In [1]:
```
import pandas as pd
import numpy as np
import plotly.express as px
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
df = pd.read_csv("BRCA Data - BRCA Data.csv")
df.head()
```

Out[2]:

| | Patient_ID | Age | Gender | Protein1 | Protein2 | Protein3 | Protein4 | Tumour_Stage | Histology | ER status | PR status | HER2 status | Surge |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | TCGA-D8-A1XD | 36 | FEMALE | 0.080353 | 0.42638 | 0.54715 | 0.273680 | III | Infiltrating Ductal Carcinoma | Positive | Positive | Negative | M Mast |
| 1 | TCGA-EW-A1OX | 43 | FEMALE | -0.420320 | 0.57807 | 0.61447 | -0.031505 | II | Mucinous Carcinoma | Positive | Positive | Negative | Lump |
| 2 | TCGA-A8-A079 | 69 | FEMALE | 0.213980 | 1.31140 | -0.32747 | -0.234260 | III | Infiltrating Ductal Carcinoma | Positive | Positive | Negative | |
| 3 | TCGA-D8-A1XR | 56 | FEMALE | 0.345090 | -0.21147 | -0.19304 | 0.124270 | II | Infiltrating Ductal Carcinoma | Positive | Positive | Negative | M Mast |
| 4 | TCGA-BH-A0BF | 56 | FEMALE | 0.221550 | 1.90680 | 0.52045 | -0.311990 | II | Infiltrating Ductal Carcinoma | Positive | Positive | Negative | |

In [3]:
```
df.isnull().sum()
```

Out[3]:
```
Patient_ID            0
Age                   0
Gender                0
Protein1              0
Protein2              0
Protein3              0
Protein4              0
Tumour_Stage          0
Histology             0
ER status             0
PR status             0
HER2 status           0
Surgery_type          0
Date_of_Surgery       0
Date_of_Last_Visit    17
Patient_Status        13
dtype: int64
```

