Capstone Project- 3

Mobile Price Range Prediction

Presented

by

SUDHANSHU KUMAR

(Sudhanshu.mriu@gmail.com)

Contents:

Problem statement

- For any product-based company, their objective is to increase the value of the product.

$$\text{Value} = \frac{\text{Functions(Features)}}{\text{Price}}$$

- A lot of factors take into consideration before buying a mobile device because we use them for a plethora of purposes, like keeping in touch with family (calling), playing games, and capturing pictures to save our memories.

- Our objective is to help the enterprise to determine the pricing range of mobiles based on features using the various machine learning algorithm, which will increase the value of the phone at the lowest possible cost and with the most possible profit.

Data description

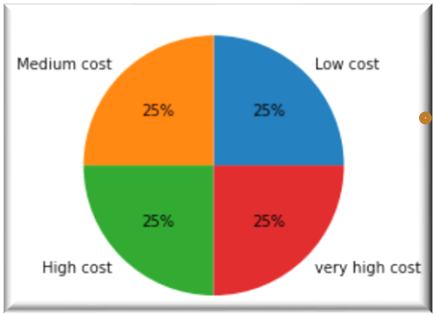
```
mob_df.shape
```

→ `(2000, 21)`

Our data set have 2000 rows and 21 features

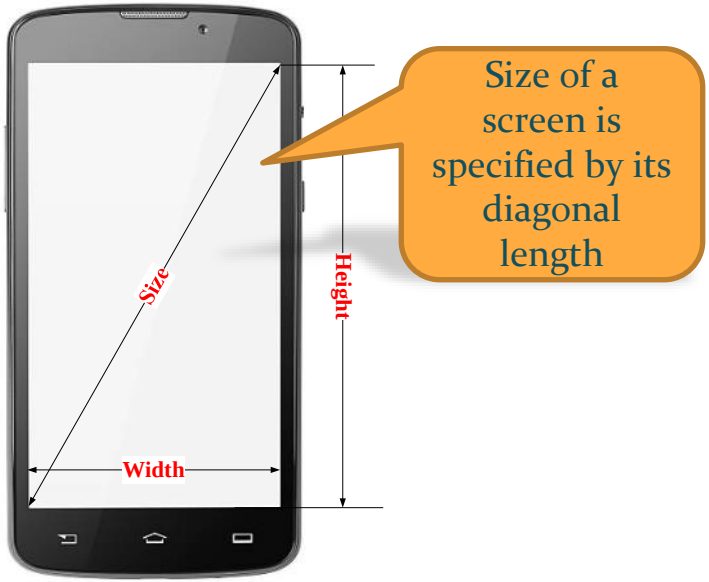
21 features are:

1) Battery_power (in mAh)
2) Blue
3) Clock_speed
4) Dual_sim
5) Fc
6) Four_g
7) Int_memory (in GB)
8) M_dep (Mobile Depth in cm)
9) Mobile_wt
10) N_cores
11) Pc (Primary Camera mega pixels)
12) Px_height (Pixel Resolution Height)
13) Px_width (Pixel Resolution Width)
14) Ram (in Mega Bytes)
15) Sc_h (in cm)
16) Sc_w (in cm)
17) Talk_time
18) Three_g
19) Touch_screen
20) Wifi
21) Price_range – 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).



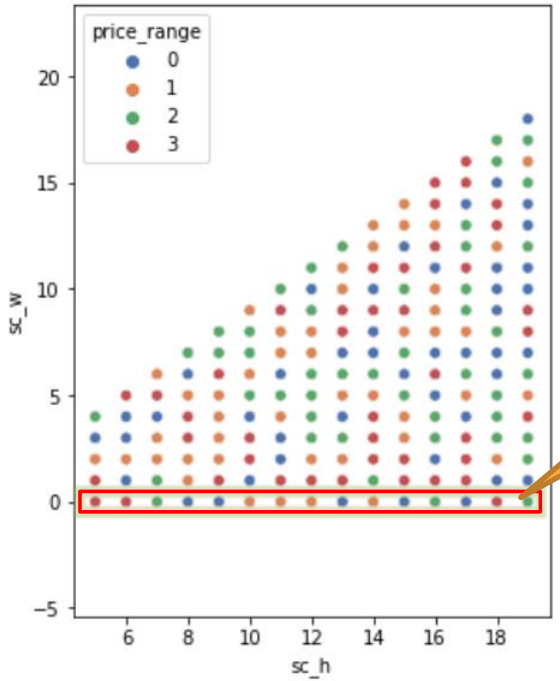
Balanced dataset

Data Discrepancy and Feature Engineering



Size of a screen is specified by its diagonal length

We created a new feature “screen size” and dropping either screen width or screen height to avoid multicollinearity.



Every mobile must have some positive screen width and screen height values

179

```
print(mob_df[mob_df['sc_w']==0].shape[0])
```

```
mob_df['sc_size'] = ((mob_df['sc_h']**2)+(mob_df['sc_w']**2))**0.5
```

```
mob_df=mob_df.drop(['sc_h'], axis=1)
```

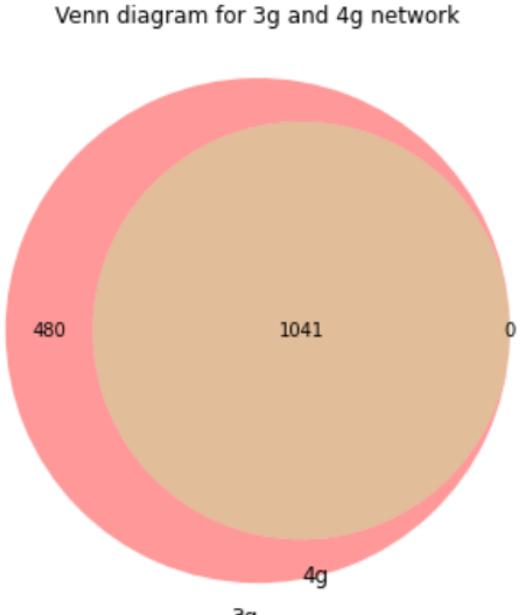
Dropping sc_h because Screen size is highly correlated with **sc_h**

Exploratory data analysis

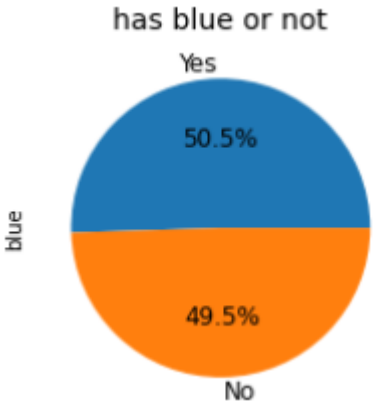
Venn Diagram for various network compatibilities

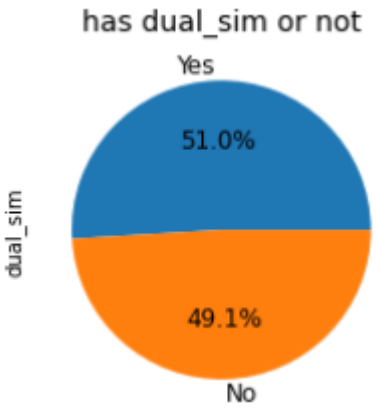
- mobile which have 4g network compatibility also have 3g network compatibility.
- 480 phone have only 3g features.
- 477 phone have neither 3g nor 4g network compatibility.

