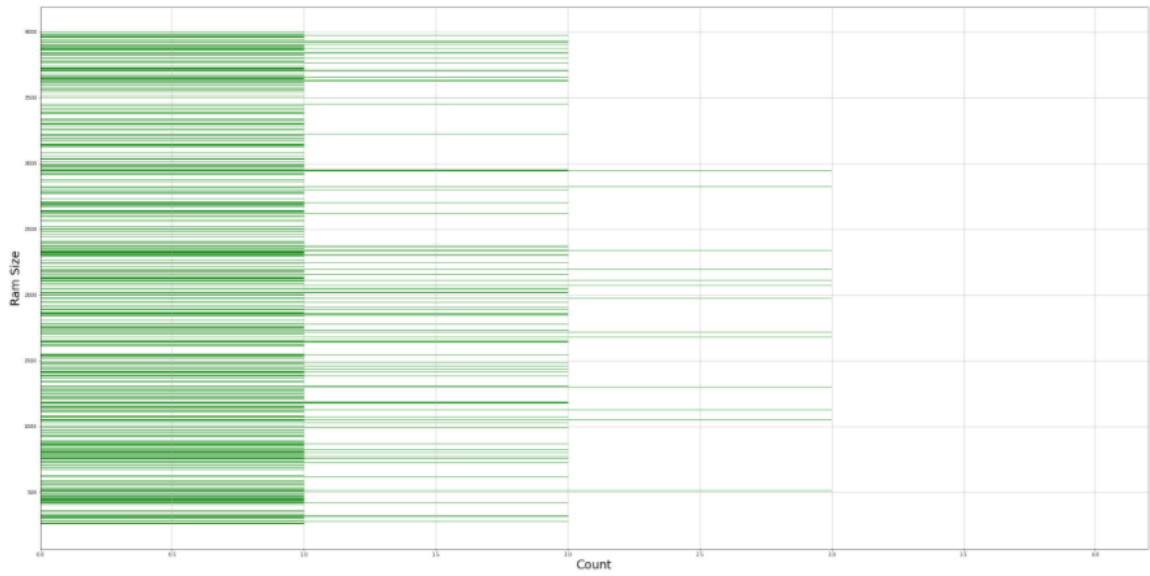


Figure 20: Ram

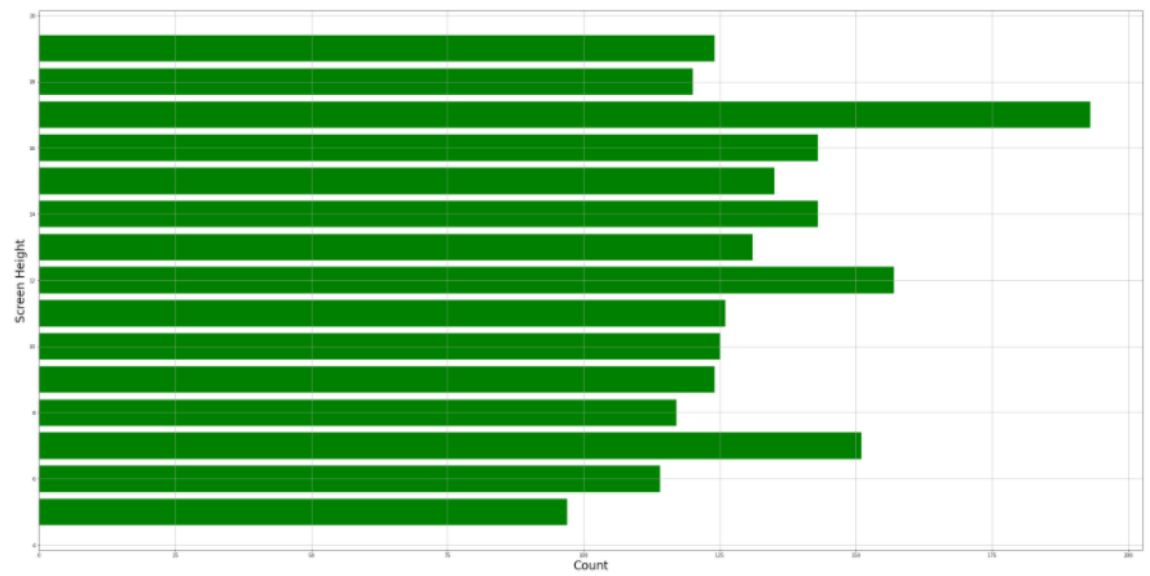


Figure 21: Screen Height

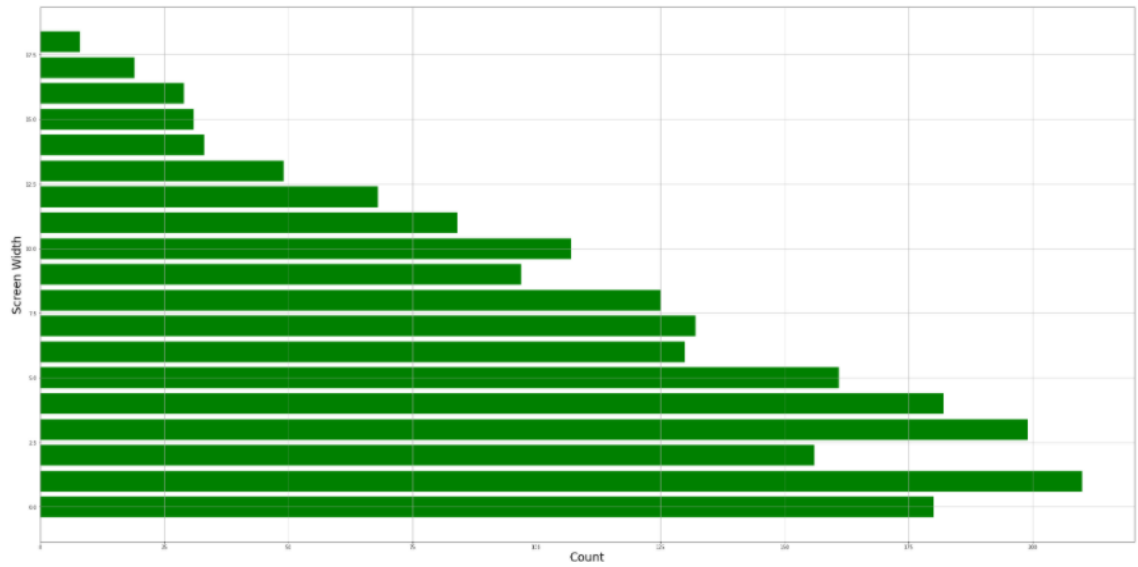


Figure 22: Screen Width

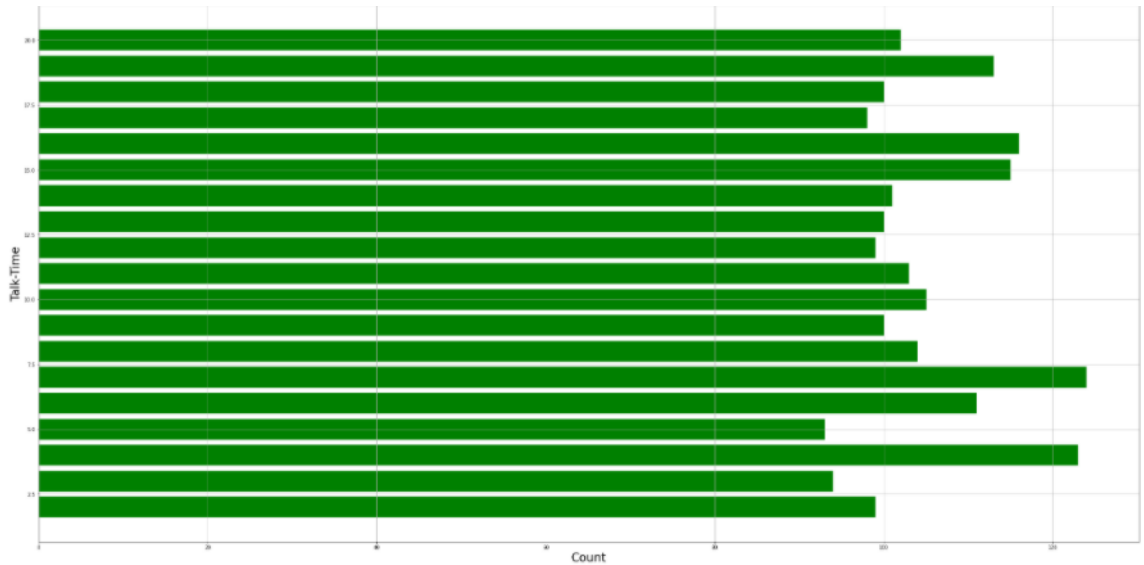


Figure 23: Talk-Time

And from Figure 10 to 23 it clear, how many types of data and for each type how many data contain in dataset. Suppose, ram can be 8192MB, 4096MB, 2048MB and for each of them there are 25,50,40 data in the dataset.

Correlation

This is a measure of the mutual relationship between attributes. By this measure the impact of an attribute depends on another attribute. Suppose, when it is summer people use to buy ice-cream (*Correlation for Data Science / Towards Data Science*, n.d.). Correlation gives a better understanding to a dataset (*Correlation – Towards Data Science*, n.d.). For example, from Fig-24 we can assume that pixel height and pixel width are mutually related to each other.