In [4]:
```
df.isnull().sum().sum()
```

Out[4]:
```
30
```

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 334 entries, 0 to 333
Data columns (total 16 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Patient_ID          334 non-null    object 
 1   Age                 334 non-null    int64  
 2   Gender              334 non-null    object 
 3   Protein1            334 non-null    float64
 4   Protein2            334 non-null    float64
 5   Protein3            334 non-null    float64
 6   Protein4            334 non-null    float64
 7   Tumour_Stage        334 non-null    object 
 8   Histology           334 non-null    object 
 9   ER status           334 non-null    object 
 10  PR status           334 non-null    object 
 11  HER2 status         334 non-null    object 
 12  Surgery_type        334 non-null    object 
 13  Date_of_Surgery     334 non-null    object 
 14  Date_of_Last_Visit  317 non-null    object 
 15  Patient_Status      321 non-null    object 
dtypes: float64(4), int64(1), object(11)
memory usage: 41.9+ KB
```

In [6]:
```
df.describe()
```

Out[6]:

| | Age | Protein1 | Protein2 | Protein3 | Protein4 |
|---|---|---|---|---|---|
| **count** | 334.000000 | 334.000000 | 334.000000 | 334.000000 | 334.000000 |
| **mean** | 58.886228 | -0.029991 | 0.946896 | -0.090204 | 0.009819 |
| **std** | 12.961212 | 0.563588 | 0.911637 | 0.585175 | 0.629055 |
| **min** | 29.000000 | -2.340900 | -0.978730 | -1.627400 | -2.025500 |
| **25%** | 49.000000 | -0.358888 | 0.362173 | -0.513748 | -0.377090 |
| **50%** | 58.000000 | 0.006129 | 0.992805 | -0.173180 | 0.041768 |
| **75%** | 68.000000 | 0.343598 | 1.627900 | 0.278353 | 0.425630 |
| **max** | 90.000000 | 1.593600 | 3.402200 | 2.193400 | 1.629900 |

In [7]:
```
from pandas_profiling import ProfileReport
```

In [8]:
```
ProfileReport(df)
```

Summarize dataset: 100% 51/51 [00:11<00:00, 2.74it/s, Completed]

Generate report structure: 100% 1/1 [00:11<00:00, 11.44s/it]

Render HTML: 100% 1/1 [00:02<00:00, 2.50s/it]

# Overview

## Dataset statistics

| | |
|---|---|
| **Number of variables** | 16 |
| **Number of observations** | 334 |
| **Missing cells** | 30 |
| **Missing cells (%)** | 0.6% |
| **Duplicate rows** | 0 |
| **Duplicate rows (%)** | 0.0% |
| **Total size in memory** | 41.9 KiB |
| **Average record size in memory** | 128.4 B |

## Variable types

| | |
|---|---|
| **Text** | 1 |
| **Numeric** | 5 |
| **Categorical** | 8 |
| **DateTime** | 2 |

## Alerts

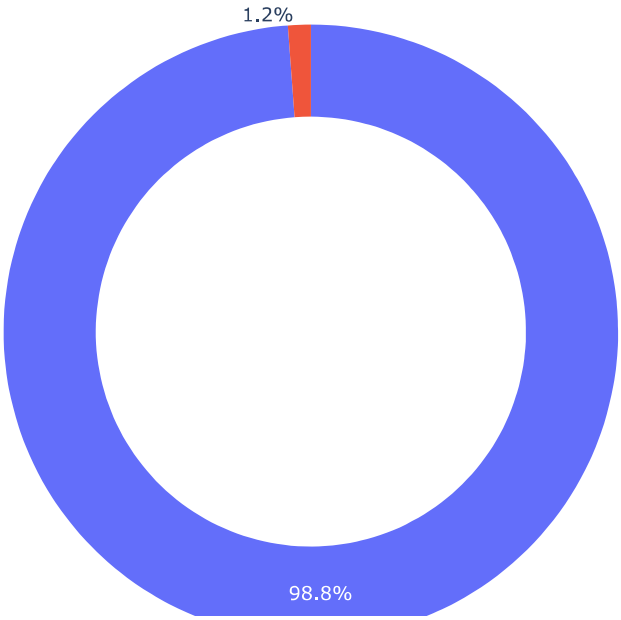
| | |
|---|---|
| `ER status` has constant value "" | Constant |
| `PR status` has constant value "" | Constant |
| `Gender` is highly imbalanced (90.6%) | Imbalance |
| `HER2 status` is highly imbalanced (57.4%) | Imbalance |
| `Date_of_Last_Visit` has 17 (5.1%) missing values | Missing |
| `Patient_Status` has 13 (3.9%) missing values | Missing |
| `Patient_ID` has unique values | Unique |

Out[8]:

In [9]:
```
df.Gender.value_counts()
```

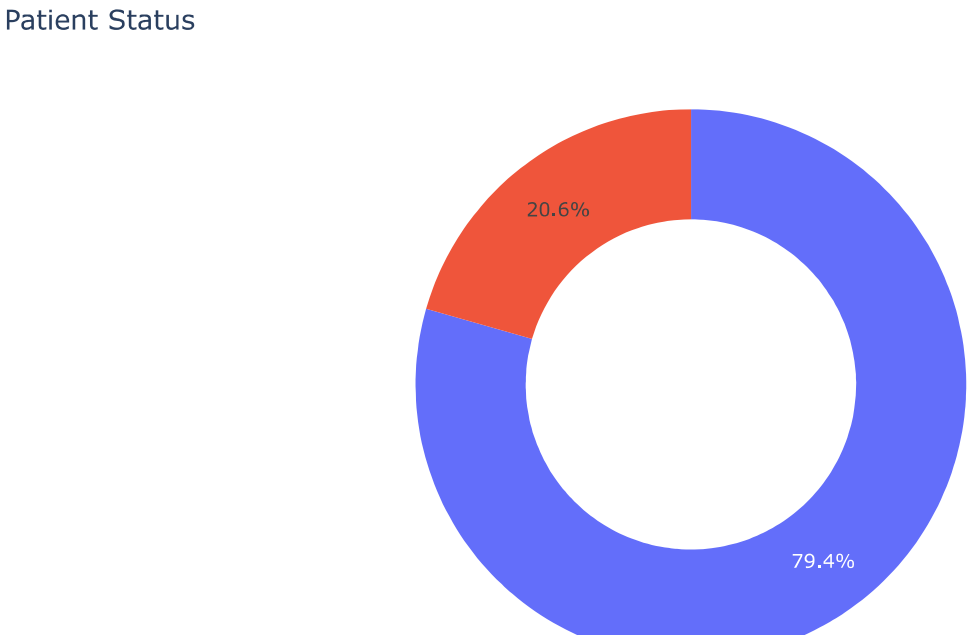
Out[9]:
```
FEMALE    330
MALE        4
Name: Gender, dtype: int64
```

In [11]:
```
p=px.pie(df,
       values=df['Gender'].value_counts().values,
       names=df['Gender'].value_counts().index,
       hole=.7 )
p.show()
```



In [12]:
```
Patient_Status = df['Patient_Status'].value_counts()
transactions = Patient_Status.index
quantity = Patient_Status.values