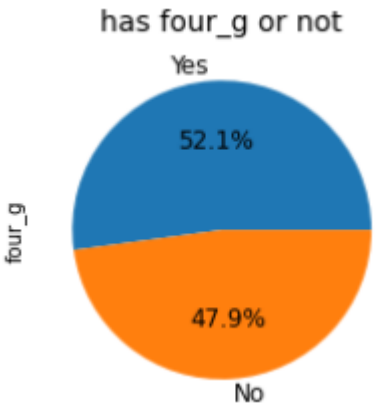
Network	Counts	Percentage
only 4g	0	0.000
only3g	480	24.024
3g and 4g	1041	52.102
No 3g and No 4g	477	23.874

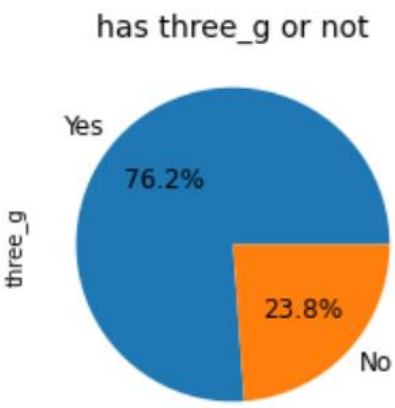


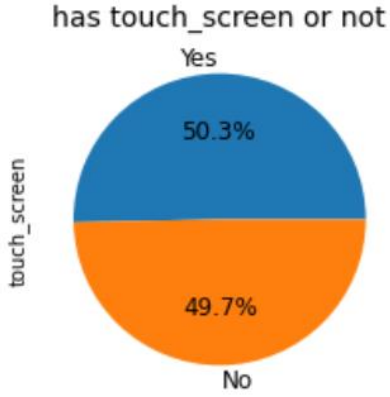
Exploratory data analysis

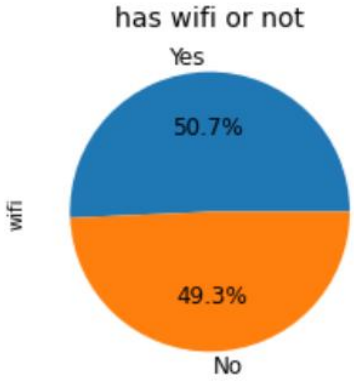






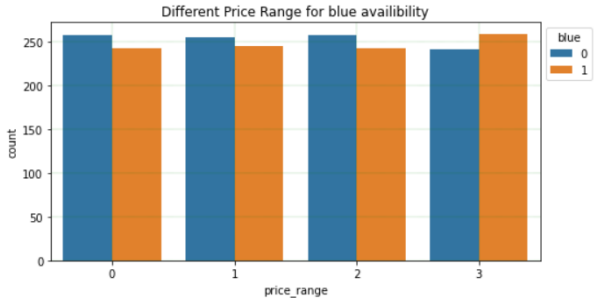


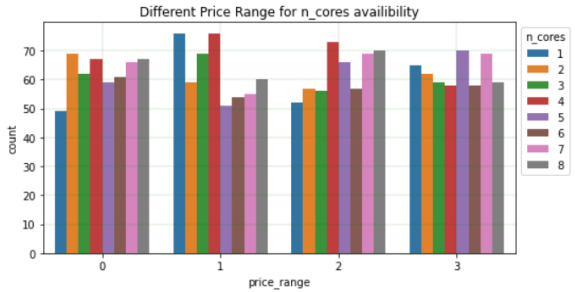


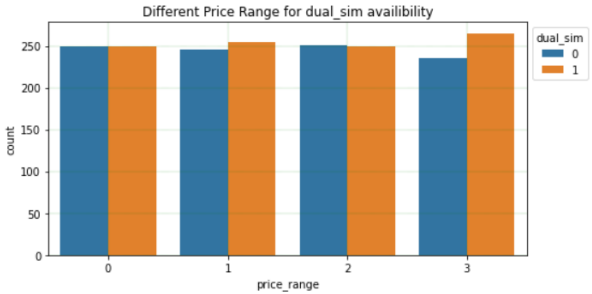


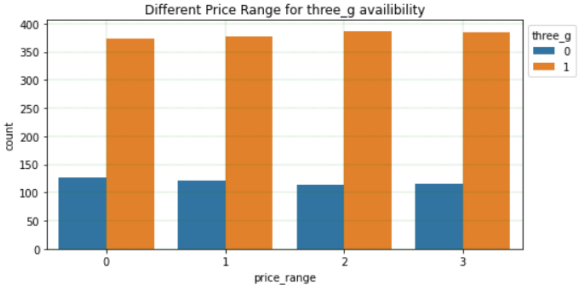
- Almost 50% of mobile given in dataset having Bluetooth, dual sim,wifi,4g network compatibility, touch screen
- 76.2% of mobile given in dataset having 3g network compatibility.

