Breast Cancer Survival Prediction

Nakisa Abbasi

Introduction

❖ Why and Who Cares?

- Breast cancer is among the most common cancers and a common cause of death among women

- The Patients want to know is how long they will survive

- Patients with different criteria can have very different survival rates.

- We will predict survival rate for patients with different diagnostic criteria.

Data Information

- Data is acquired from <u>Seer Breast Cancer Data</u>
- Was obtained from the 2017 November update of the SEER, from cancer patients
- From female patients with infiltrating duct and lobular carcinoma breast cancer diagnosed in 2006-2010.

Data Information

Each record is consist of:

- **age, race, marital status, t stage, n stage,a stage, tumor size, estrogen status, progesterone status,regional nodes examined, regional nodes positive and status of the patient.**

Data wrangling

Data cleaning?

This dataset is not so dirty, no Null records. Some text cleaning

- Moderately differentiated; Grade II: → Grade II
- Poorly differentiated; Grade III: → Grade III
- . . .

Data wrangling

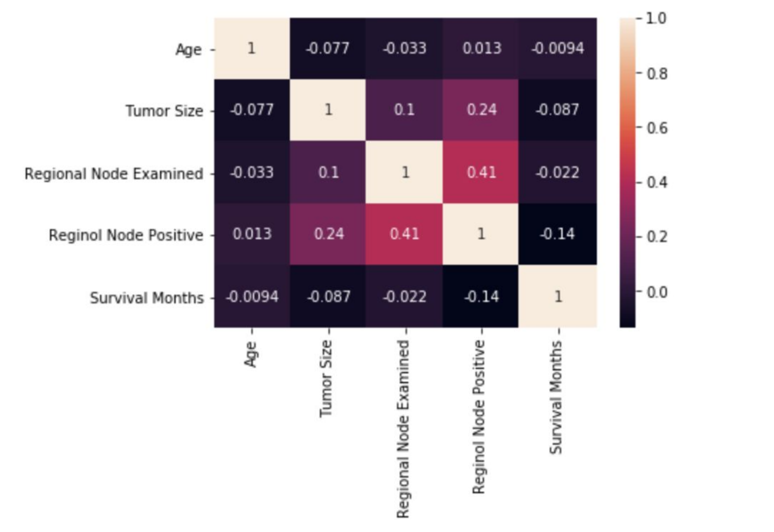
Categorical data to numeric

- **For features “one hot encoding”**
    - features = pd.get_dummies(features,drop_first = True)

- **For target “LabelEncoder”**
    - target = pd.DataFrame(le.fit_transform(target),columns = ['Status'])

Data wrangling

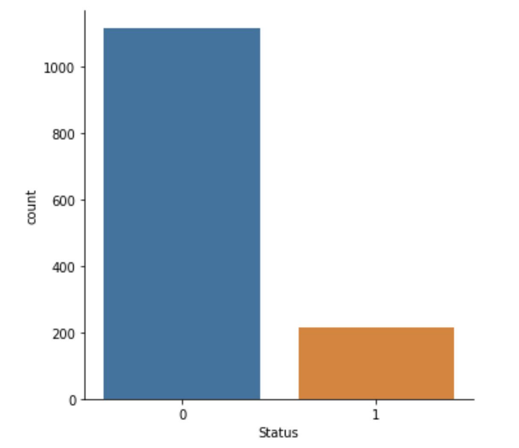
Any correlation?