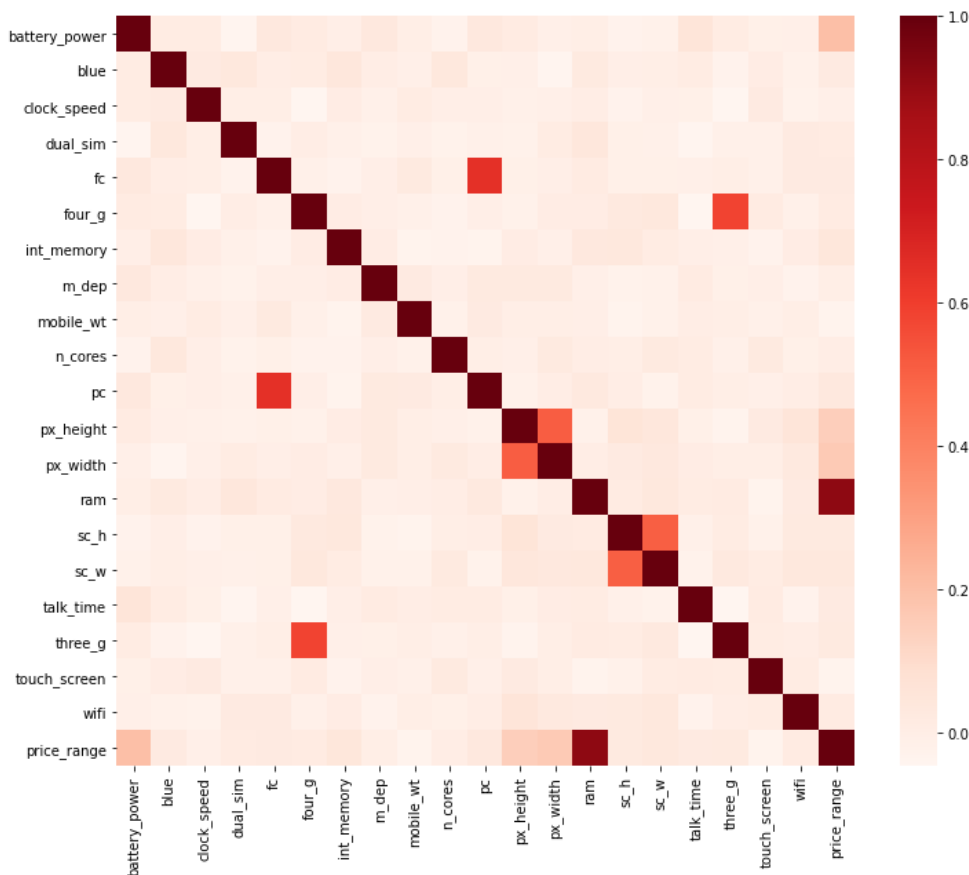


Figure 24: Correlation Heatmap

Selecting Features

There are many ways to select the best feature from a dataset. In this dataset, SelectKBest method is used and select 10 features according to their score. There are many features which are irrelevant or less important and for this sometimes the accuracy of the model has been decreased. So, for avoiding this problem best feature selecting technique is used.

Specs	ram	px_height	battery_power	px_width	mobile_wt	int_memory	sc_w	talk_time	fc	sc_h
Score	931268	17363.6	14129.9	9810.59	95.9729	89.8391	16.4803	13.2364	10.1352	9.61488

Figure 25: Best Features Score

Algorithms Applied

Decision Tree

Decision Trees are one of the most useful supervised learning algorithms. In supervised learning, data is already labeled and the target is known. DTs algorithms are best to solve classification problems(*The Complete Guide to Decision Trees | by Diego Lopez Yse | Towards Data Science*, n.d.). Decision tree algorithm and other classification algorithm seek to calculate the value of target attribute based on input attribute (Nwulu, 2017). There is a term called Entropy which is calculated by the probability of classes and then Information gain is calculated for identifying next node. Decision trees are definite nodes, branches and leaves. Each node defines the feature/attribute, branch defines the decisions and leaf defines the outcome. And also max depth based on the number of levels(*The Complete Guide to Decision Trees | by Diego Lopez Yse | Towards Data Science*, n.d.) Decision trees use entropy and information gain to solve the problem.

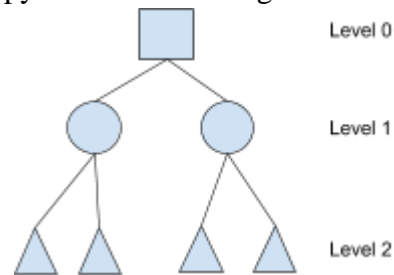


Figure 26: Decision Tree Levels

Entropy mathematical term

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

Equation 1: Entropy

Where, 'Pi' is the frequentist probability of an element/class 'i' in data(*Entropy: How Decision Trees Make Decisions | by Sam T | Towards Data Science*, n.d.).

Or

$$\text{Entropy}(S) = -P(Y)\log_2 P(Y) - P(N)\log_2 P(N)$$

Where, S=total number of samples, P(Y)=probability of Yes, P(N)=probability of No.

Information Gain is a measure of change in entropy which helps to differentiate the node and construct a decision tree.

$$\text{Gain} = \text{Entropy}(S) - [(\text{Weighted Avg}) * \text{Entropy}(\text{feature})]$$

Test Train Accuracy Decision Tree

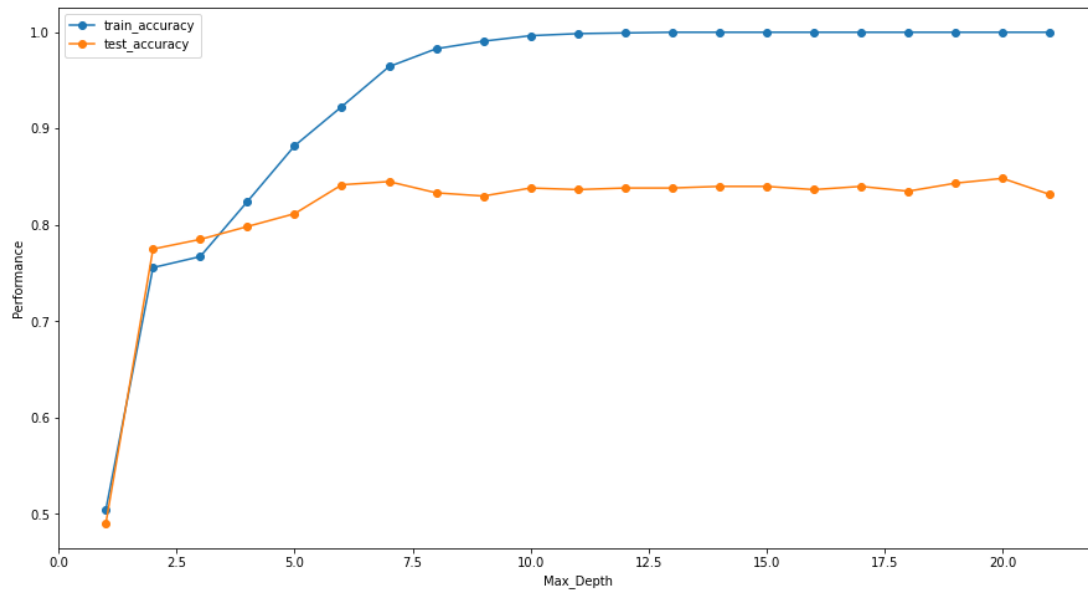


Figure 27: Test and Train Accuracy Using max_depth

From Fig-27, it can be said that when the depth is 20 it gives the best accuracy 84.83% for test results.

Confusion Matrix

Confusion Matrix is a summary of prediction results on a classification problem (*What Is a Confusion Matrix in Machine Learning*, n.d.). Using this matrix, the right and errors of the classification model can be visualized. The diagonal values are the accuracy of true positive the others are errors in this Fig-28. Based on this matrix classification report is created which contain accuracy, precision, recall/sensitivity.