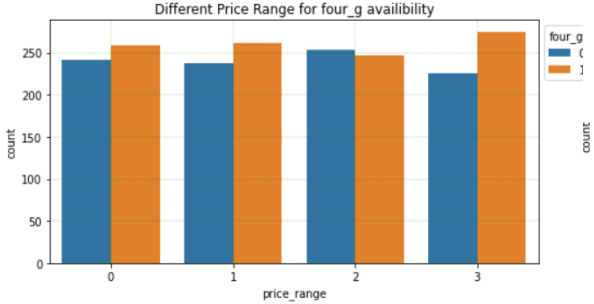
Exploratory data analysis

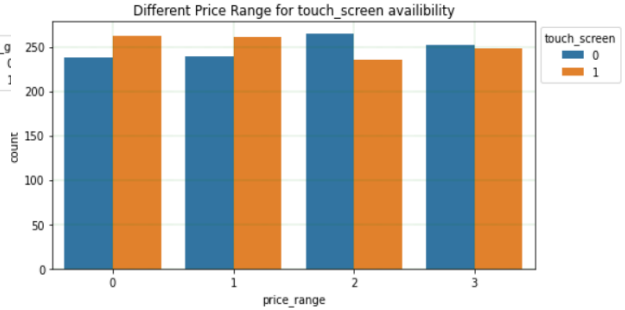






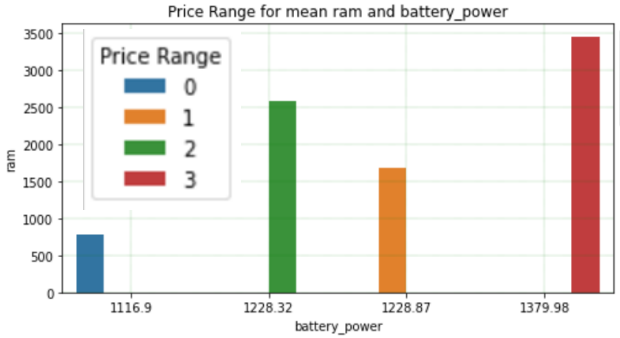


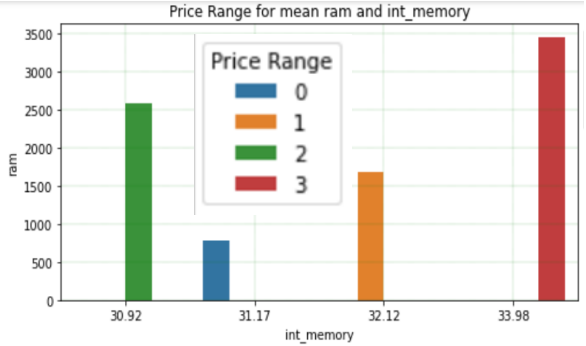


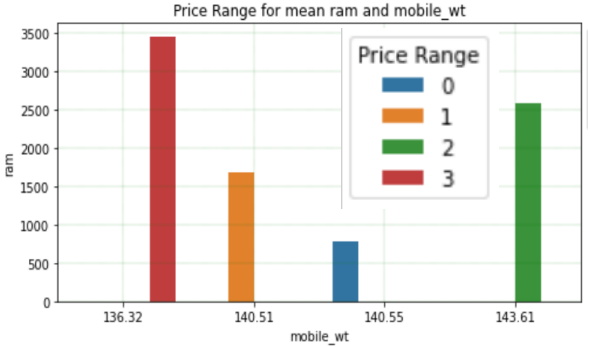


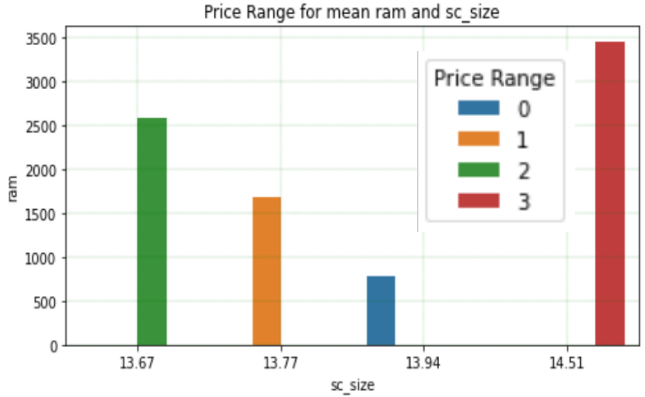
- Low and moderate price range phone having touch screen feature,most of the high price range phone have not touchscreen.
- 3g and 4g phone is more popular among each price range.
- In lower price range phone, most of the phone do not have wifi feature.
- Most of the Higher price range phone having dual sim compatibility.
- Bluetooth in high price range phone is more popular.

Exploratory Data Analysis





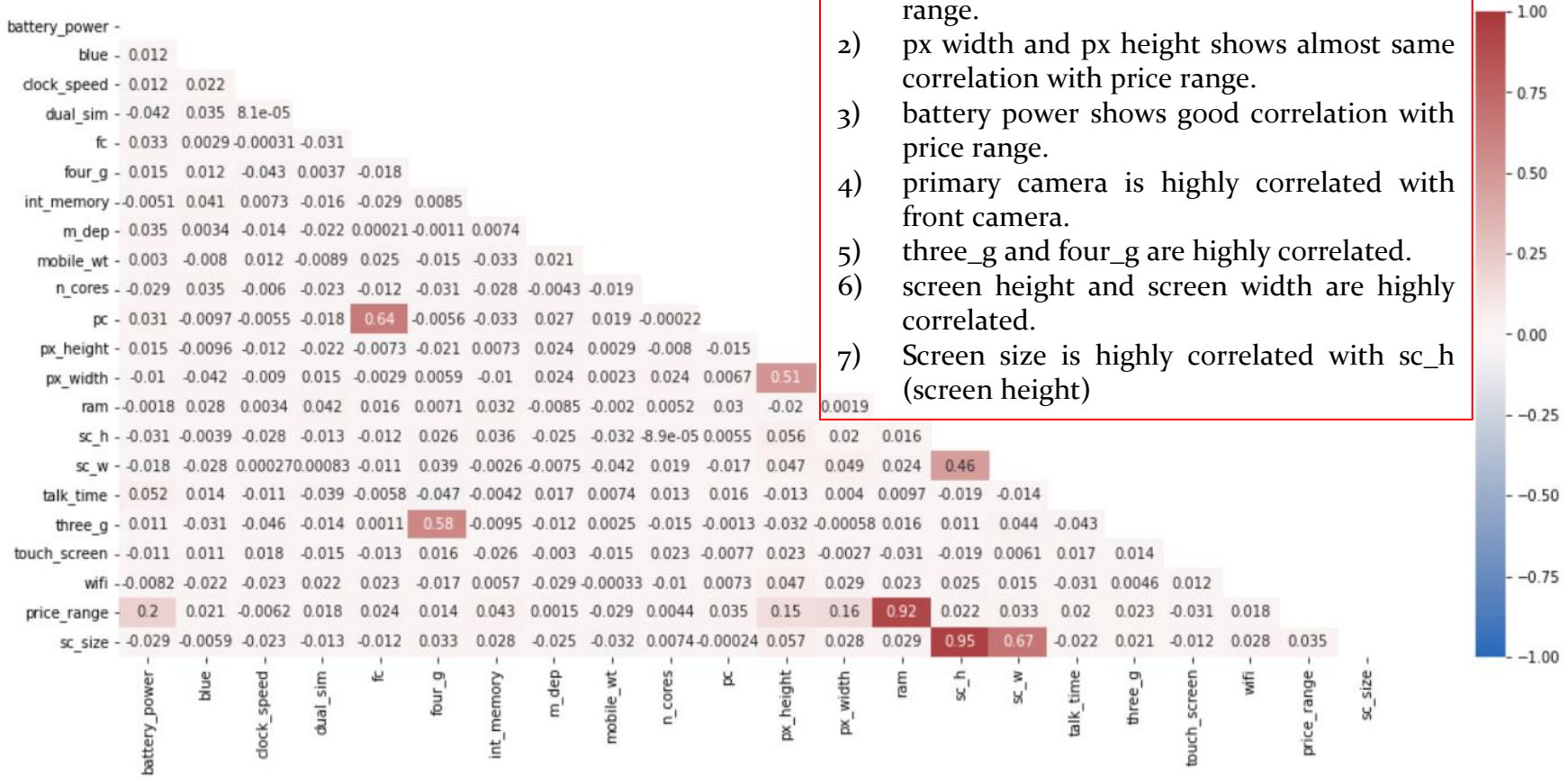




On the basis of above plot we can differentiate low price range and very high price range mobile

Low Price Range Mobile	Very High Price Mobile
1) It has average RAM size close to 750 MB. 2) It has low battery power (around 1000 mAh). 3) It has internal memory around 30GB. 4) It is moderate weight. 5) Moderate screen size.	1) It has average RAM size approximately 3500 MB. 2) It has high battery power. 3) It has internal memory OF 34GB. 4) It is lighter in weight(around 137gm). 5) Larger screen size.

Exploratory Data Analysis



1) Ram shows high correlation with price range.
2) px width and px height shows almost same correlation with price range.
3) battery power shows good correlation with price range.
4) primary camera is highly correlated with front camera.
5) three_g and four_g are highly correlated.
6) screen height and screen width are highly correlated.
7) Screen size is highly correlated with sc_h (screen height)

Data Preprocessing

1. Selecting top 15 important feature using chi2 of sklearn library.
2. Train test split of given data.
3. Scaled the data using StandardScaler of sklearn library.