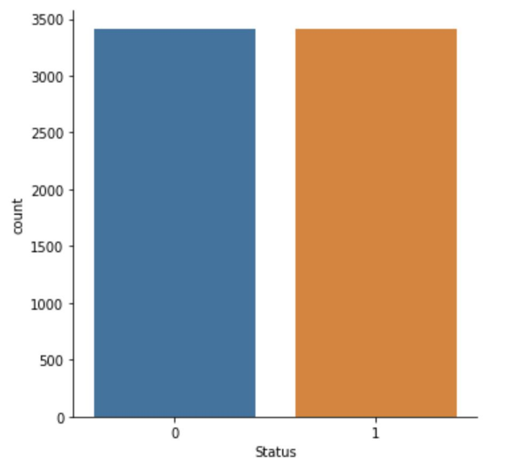
	Age	Tumor Size	Regional Node Examined	Reginol Node Positive	Survival Months
Age	1	-0.077	-0.033	0.013	-0.0094
Tumor Size	-0.077	1	0.1	0.24	-0.087
Regional Node Examined	-0.033	0.1	1	0.41	-0.022
Reginol Node Positive	0.013	0.24	0.41	1	-0.14
Survival Months	-0.0094	-0.087	-0.022	-0.14	1

Data wrangling

Is the target data balanced?



Apply SMOTE

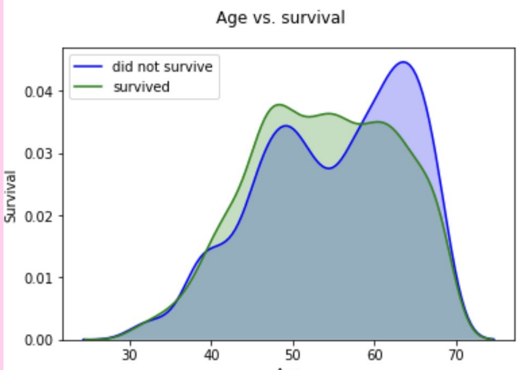


EDA

Continuous data:

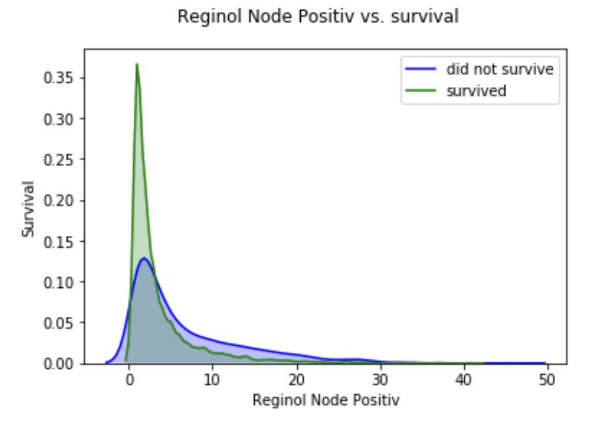
1. Age VS Survival

There is and increase rate in death numbers by age increase



2. Regional Node Positive VS Survival

Most of the survived patients have less number of positive nodes

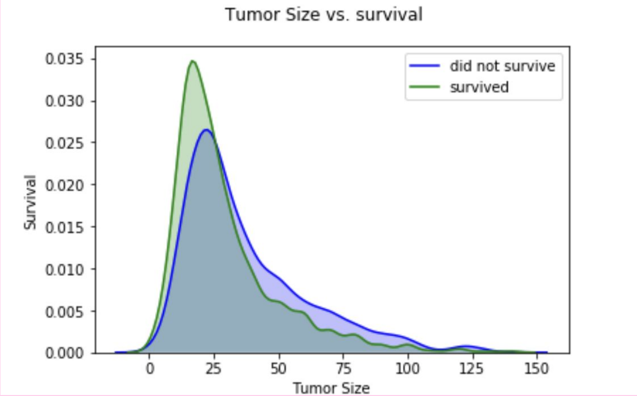


EDA

Continuous data continued:

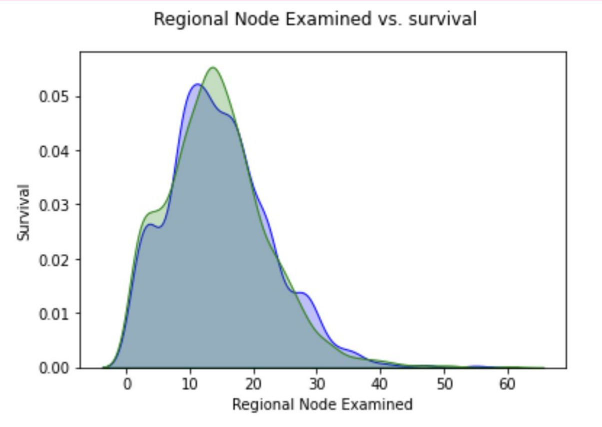
3. Tumor Size VS Survival

Most of the survived patients have smaller tumor size



4. Regional Node Examined VS Survival

The number of nodes being examined seems to be equal in most of the patient

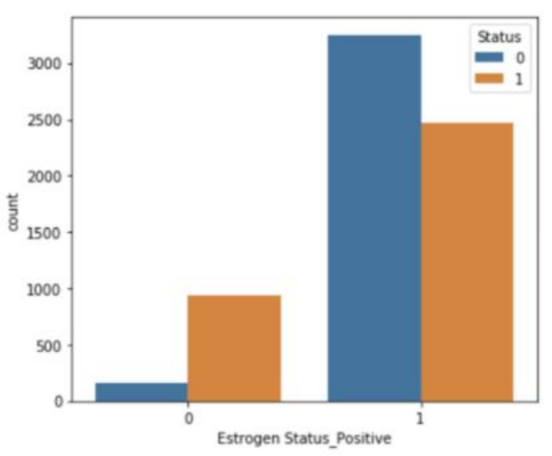


EDA

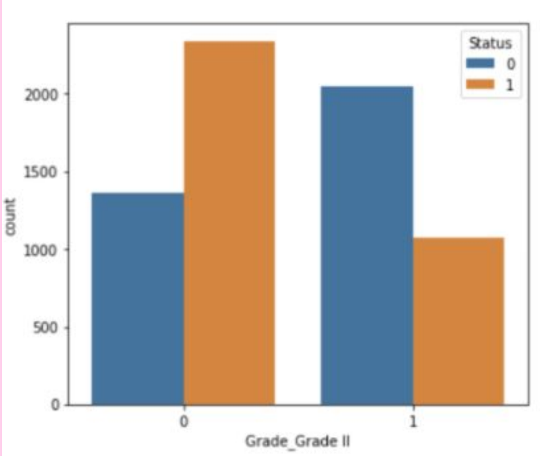
Categorical data continued:

Most important ones:

Apparently the number of survived patients is higher with Estrogen positive



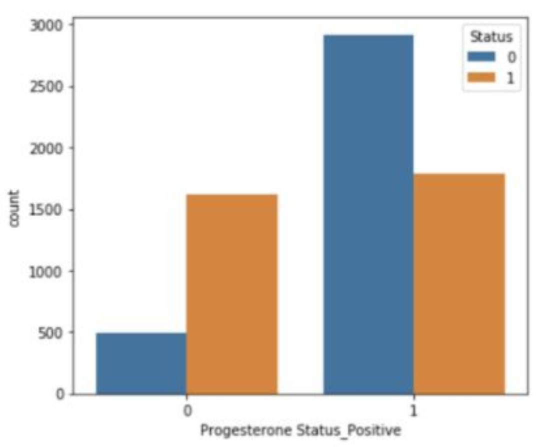
The survival rate is higher with Grade 2



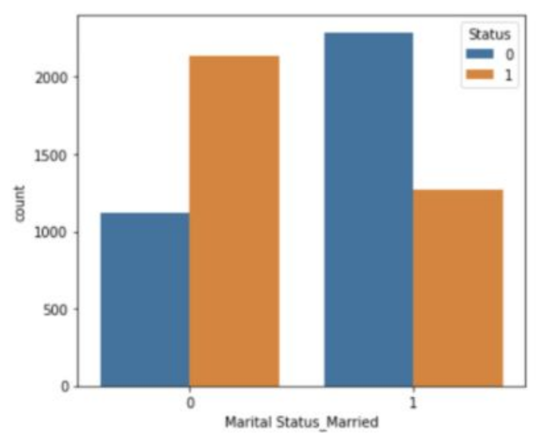
EDA

Categorical data continued:

The number of survived patients is higher with progesterone positive



Married patients have higher chance to survive

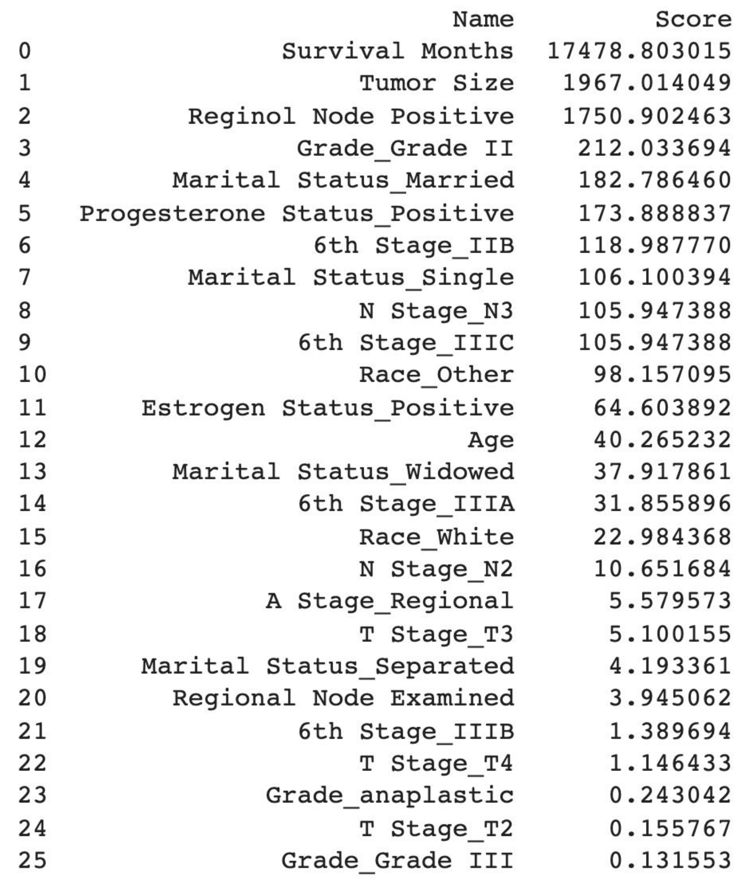


Preprocessing

- **Which Features are more important?**

  Chi-squared function was used with SelectKBest from Sklearn to calculate the highest scores:

1. Survival Months
2. Tumor Size
3. Regional Node Positive
4. Grade_Grade II
5. Progesterone Status_Positive
6. 6th Stage_IIIC
7. Marital Status_Married
8. Marital Status_Single
9. Race_Other
10. Estrogen Status_Positive



```
                             Name         Score
0                 Survival Months  17478.803015
1                      Tumor Size   1967.014049
2          Reginol Node Positive    1750.902463
3                  Grade_Grade II    212.033694
4          Marital Status_Married    182.786460
5    Progesterone Status_Positive    173.888837
6                  6th Stage_IIB     118.987770
7           Marital Status_Single    106.100394
8                      N Stage_N3    105.947388
9                 6th Stage_IIIC     105.947388
10                     Race_Other     98.157095
11       Estrogen Status_Positive     64.603892
12                            Age     40.265232
13         Marital Status_Widowed     37.917861
14                6th Stage_IIIA      31.855896
15                     Race_White     22.984368
16                     N Stage_N2     10.651684
17              A Stage_Regional       5.579573
18                     T Stage_T3      5.100155
19       Marital Status_Separated      4.193361
20        Regional Node Examined       3.945062
21                6th Stage_IIIB       1.389694
22                     T Stage_T4      1.146433
23              Grade_anaplastic       0.243042
24                     T Stage_T2      0.155767
25                Grade_Grade III      0.131553
```