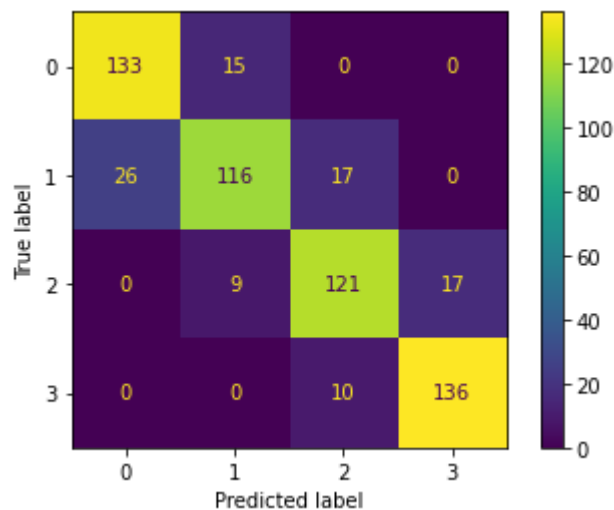


Figure 28: Confusion Matrix

Feature Importance

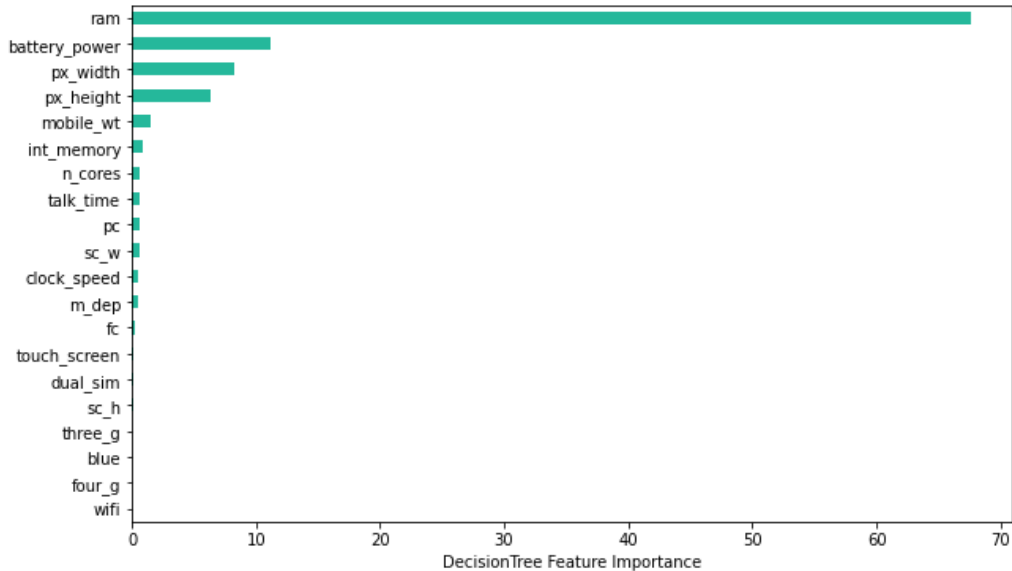


Figure 29: Feature Importance

From Fig-29 we can identify the importance of the feature to predict the price range of mobile.

Actually, it is not necessary after selecting best features. But it can visualize the attributes which plays best role in this dataset. It is clear that ram, battery power resolution, processor core, internal memory is much important.

Decision Tree Visualization

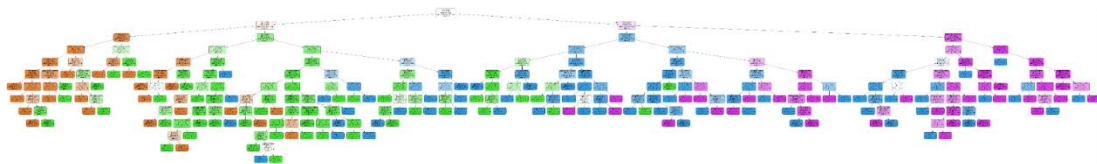


Figure 30: Decision Tree Graph

For better quality, go to this link: [DecisionTreeGraph\(Decision_tree_graphivz.Png - Google Drive, n.d.\)](#)



Figure 31: Pruned Decision Tree(entropy,depth=18)

Actual and Predictive value Visualization

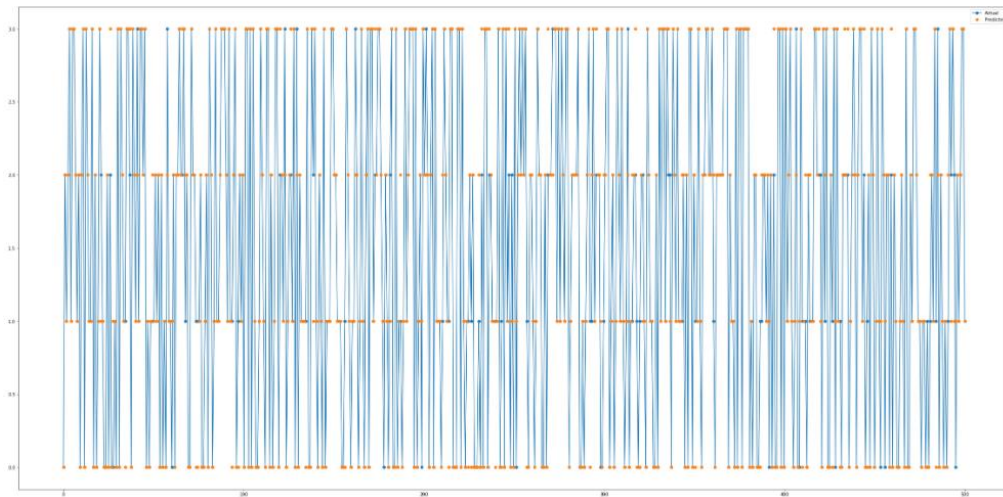


Figure 32: Actual vs Predicted values

From Fig-32 we can see the predicted and actual value. We can understand the error from this graph. In this graph blue lines are the actual value and orange dots are predicted value. So, where the blue line does not meet the orange dot there is the prediction error.