Performance Parameter

1. Accuracy score of train and test data
2. Classification report of test and train data
3. ROC curve

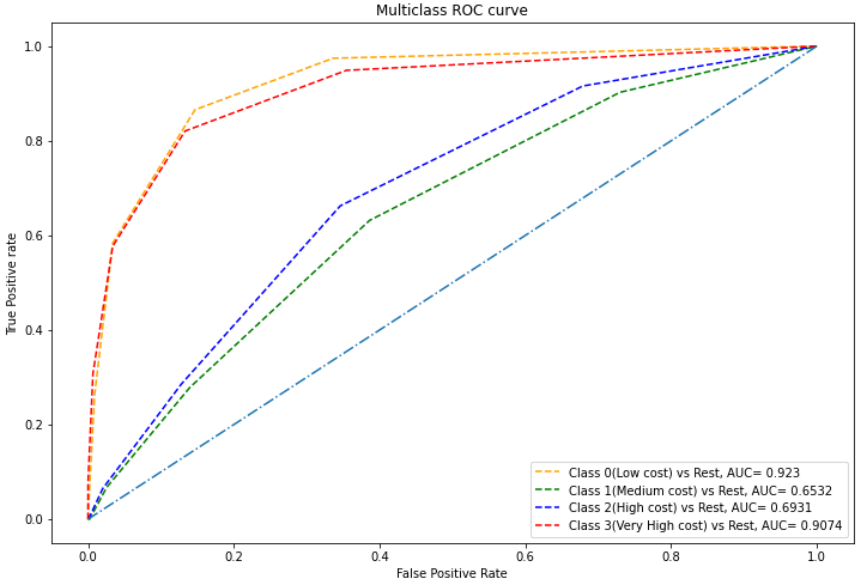
KNN classifier

X_train_scaled,
y_train
Default
parameter

→ KNN classifier →

Train accuracy: 0.7313
Test accuracy: 0.5592
Train F1 score(avg): 0.73
Test F1 score(avg): 0.56

- Training accuracy is 0.73 and test accuracy is 0.559 indicate that model is not able to given data very well. hence, we have to try some other complex algorithm.
- Training accuracy is more than test accuracy, difference between them is very large, hence it is a case of overfitting.
- From ROC curve, medium and high cost phone perform very poor with this data, as AUC close to 0.5.
- Average F1-score for test data 0.56, which is also perform poor with this given data.

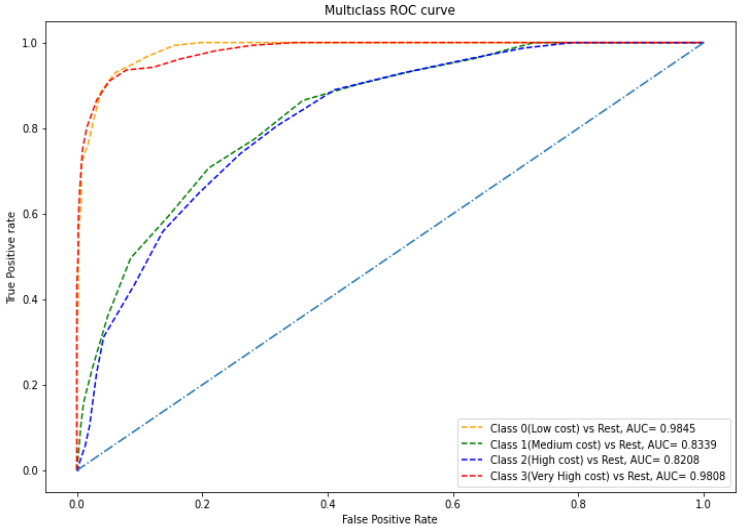


KNN classifier with GridSearchCV

Model, Parameter → **GridSearchCV** → **X_train_scaled,y_train**
Optimum parameter:
'algorithm'= 'ball_tree',
'leaf_size'= 4,
'metric'= 'manhattan',
'n_neighbors'=41
→ KNN classifier

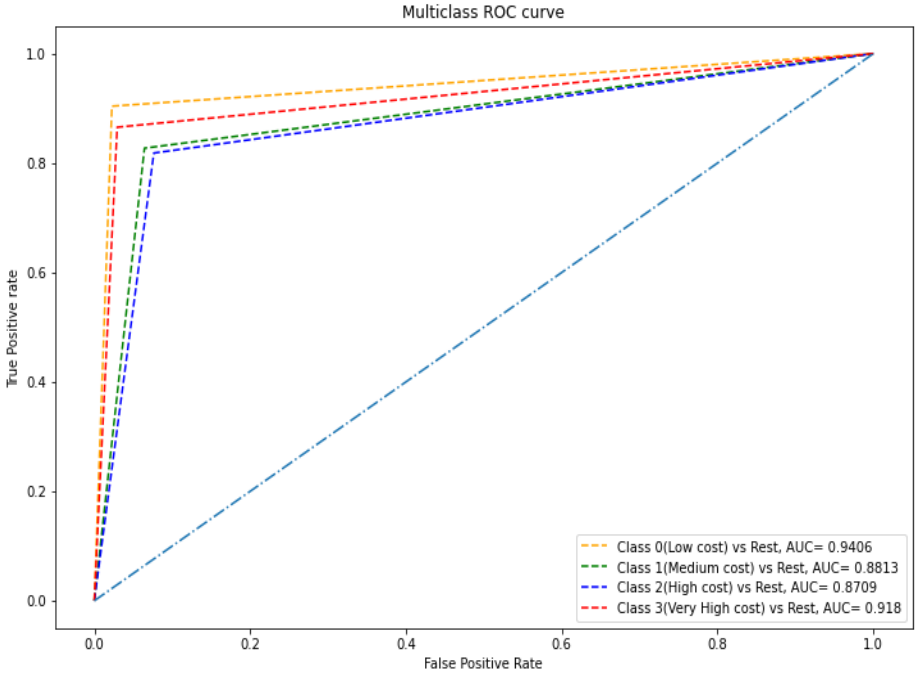
Train accuracy: 0.776
Test accuracy: 0.735
Train F1 score(avg): 0.78
Test F1 score(avg): 0.73

- Training accuracy is 0.77 and test accuracy is 0.73 indicate that model is not able to learn the given data very well.
- Training accuracy is almost equal test accuracy, difference between them is very small, hence model is generalized well but we have have to use some other complex model to improve the accuracy.
- From ROC curve, medium and high cost phone perform very poor with this data.
- Average F1-score for test data 0.73, which is also perform poor with this given data.