Preprocessing

- **Are the data in the same range?**

NO

Used MinMaxScaler to scale our data in the range of 0-5

Feature	**Min**	**Max**
Age	30	69
Tumor size (mm)	1	140
Regional node examined	1	61
Regional node positive f	1	46
Survival Months	1	107

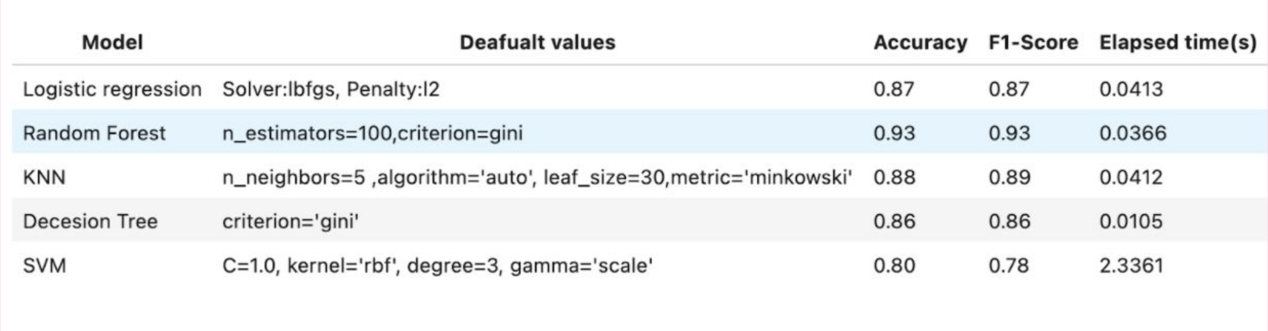
Modeling

There are 5 models that we will investigate

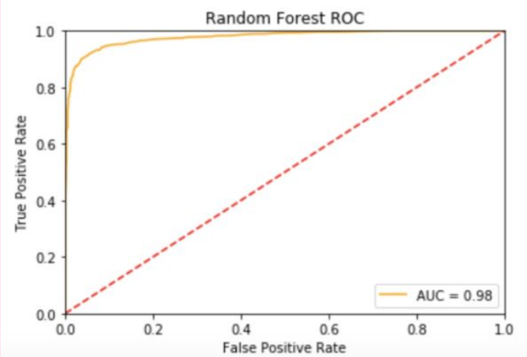
1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest
4. KNeighborsClassifier
5. Support Vector Machine