Random Forest

Random forest used to solve classification and regression problems. Random vectors are constructed ensembles of trees and decide the improved classification accuracy (Dogru & Subasi, 2018).

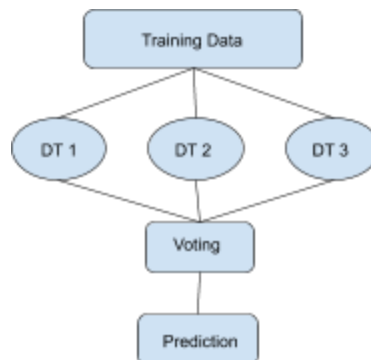


Figure 33: Random Forest

Actually, while random forest is implemented to a dataset it creates multiple trees with different score. Then it merge the trees and gives the best accuracy among them.

Tree Visualization

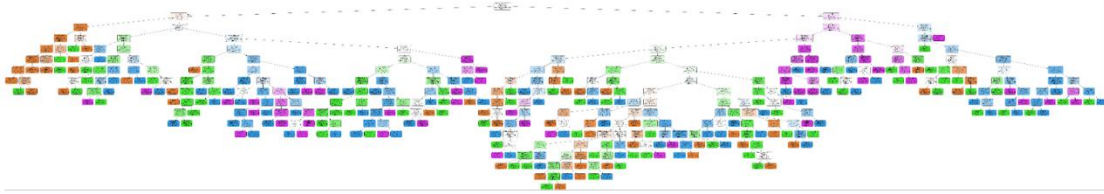


Figure 34: Tree of Random Forest

Mathematical Terms

When using Random forest for regression problems, root means square used to see how data branches from each node (*Random Forest Algorithm for Machine Learning* / by Madison Schott / *Capital One Tech* / Medium, n.d.).

$$MSE = \frac{1}{N} \sum_{i=1}^N (f_i - y_i)^2$$

Equation 2: Root Mean Square

Where, N is the number of data points, f_i is the value return by data model, y_i is the actual value for data point i .

For gini,

$$Gini = 1 - \sum_{i=1}^C (p_i)^2$$

Equation 3: Gini Impurity

Confusion Matrix

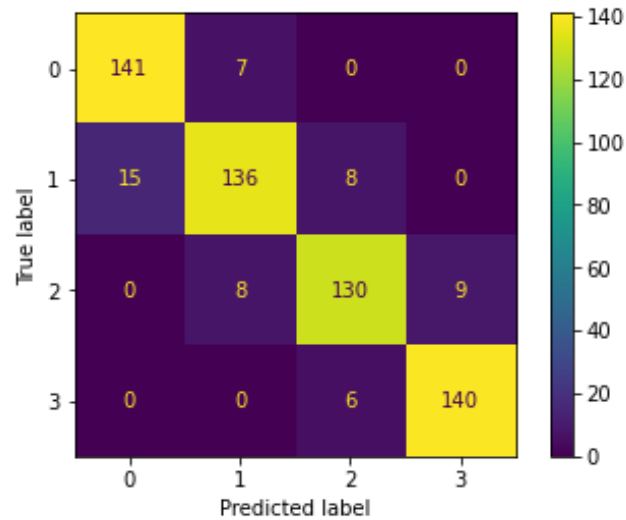


Figure 35: Confusion Matrix Random Forest

Actual vs Predictive Values

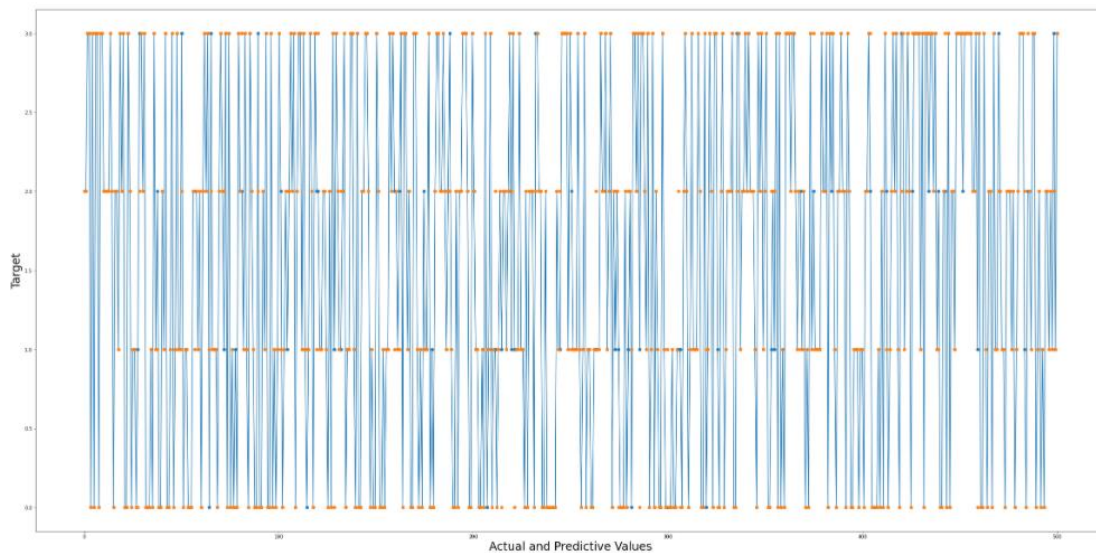


Figure 36: Actual vs Predictive Values

From Fig-36 we can see the predicted and actual value of Random Forest classifier. We can understand the error from this graph. In this graph blue lines are the actual value and orange dots are predicted value. So, where the blue line does not meet the orange dot there is the prediction error.