figure = px.pie(df,
             values=quantity,
             names=transactions, hole =.60,
             title="Patient Status")
figure.show()
```




In [13]:
```
df.dropna(inplace=True)
```

we drop value cause we before seen in target coloumn there are some null value

In [14]:
```
df.head()
```

Out[14]:

| | Patient_ID | Age | Gender | Protein1 | Protein2 | Protein3 | Protein4 | Tumour_Stage | Histology | ER status | PR status | HER2 status | Surge |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | TCGA-D8-A1XD | 36 | FEMALE | 0.080353 | 0.42638 | 0.54715 | 0.273680 | III | Infiltrating Ductal Carcinoma | Positive | Positive | Negative | M Mast |
| 1 | TCGA-EW-A1OX | 43 | FEMALE | -0.420320 | 0.57807 | 0.61447 | -0.031505 | II | Mucinous Carcinoma | Positive | Positive | Negative | Lump |
| 2 | TCGA-A8-A079 | 69 | FEMALE | 0.213980 | 1.31140 | -0.32747 | -0.234260 | III | Infiltrating Ductal Carcinoma | Positive | Positive | Negative | |
| 3 | TCGA-D8-A1XR | 56 | FEMALE | 0.345090 | -0.21147 | -0.19304 | 0.124270 | II | Infiltrating Ductal Carcinoma | Positive | Positive | Negative | M Mast |
| 4 | TCGA-BH-A0BF | 56 | FEMALE | 0.221550 | 1.90680 | 0.52045 | -0.311990 | II | Infiltrating Ductal Carcinoma | Positive | Positive | Negative | |

there are some unnecessary column that wasn't relate with our target

In [15]:
```
df.drop(['Patient_ID','ER status','PR status','Date_of_Surgery','Date_of_Last_Visit'],axis=1,inplace=True
```

In [16]:
```
df.shape
```

Out[16]: (317, 11)

In [17]:
```
df.head()
```

Out[17]:

| | Age | Gender | Protein1 | Protein2 | Protein3 | Protein4 | Tumour_Stage | Histology | HER2 status | Surgery_type | Patient_Status |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 36 | FEMALE | 0.080353 | 0.42638 | 0.54715 | 0.273680 | III | Infiltrating Ductal Carcinoma | Negative | Modified Radical Mastectomy | Alive |
| **1** | 43 | FEMALE | -0.420320 | 0.57807 | 0.61447 | -0.031505 | II | Mucinous Carcinoma | Negative | Lumpectomy | Dead |
| **2** | 69 | FEMALE | 0.213980 | 1.31140 | -0.32747 | -0.234260 | III | Infiltrating Ductal Carcinoma | Negative | Other | Alive |
| **3** | 56 | FEMALE | 0.345090 | -0.21147 | -0.19304 | 0.124270 | II | Infiltrating Ductal Carcinoma | Negative | Modified Radical Mastectomy | Alive |
| **4** | 56 | FEMALE | 0.221550 | 1.90680 | 0.52045 | -0.311990 | II | Infiltrating Ductal Carcinoma | Negative | Other | Dead |

In [18]:
```
from sklearn.preprocessing import LabelEncoder
```

In [19]:
```
le=LabelEncoder()
```

In [20]:
```
from pandas.core.dtypes.common import is_numeric_dtype
```

In [21]:
```
for i in df.columns:
    if is_numeric_dtype(df[i]):
        continue
    else:
        df[i]=le.fit_transform(df[i])
```

In [38]:
```
x=df.drop('Patient_Status',axis=1)
y=df[['Patient_Status']]
```

In [39]:
```
from imblearn.over_sampling import RandomOverSampler
```

In [40]:
```
ros=RandomOverSampler(random_state=42)
```

In [41]:
```
new_x,new_y=ros.fit_resample(x,y)
```

In [56]:
```
new_x.head()
```

Out[56]:

| | Age | Gender | Protein1 | Protein2 | Protein3 | Protein4 | Tumour_Stage | Histology | HER2 status | Surgery_type |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 36 | 0 | 0.080353 | 0.42638 | 0.54715 | 0.273680 | 2 | 0 | 0 | 1 |
| **1** | 43 | 0 | -0.420320 | 0.57807 | 0.61447 | -0.031505 | 1 | 2 | 0 | 0 |
| **2** | 69 | 0 | 0.213980 | 1.31140 | -0.32747 | -0.234260 | 2 | 0 | 0 | 2 |
| **3** | 56 | 0 | 0.345090 | -0.21147 | -0.19304 | 0.124270 | 1 | 0 | 0 | 1 |
| **4** | 56 | 0 | 0.221550 | 1.90680 | 0.52045 | -0.311990 | 1 | 0 | 0 | 2 |

In [57]:
```
df2 = pd.concat([new_x, new_y], axis=1)
df2.shape
```

Out[57]: (510, 11)

In [58]:
```
df.shape
```

Out[58]:
```
(510, 11)
```

In [61]:
```
df2.head()
```

Out[61]:

| | Age | Gender | Protein1 | Protein2 | Protein3 | Protein4 | Tumour_Stage | Histology | HER2 status | Surgery_type | Patient_Status |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 36 | 0 | 0.080353 | 0.42638 | 0.54715 | 0.273680 | 2 | 0 | 0 | 1 | 0 |
| **1** | 43 | 0 | -0.420320 | 0.57807 | 0.61447 | -0.031505 | 1 | 2 | 0 | 0 | 1 |
| **2** | 69 | 0 | 0.213980 | 1.31140 | -0.32747 | -0.234260 | 2 | 0 | 0 | 2 | 0 |
| **3** | 56 | 0 | 0.345090 | -0.21147 | -0.19304 | 0.124270 | 1 | 0 | 0 | 1 | 0 |
| **4** | 56 | 0 | 0.221550 | 1.90680 | 0.52045 | -0.311990 | 1 | 0 | 0 | 2 | 1 |

In [62]:
```
from pycaret.classification import *
```

In [64]:
```
setup(data=df2,target='Patient_Status',
      fix_imbalance=True,
      fix_imbalance_method='randomoversampler',