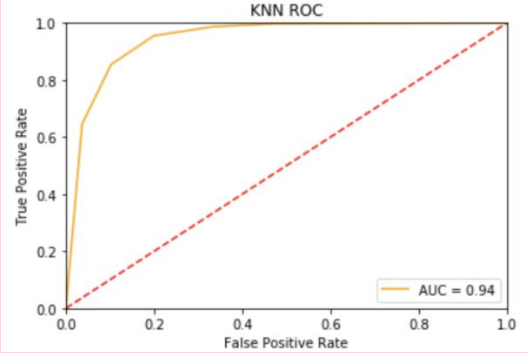
Modeling

- Default Model results for **all available data**



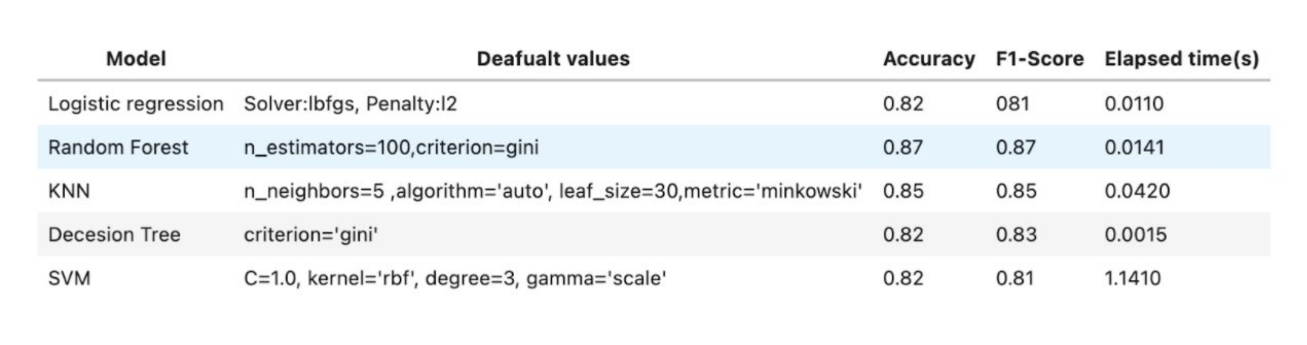
Model	Deafualt values	Accuracy	F1-Score	Elapsed time(s)
Logistic regression	Solver:lbfgs, Penalty:l2	0.87	0.87	0.0413
Random Forest	n_estimators=100,criterion=gini	0.93	0.93	0.0366
KNN	n_neighbors=5 ,algorithm='auto', leaf_size=30,metric='minkowski'	0.88	0.89	0.0412
Decesion Tree	criterion='gini'	0.86	0.86	0.0105
SVM	C=1.0, kernel='rbf', degree=3, gamma='scale'	0.80	0.78	2.3361



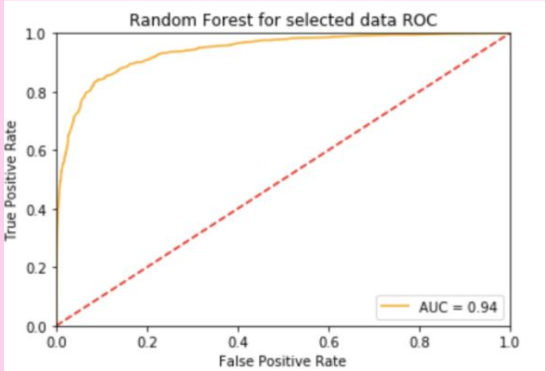


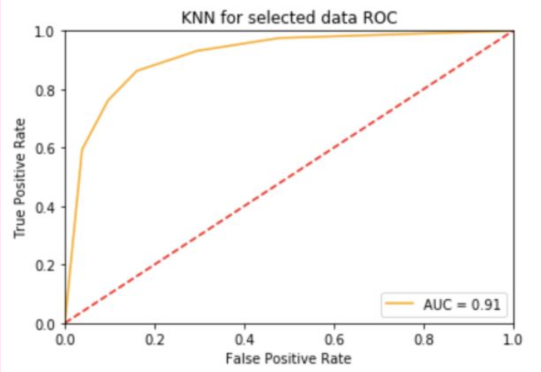
Modeling

- Default Model results for **selected data**



Model	Deafualt values	Accuracy	F1-Score	Elapsed time(s)
Logistic regression	Solver:lbfgs, Penalty:l2	0.82	081	0.0110
Random Forest	n_estimators=100,criterion=gini	0.87	0.87	0.0141
KNN	n_neighbors=5 ,algorithm='auto', leaf_size=30,metric='minkowski'	0.85	0.85	0.0420
Decesion Tree	criterion='gini'	0.82	0.83	0.0015
SVM	C=1.0, kernel='rbf', degree=3, gamma='scale'	0.82	0.81	1.1410