Tuning Random Forest

```
max_depth:10,estimator:25,trainacc:0.9985714285714286,testacc:0.9133333333333333
max_depth:10,estimator:26,trainacc:0.9985714285714286,testacc:0.9133333333333333
max_depth:10,estimator:27,trainacc:0.9985714285714286,testacc:0.9183333333333333
max_depth:10,estimator:28,trainacc:0.9992857142857143,testacc:0.9133333333333333
max_depth:10,estimator:29,trainacc:0.9992857142857143,testacc:0.915
max_depth:10,estimator:30,trainacc:0.9992857142857143,testacc:0.9133333333333333
max_depth:10,estimator:31,trainacc:0.9992857142857143,testacc:0.915
max_depth:10,estimator:32,trainacc:0.9992857142857143,testacc:0.92
max_depth:10,estimator:33,trainacc:0.9985714285714286,testacc:0.9216666666666666
max_depth:10,estimator:34,trainacc:0.9985714285714286,testacc:0.9166666666666666
max_depth:11,estimator:1,trainacc:0.835,testacc:0.67
max_depth:11,estimator:2,trainacc:0.8814285714285715,testacc:0.7166666666666667
max_depth:11,estimator:3,trainacc:0.9557142857142857,testacc:0.8133333333333334
max_depth:11,estimator:4,trainacc:0.9664285714285714,testacc:0.83
max_depth:11,estimator:5,trainacc:0.9742857142857143,testacc:0.8366666666666667

max_depth:11,estimator:6,trainacc:0.9807142857142858,testacc:0.855
max_depth:11,estimator:7,trainacc:0.9871428571428571,testacc:0.855
max_depth:11,estimator:8,trainacc:0.9878571428571429,testacc:0.8666666666666667
max_depth:11,estimator:9,trainacc:0.9921428571428571,testacc:0.87
```

Figure 37 : Tuned Random Forest Accuracy

From Fig37, it is clear to be said that if Random forest is pruned with max_depth 10 and estimator 33 it will give the best accuracy.

Chapter 4

Performance Evaluation

After implementing models and getting output, the next step is to find out the effectiveness of the model based on some metric using test datasets. Different performance metrics used to evaluate different models. For evaluating we will choose precision, recall and accuracy.

Accuracy

Accuracy is the ratio of correct prediction respective to all data. It is a faster way to evaluate a set of predictions in a classification problem.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{All Samples}}$$

Equation 4: Accuracy

Precision

Precision is the ratio of correct prediction respected to correct and incorrect prediction. It defines that the correct prediction it makes is actually correct or incorrect based on data. Suppose, algorithm identify a number of people who has cancer and actually how many of them has cancer, the ration between these terms can be said precision.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Equation 5: Precision

Recall/Sensitivity

It means that a model exactly predicts correctly from true value.

Suppose, algorithm correctly predict a number of people who actually has cancer and actually how many of them has cancer, the ratio between two terms can be said recall.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negative}}$$

Equation 6: Recall

ROC Curve Decision Tree

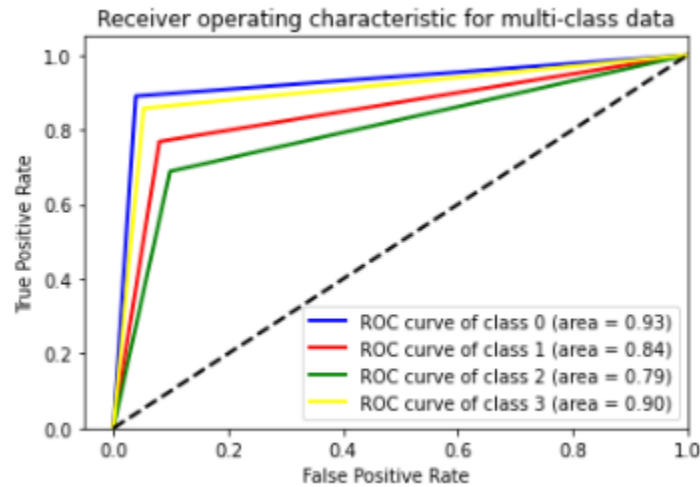


Figure 38: ROC Curve Decision Tree

From this curve it is clear that the curve for class 0 is 0.93, 1 is 0.84, 2 is 0.79, 3 is 0.90 using True positive rate and False positive rate.

Random Forest

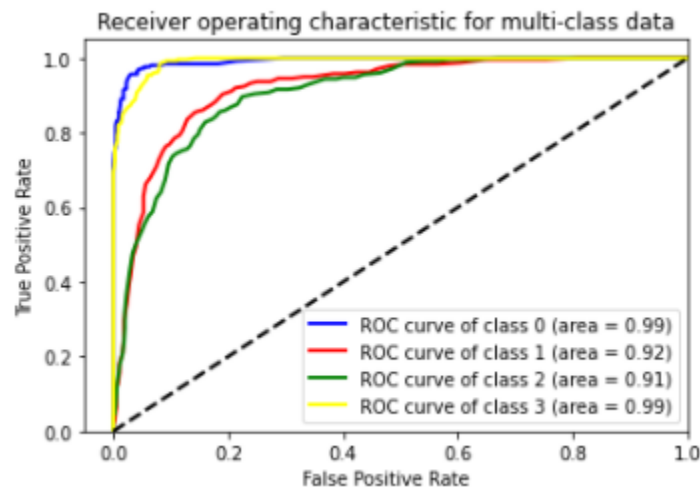


Figure 39: ROC Curve Random Forest

From Fig-24, it is seen that the curve for class 0 is 0.99, 1 is 0.92, 2 is 0.91, 3 is 0.99 in the perspective of True positive rate and False positive rate.