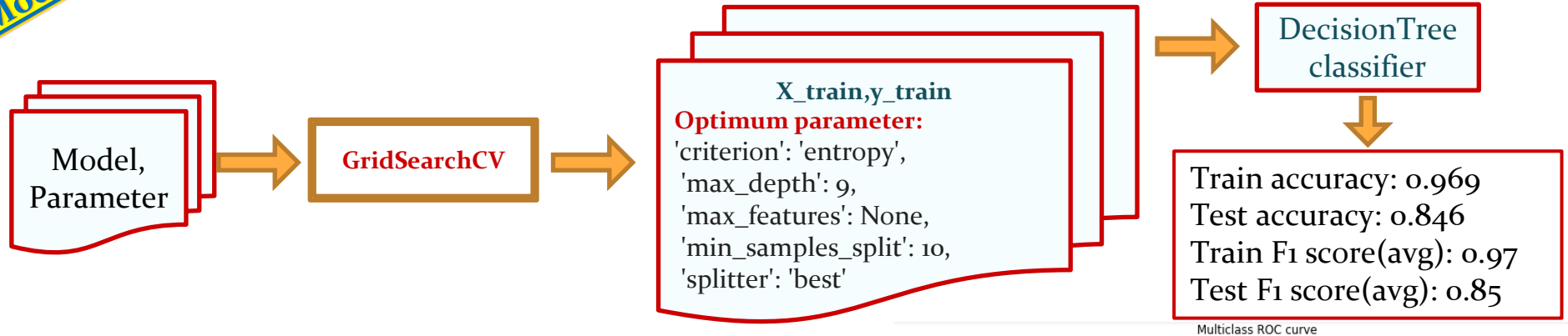


Decision Tree

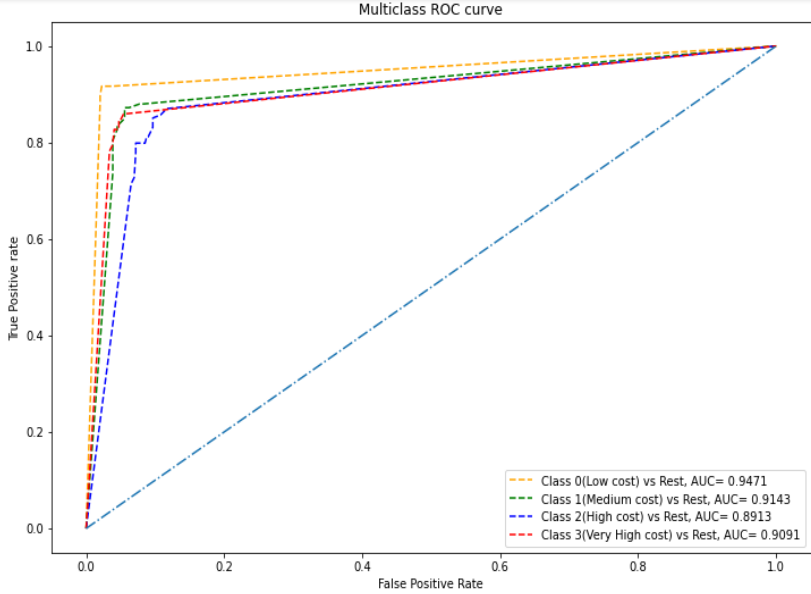
- Training accuracy is 1 and test accuracy is 0.85 indicate that model is train with the given data very well.
- Training accuracy is more than test accuracy, hence model is overfitted with the given data, we have have to use some other complex model to improve the accuracy.
- From ROC curve, medium and high cost phone perform very poor with this data.
- Average F1-score for test data 0.85, which is also perform poor with this given data.



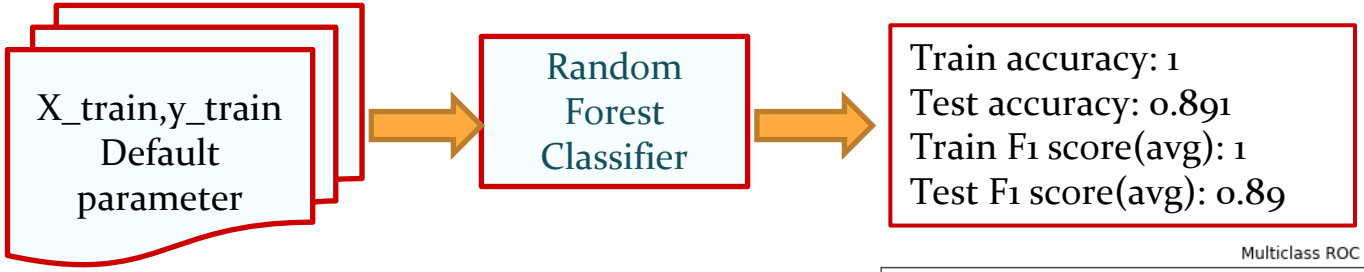
Decision Tree with GridSearchCV



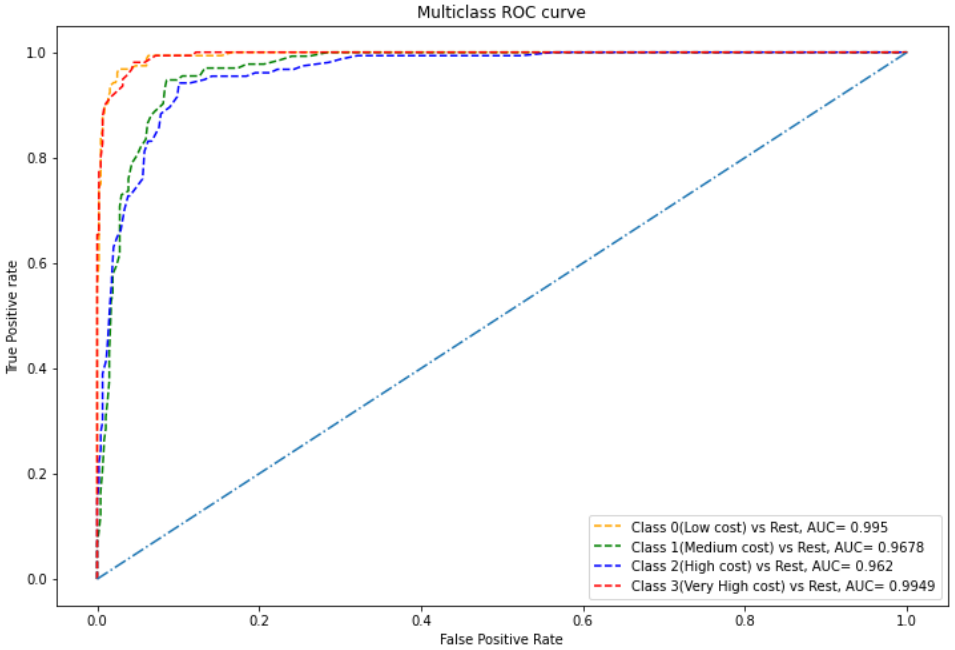
- Training accuracy is 0.97 and test accuracy is 0.84 indicate that model is train with the given data very well.
- Training accuracy is more than test accuracy, difference between them is large, hence model is train with some percentage of overfitting but we have to use some other complex model to reduce this overfitting.
- From ROC curve, low cost phone have high area under curve as compare other three phone.
- Average F1-score for test data 0.85, which is also perform poor with this given data.