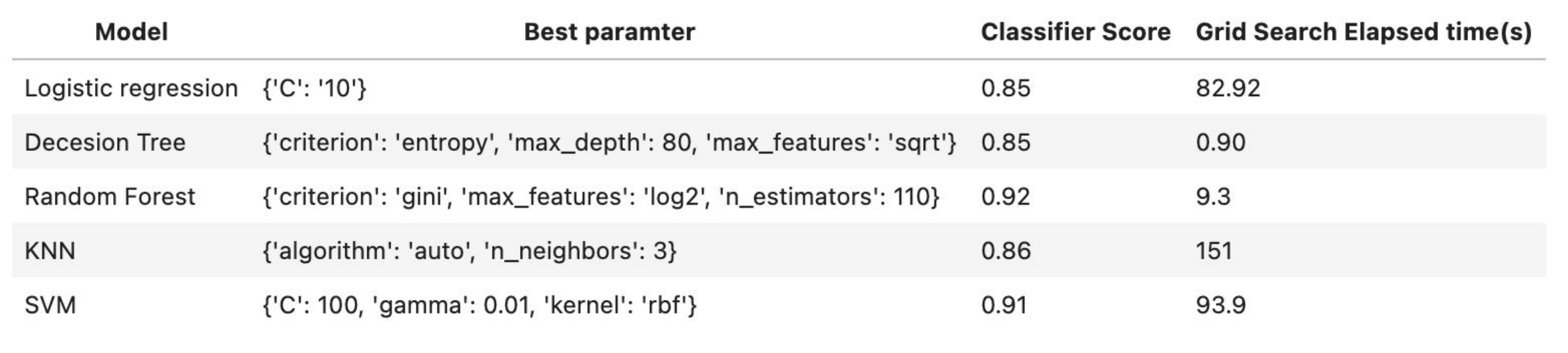


Modeling

- **Default Model conclusion:**
  - All models have very close Accuracy and F1-score when using all the data
  - Elapsed time is much higher in the SVM model
  - Accuracy declined in all models while selected data used
    - ('Grade_Grade II', 'Progesterone Status_Positive', '6th Stage_IIIC', 'Marital Status_Married', 'Estrogen Status_Positive')
  - The elapsed time for running and predicting has decreased a lot.
  - In conclusion, as the data is not so huge and the elapsed time does not seem to be a problem for the amount of the data we have, and the accuracy is better with all the data, we will use all the data for deeper analysis.

Modeling

❖ **Models with hyper parameter tuning**

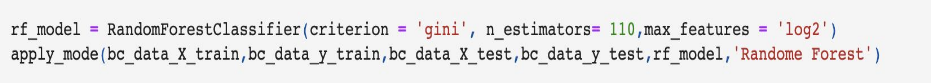


Model	Best paramter	Classifier Score	Grid Search Elapsed time(s)
Logistic regression	{'C': '10'}	0.85	82.92
Decesion Tree	{'criterion': 'entropy', 'max_depth': 80, 'max_features': 'sqrt'}	0.85	0.90
Random Forest	{'criterion': 'gini', 'max_features': 'log2', 'n_estimators': 110}	0.92	9.3
KNN	{'algorithm': 'auto', 'n_neighbors': 3}	0.86	151
SVM	{'C': 100, 'gamma': 0.01, 'kernel': 'rbf'}	0.91	93.9

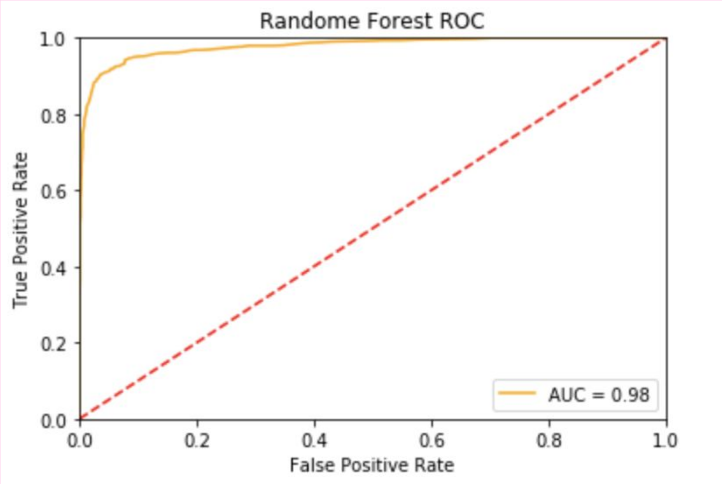
According to the different models' score result, **Random Forest** and **SVM** have the highest score equal to 92% AND 91%. Let's compare the performance of these models with the selected parameter.