Actually, ROC curve can explain and visualize the performance of each class for different model that's why it is used.

Chapter 5

Result and Discussion

In this study, two classification techniques were implemented. In this section, all the experiment results will be cast and discussed.

Table 2: Performance Comparison between models

Evaluation Metrics	Accuracy	Precision	Recall																
Decision Tree (All Features)	83%	<table><tr><td>0</td><td>0.85</td></tr><tr><td>1</td><td>0.83</td></tr><tr><td>2</td><td>0.79</td></tr><tr><td>3</td><td>0.84</td></tr></table>	0	0.85	1	0.83	2	0.79	3	0.84	<table><tr><td>0</td><td>0.83</td></tr><tr><td>1</td><td>0.75</td></tr><tr><td>2</td><td>0.78</td></tr><tr><td>3</td><td>0.90</td></tr></table>	0	0.83	1	0.75	2	0.78	3	0.90
0	0.85																		
1	0.83																		
2	0.79																		
3	0.84																		
0	0.83																		
1	0.75																		
2	0.78																		
3	0.90																		
Decision Tree (Best Features)	85% 87% (Pruning Accuracy using entropy, max_depth18)	<table><tr><td>0</td><td>0.85</td></tr><tr><td>1</td><td>0.84</td></tr><tr><td>2</td><td>0.82</td></tr><tr><td>3</td><td>0.88</td></tr></table>	0	0.85	1	0.84	2	0.82	3	0.88	<table><tr><td>0</td><td>0.91</td></tr><tr><td>1</td><td>0.74</td></tr><tr><td>2</td><td>0.81</td></tr><tr><td>3</td><td>0.94</td></tr></table>	0	0.91	1	0.74	2	0.81	3	0.94
0	0.85																		
1	0.84																		
2	0.82																		
3	0.88																		
0	0.91																		
1	0.74																		
2	0.81																		
3	0.94																		
Random Forest (All Features)	90%	<table><tr><td>0</td><td>0.90</td></tr><tr><td>1</td><td>0.89</td></tr><tr><td>2</td><td>0.89</td></tr><tr><td>3</td><td>0.91</td></tr></table>	0	0.90	1	0.89	2	0.89	3	0.91	<table><tr><td>0</td><td>0.95</td></tr><tr><td>1</td><td>0.86</td></tr><tr><td>2</td><td>0.83</td></tr><tr><td>3</td><td>0.95</td></tr></table>	0	0.95	1	0.86	2	0.83	3	0.95
0	0.90																		
1	0.89																		
2	0.89																		
3	0.91																		
0	0.95																		
1	0.86																		
2	0.83																		
3	0.95																		
Random Forest (Best Features)	90% 92.16% (Pruned using max_depth 10, estimator 33)	<table><tr><td>0</td><td>0.90</td></tr><tr><td>1</td><td>0.89</td></tr><tr><td>2</td><td>0.88</td></tr><tr><td>3</td><td>0.92</td></tr></table>	0	0.90	1	0.89	2	0.88	3	0.92	<table><tr><td>0</td><td>0.95</td></tr><tr><td>1</td><td>0.86</td></tr><tr><td>2</td><td>0.87</td></tr><tr><td>3</td><td>0.94</td></tr></table>	0	0.95	1	0.86	2	0.87	3	0.94
0	0.90																		
1	0.89																		
2	0.88																		
3	0.92																		
0	0.95																		
1	0.86																		
2	0.87																		
3	0.94																		

From Table-2, it is clear that random forest has the highest performance in terms of the Evaluation metrics. The accuracy of Random Forest is 91% where Decision Tree has the accuracy of 84%.

Table 3: ROC Curve Comparison

ML Technique	Class 0	Class 1	Class 2	Class 3
Decision Tree	0.93	0.84	0.79	0.90
Random Forest	0.99	0.92	0.91	0.99

Chapter 6

Conclusion and Future Work

Findings

In this study, we performed correlation to see how the features are related to each other and also for better understanding it is shown in a heatmap. Here, 10 best features selected which are mostly related for further work. Three machine learning techniques implemented for this study are Naïve Bayes, Decision Tree and Random Forest. From these three Random Forest gives the best accuracy of 91% and Decision Tree gives 84% accuracy because I used 10 attributes which are mostly related or connected. But, the others used less attribute for which they didn't get a good accuracy and one of the researchers(Asim & Khan, 2018) (*GSMArena.Com - Mobile Phone Reviews, News, Specifications and More...*, n.d.) dataset was different from the others(Balakumar et al., n.d.; *Kaggle: Your Home for Data Science*, n.d.-a; Pipalia & Bhadja, n.d.).

Limitation

Nowadays, well configured phones are getting at a lower price so the price level depending on quality has been changed.

Future Work

It can be taken to application level. More sophisticated NN or combining different algorithms with greater accuracy can be found.