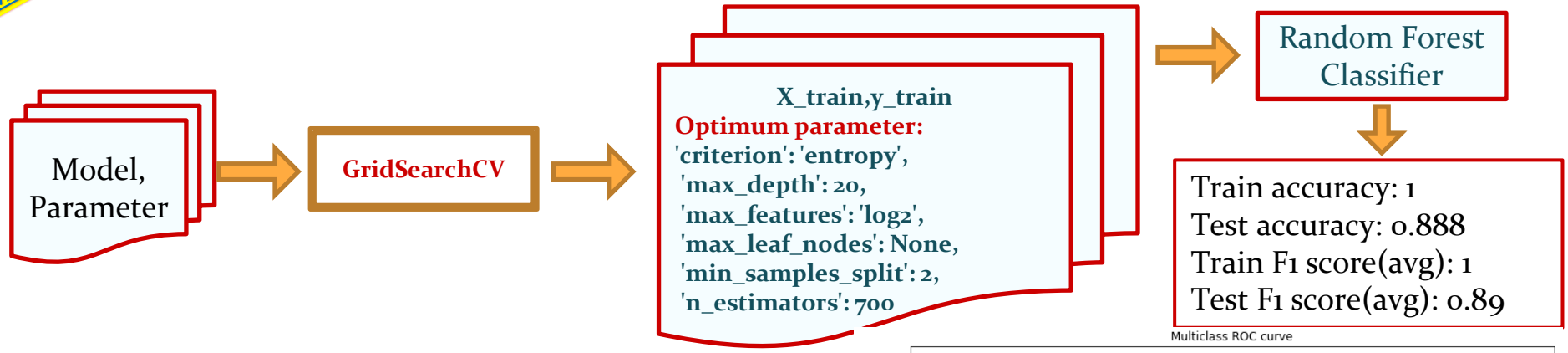
Random Forest Classifier



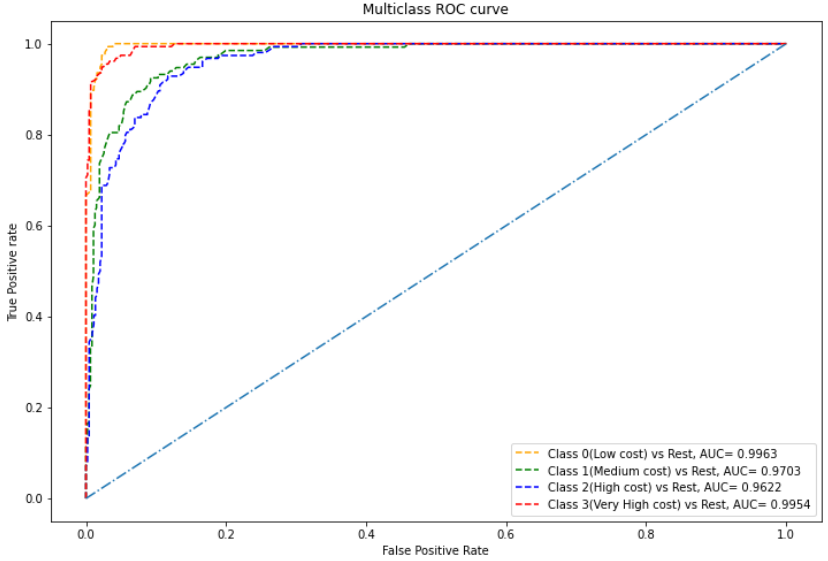
- Training accuracy is 1 and test accuracy is 0.8898 indicate that model is train with the given data very well.
- Training accuracy is more than test accuracy, difference between them is large, hence model is train with some percentage of overfitting but we have to use some other complex model to reduce this overfitting.
- From ROC curve, low cost and very high cost phone have high area under curve as compare other two phone.
- Average F1-score for test data is equal to 0.89.



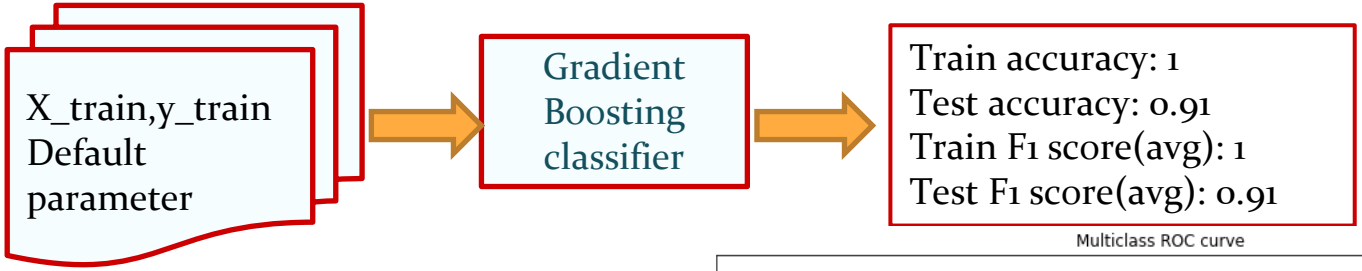
RandomForestClassifier with GridSearchCV



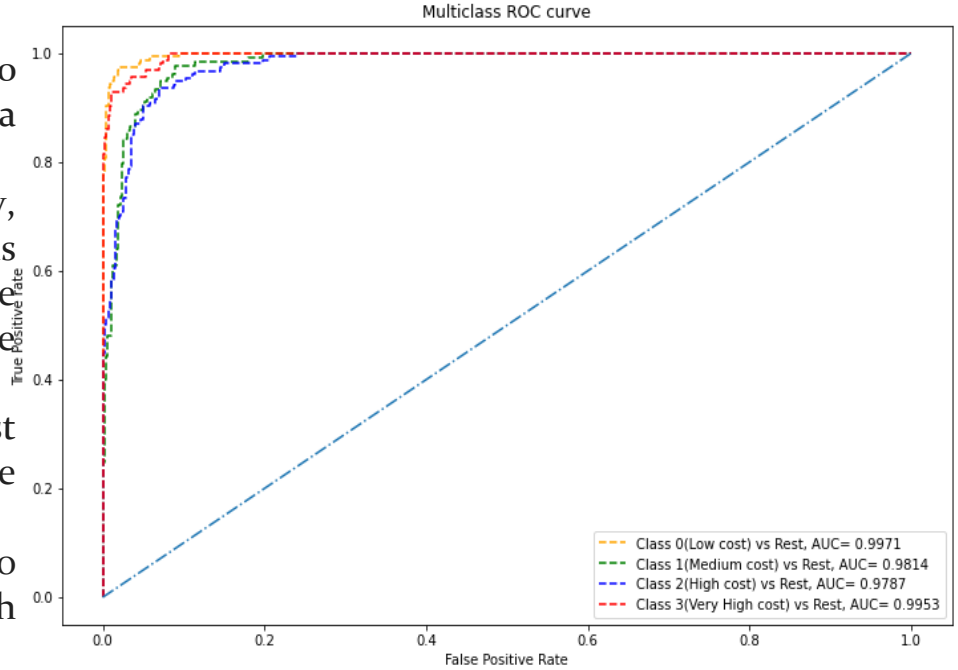
- Training accuracy is 1 and test accuracy is 0.89 indicate that model is train with the given data very well.
- Training accuracy is more than test accuracy, difference between them is large, hence model is train with some percentage of overfitting but we have to use some other complex model to reduce this overfitting.
- From ROC curve, low cost and very high cost phone have high area under curve as compare other two phone.
- Average F1-score for test data 0.90, which is also needs improvement. Hence, we have to play with some other model.



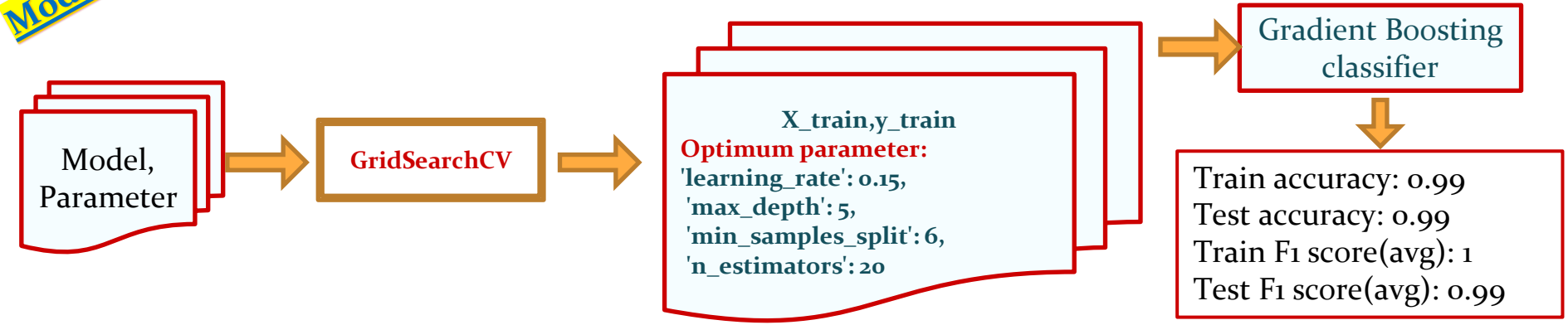
Gradient Boosting classifier



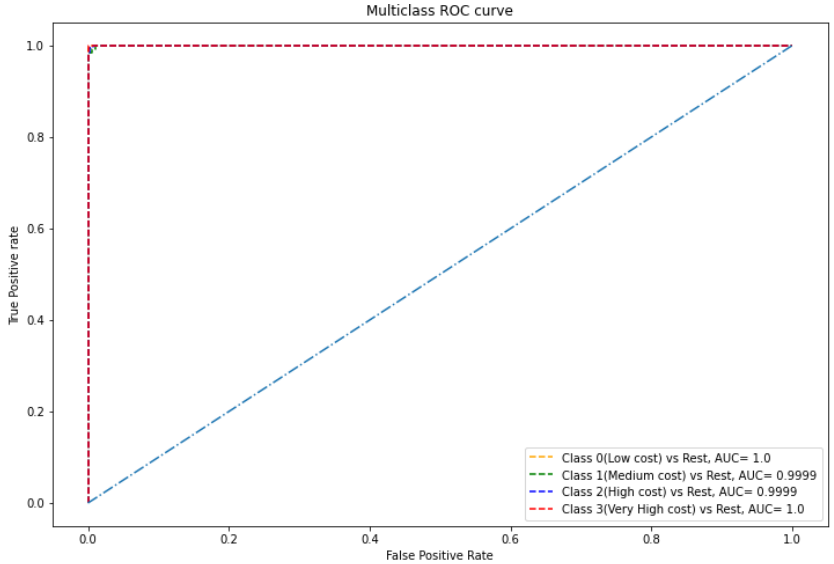
- Training accuracy is 1 and test accuracy is 0.90 indicate that model is train with the given data very well.
- Training accuracy is more than test accuracy, difference between them is large, hence model is train with some percentage of overfitting but we have to use some other complex model to reduce this overfitting.
- From ROC curve, low cost and very high cost phone have high area under curve as compare other two phone.
- Average F1-score for test data 0.91, which is also needs improvement. Hence, we have to play with some other model.