Modeling

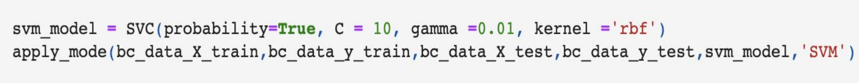
- **Selected Model 1: RandomForestClassifier:**



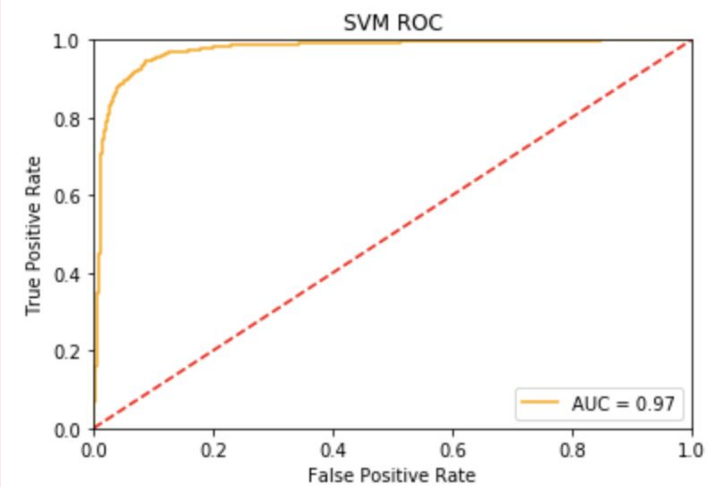
```
rf_model = RandomForestClassifier(criterion = 'gini', n_estimators= 110,max_features = 'log2')
apply_mode(bc_data_X_train,bc_data_y_train,bc_data_X_test,bc_data_y_test,rf_model,'Randome Forest')
```



- **Selected Model 2: SupportVectorMachineClassifier**

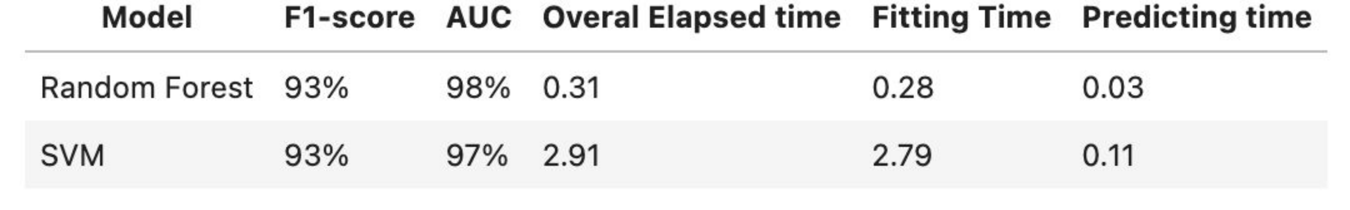


```
svm_model = SVC(probability=True, C = 10, gamma =0.01, kernel ='rbf')
apply_mode(bc_data_X_train,bc_data_y_train,bc_data_X_test,bc_data_y_test,svm_model,'SVM')
```



Modeling

Final Conclusion



Model	F1-score	AUC	Overal Elapsed time	Fitting Time	Predicting time
Random Forest	93%	98%	0.31	0.28	0.03
SVM	93%	97%	2.91	2.79	0.11

- RF slightly better AUC (98%) compared to SVM (97%)
- They both have an equal (93%) score
- Decision Tree took much much less time to fit (0.2877) compared to SVM fitting time (2.79)
- Less time to predict the test data (0.03) compared to the SVM (0.11 s).
- For our purpose, which is to predict the survival rate of cancer patients, we select **Random Forest** as our winning model.