References

- 29+ Smartphone Usage Statistics: Around the World in 2020. (n.d.).
<https://lefronic.com/smartphone-usage-statistics/>
- Al-Daour, A. F., Al-Shawwa, M. O., & Abu-Naser, S. S. (2020). *Banana Classification Using Deep Learning*.
- Al-Shawwa, M. O., Al-Absi, A. A.-R., Hassanein, S. A., Baraka, K. A., & Abu-Naser, S. S. (2018). *Predicting Temperature or Humidity in the Surrounding Environment Using Artificial Neural Network*.
- Alghoul, A., Al Ajrami, S., Al Jarousha, G., Harb, G., & Abu-Naser, S. S. (2018). *Email Classification Using Artificial Neural Network*.
- Asim, M., & Khan, Z. (2018). Mobile Price Class prediction using Machine Learning Techniques. *International Journal of Computer Applications*, 975, 8887.
- Balakumar, B., Raviraj, P., & Gowsalya, V. (n.d.). *Mobile Price prediction using Machine Learning Techniques*.
- Correlation – Towards Data Science*. (n.d.). <https://towardsdatascience.com/tagged/correlation>
- Correlation for data science | Towards Data Science*. (n.d.).
<https://towardsdatascience.com/what-it-takes-to-be-correlated-ce41ad0d8d7f>
- decision_tree_graphivz.png - Google Drive*. (n.d.). https://drive.google.com/file/d/1ZML_-0PFdfnePoCUCUR4pQao029sVpvd/view
- Di Persio, L., & Honchar, O. (2016). Artificial neural networks architectures for stock price prediction: Comparisons and applications. *International Journal of Circuits, Systems and Signal Processing*, 10(2016), 403–413.
- Dogru, N., & Subasi, A. (2018). Traffic accident detection using random forest classifier. *2018 15th Learning and Technology Conference (L&T)*, 40–45.
- Ebrahimian, H., Barmayoon, S., Mohammadi, M., & Ghadimi, N. (2018). The price prediction for the energy market based on a new method. *Economic Research-Ekonomska Istraživanja*, 31(1), 313–337.
- Entropy: How Decision Trees Make Decisions | by Sam T | Towards Data Science*. (n.d.).
<https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8>
- GSMArena.com - mobile phone reviews, news, specifications and more...* (n.d.). Retrieved February 3, 2021, from <https://www.gsmarena.com/>
- Kaggle: Your Home for Data Science*. (n.d.-a). Retrieved January 10, 2021, from <https://www.kaggle.com/>
- Kaggle: Your Home for Data Science*. (n.d.-b). <https://www.kaggle.com/>
- Khillha, F., & Shawwa, N. (2020). *ANN for Predicting Mobile Phone Price Range*.
- Lim, W. T., Wang, L., Wang, Y., & Chang, Q. (2016). Housing price prediction using neural

- networks. *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 518–522.
- Mobile technology* | IBM. (n.d.). <https://www.ibm.com/topics/mobile-technology>
- Nwulu, N. I. (2017). A decision trees approach to oil price prediction. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–5.
- Pipalia, K., & Bhadja, R. (n.d.). *Performance Evaluation of Different Supervised Learning Algorithms for Mobile Price Classification*.
- Random Forest Algorithm for Machine Learning* | by Madison Schott | Capital One Tech | Medium. (n.d.). <https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb>
- The Complete Guide to Decision Trees* | by Diego Lopez Yse | Towards Data Science. (n.d.). <https://towardsdatascience.com/the-complete-guide-to-decision-trees-28a4e3c7be14>
- What is a Confusion Matrix in Machine Learning*. (n.d.). <https://machinelearningmastery.com/confusion-matrix-machine-learning/>
- Zehtab-Salmasi, A., Feizi-Derakhshi, A.-R., Nikzad-Khasmakhi, N., Asgari-Chenaghlu, M., & Nabipour, S. (2020). Multimodal price prediction. *ArXiv Preprint ArXiv:2007.05056*.

Appendix

1/31/2021

Turnitin

Turnitin Originality Report

Processed on: 31-Jan-2021 16:22 +06
ID: 1498067748
Word Count: 3523
Submitted: 1

171-35-1838 By Ahsanul Hoque
Sakib

Similarity Index

13%

Similarity by Source

Internet Sources: 10%
Publications: 4%
Student Papers: 9%

4% match (student papers from 04-Apr-2018)

Class: Article 2018

Assignment: Journal Article

Paper ID: [940891790](#)

1% match (Internet from 11-Dec-2020)

<https://towardsdatascience.com/the-complete-guide-to-decision-trees-28a4e3c7be14?gi=a3455f10b1b7>

1% match (Internet from 13-Jan-2021)

<http://dstore.alazhar.edu.ps/xmlui/bitstream/handle/123456789/45/NASDAN-2v1.pdf>

1% match (Internet from 01-Jan-2021)

<https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b>

1% match (student papers from 01-Sep-2020)

[Submitted to University of Surrey on 2020-09-01](#)

1% match (Internet from 02-Mar-2017)

<http://researchbank.rmit.edu.au/eserv/rmit:6372/Liu.pdf>

1% match (publications)

["Data Analytics and Management", Springer Science and Business Media LLC, 2021](#)

1% match (publications)

[Talha Ahmed Khan, Kushsairy A., Shahzad Nasim, Muhammad Alam, Zeeshan Shahid, M.S Mazliham, "Proficiency Assessment of Machine Learning Classifiers: An Implementation for the Prognosis of Breast Tumor and Heart Disease Classification", International Journal of Advanced Computer Science and Applications, 2020](#)

< 1% match (Internet from 17-Mar-2019)

<https://philpapers.org/archive/NASANN.docx>

< 1% match (Internet from 07-Nov-2020)

<https://www.coursehero.com/file/p3gcg047/Associate-Professor-Richard-Dazeley-Email-RichardDazeleydeakineduau-School-of/>

< 1% match (student papers from 20-Oct-2020)

[Submitted to University of Puerto Rico-Mayaguez on 2020-10-20](#)

< 1% match (Internet from 04-Jan-2021)

https://medium.com/@gp_pulipaka/an-essential-guide-to-classification-and-regression-trees-in-r-language-4ced657d176b#:~:text=The%20primary%20difference%20between%20classification,ordered%20val

< 1% match (publications)

https://www.turnitin.com/newreport_printview.asp?eq=1&eb=1&esm=10&oid=1498067748&sid=0&n=0&m=2&svr=48&r=22.773396245565113&lang=e... 1/8