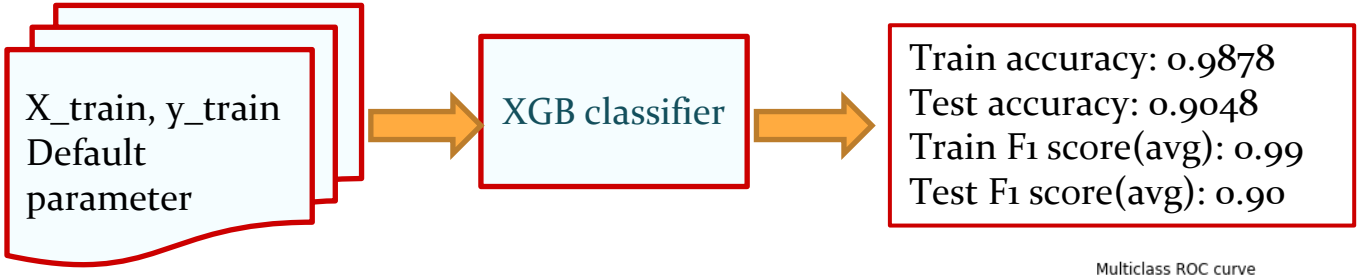
Gradient boost with GridSearchCV



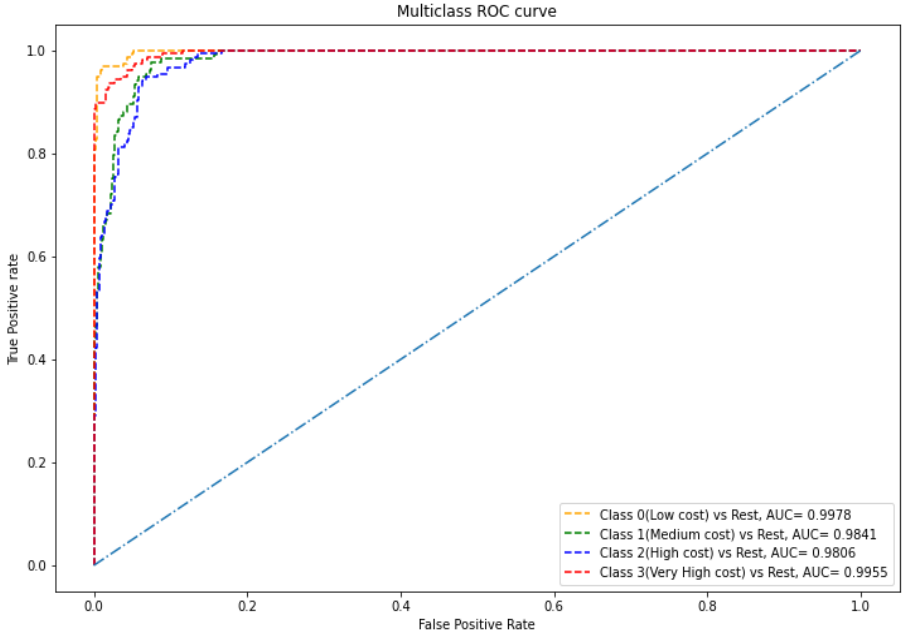
- Training accuracy is 0.99 and test accuracy is 0.99 indicate that model is train with the given data very well.
- From ROC curve, almost all class of phone overlap each other and it is closer to 1.
- Average F1-score for test data 0.99. we have to play with some other model.



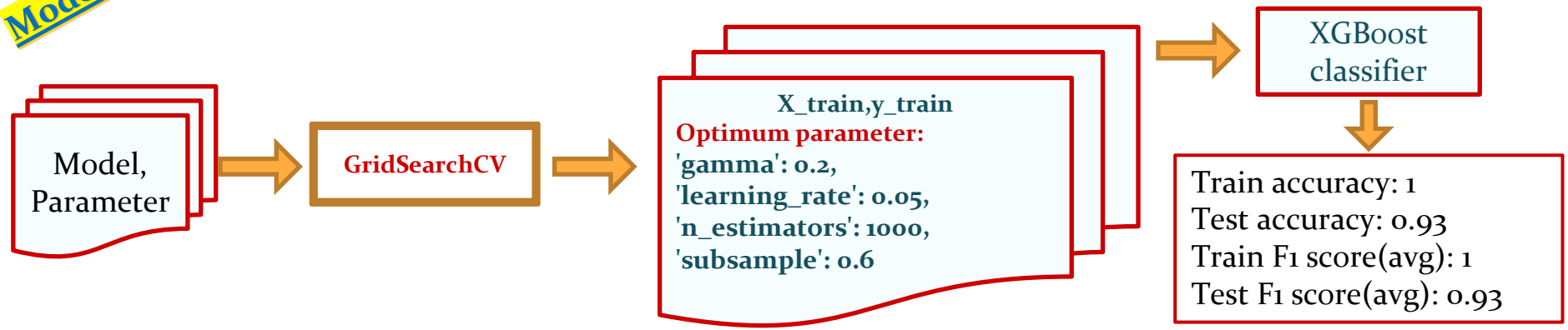
XGBClassifier



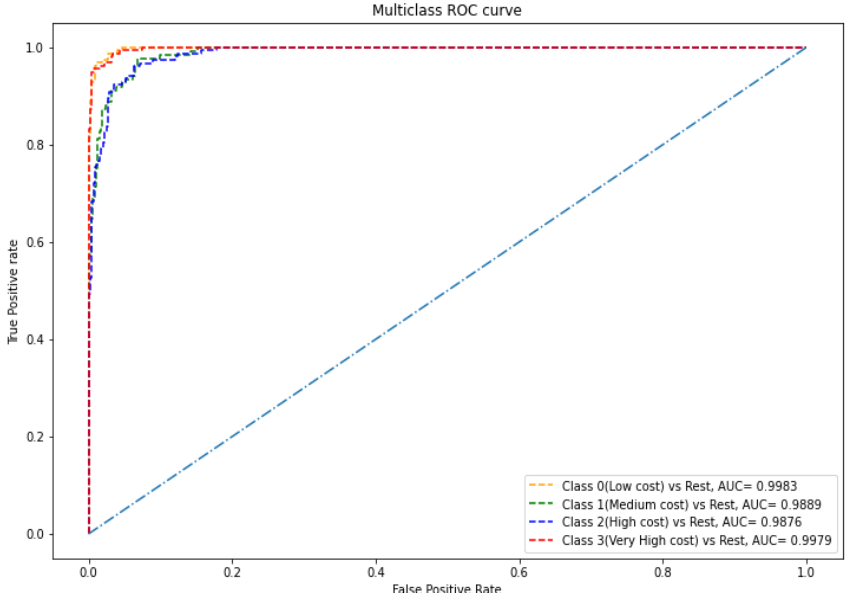
- Training accuracy is 0.98 and test accuracy is 0.90 indicate that model is train with the given data very well.
- Training accuracy is more than test accuracy, difference between them is large, hence model is train with some percentage of overfitting but we have have to search other algorithm to reduce this overfitting.
- From ROC curve, low cost and very high cost phone have high area under curve as compare other two phone.
- Average F1-score for test data 0.90, which is more than earlier model.



XGBoost with GridSearchCV



- Training accuracy is 1 and test accuracy is 0.93 indicate that model is train with the given data very well.
- Training accuracy is more than test accuracy, difference between them is large, hence model is train with some percentage of overfitting but we have search other algorithm to reduce this overfitting.
- From ROC curve, low cost and very high cost phone have high area under curve as compare other two phone.
- Average F1-score for test data 0.93, which is more than earlier model.



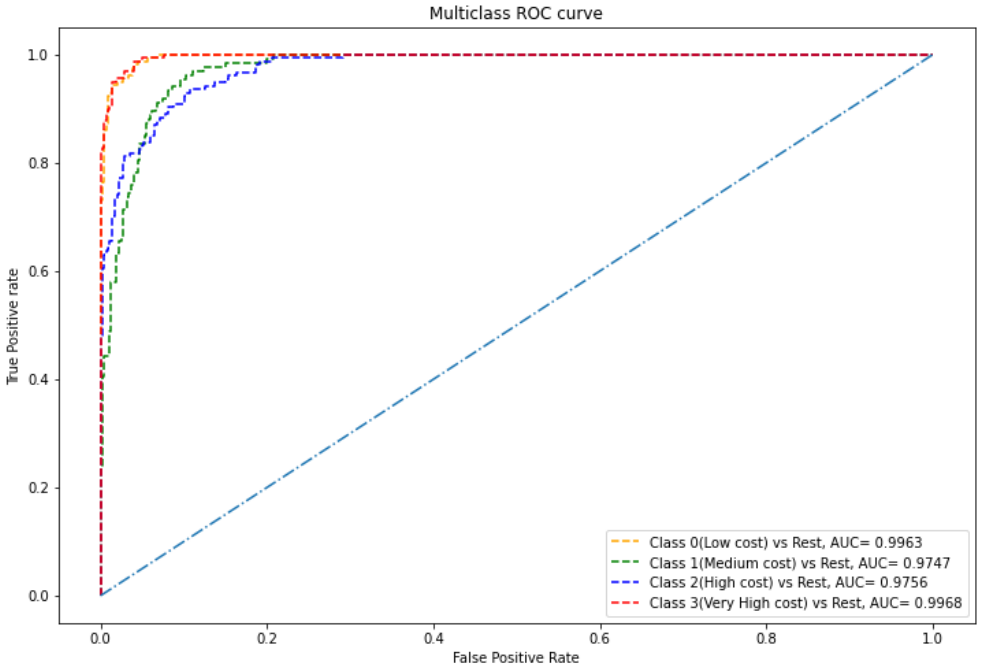
Support Vector Machine Classifier

X_train_scaled,
y_train
Default
parameter

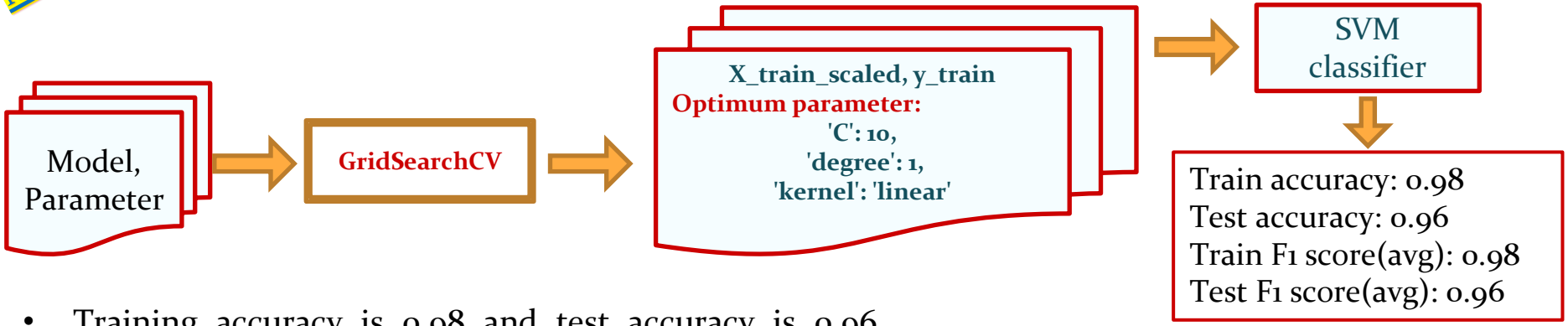
→ SVM classifier →

Train accuracy: 0.984
Test accuracy: 0.888
Train F1 score(avg): 0.98
Test F1 score(avg): 0.89

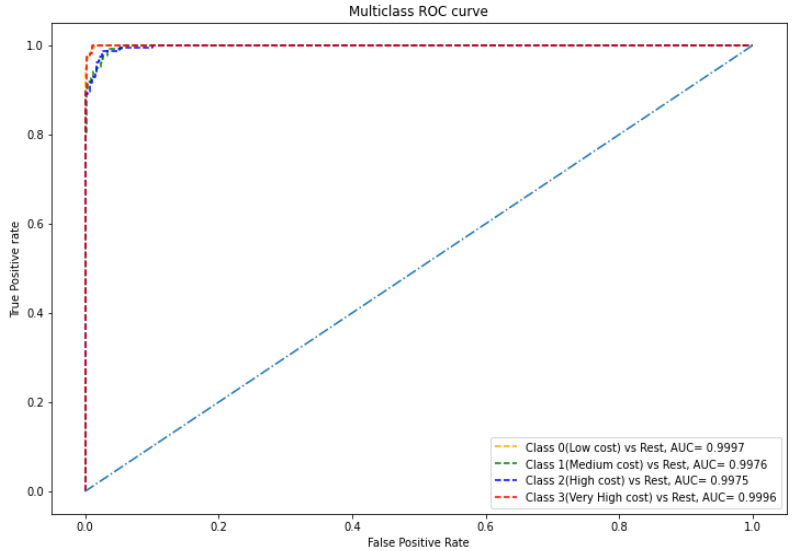
- Training accuracy is 0.98 and test accuracy is 0.88 indicate that model is train with the given data very well.
- Training accuracy is more than test accuracy, difference between them is large, hence model is train with some percentage of overfitting but we have to search optimum parameter using hyperparameter tuning to reduce this overfitting.
- From ROC curve, low cost and very high cost phone have high area under curve as compare other two phone.
- Average F1-score for test data 0.89, which is more than earlier model.



Support Vector Machine with GridSearchCV



- Training accuracy is 0.98 and test accuracy is 0.96 indicate that model is train with the given data very well.
- Training accuracy is almost equal to test accuracy, difference between them is so small that it can be neglected, hence model is generalized very well with the given data.
- From ROC curve, low cost and very high cost phone have high area under curve as compare other two phone.
- Average F1-score for test data 0.96, which is also good.



Conclusions:

- KNN with or without GridSearchCV show poor accuracy score as compare to other model.
- SVC with GridSearchCV having train accuracy score is 0.984 and test accuracy score is 0.962 which is good performer.
- GradientBoostingClassifier with GridSearchCV have test accuracy score is 0.992 and train accuracy score is 0.999, which is best among all the model. Hence we can deploy this model.


Thank You