      normalize=True,
      remove_multicollinearity=True,
      log_experiment=True,

     )
```

| | Description | Value |
|---|---|---|
| 0 | Session id | 8963 |
| 1 | Target | Patient_Status |
| 2 | Target type | Binary |
| 3 | Original data shape | (510, 11) |
| 4 | Transformed data shape | (511, 11) |
| 5 | Transformed train set shape | (358, 11) |
| 6 | Transformed test set shape | (153, 11) |
| 7 | Numeric features | 10 |
| 8 | Preprocess | True |
| 9 | Imputation type | simple |
| 10 | Numeric imputation | mean |
| 11 | Categorical imputation | mode |
| 12 | Remove multicollinearity | True |
| 13 | Multicollinearity threshold | 0.900000 |
| 14 | Fix imbalance | True |
| 15 | Fix imbalance method | randomoversampler |
| 16 | Normalize | True |
| 17 | Normalize method | zscore |
| 18 | Fold Generator | StratifiedKFold |
| 19 | Fold Number | 10 |
| 20 | CPU Jobs | -1 |
| 21 | Use GPU | False |
| 22 | Log Experiment | MlflowLogger |
| 23 | Experiment Name | clf-default-name |
| 24 | USI | 8433 |

Out[64]: `<pycaret.classification.oop.ClassificationExperiment at 0x17a1cd61290>`

In [65]:
```
compare_models()
```

| | Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC | TT (Sec) |
|---|---|---|---|---|---|---|---|---|---|
| **et** | Extra Trees Classifier | 0.9357 | 0.9678 | 0.9379 | 0.9379 | 0.9356 | 0.8714 | 0.8758 | 0.2190 |
| **rf** | Random Forest Classifier | 0.9046 | 0.9660 | 0.9379 | 0.8865 | 0.9079 | 0.8090 | 0.8177 | 0.2900 |
| **catboost** | CatBoost Classifier | 0.8824 | 0.9637 | 0.9435 | 0.8462 | 0.8895 | 0.7647 | 0.7759 | 2.5770 |
| **xgboost** | Extreme Gradient Boosting | 0.8795 | 0.9529 | 0.9490 | 0.8351 | 0.8868 | 0.7592 | 0.7706 | 0.0930 |
| **gbc** | Gradient Boosting Classifier | 0.8575 | 0.9304 | 0.9379 | 0.8128 | 0.8683 | 0.7149 | 0.7301 | 0.1610 |
| **lightgbm** | Light Gradient Boosting Machine | 0.8516 | 0.9451 | 0.9490 | 0.7961 | 0.8643 | 0.7034 | 0.7212 | 0.1600 |
| **dt** | Decision Tree Classifier | 0.8181 | 0.8185 | 0.9490 | 0.7570 | 0.8403 | 0.6365 | 0.6629 | 0.0440 |
| **ada** | Ada Boost Classifier | 0.7371 | 0.7980 | 0.8029 | 0.7116 | 0.7530 | 0.4741 | 0.4802 | 0.1280 |
| **qda** | Quadratic Discriminant Analysis | 0.6867 | 0.7438 | 0.8209 | 0.6504 | 0.7234 | 0.3730 | 0.3910 | 0.0460 |
| **knn** | K Neighbors Classifier | 0.6779 | 0.7377 | 0.7922 | 0.6456 | 0.7059 | 0.3561 | 0.3747 | 0.0620 |
| **ridge** | Ridge Classifier | 0.5884 | 0.0000 | 0.6190 | 0.5862 | 0.6000 | 0.1769 | 0.1785 | 0.0420 |
| **svm** | SVM - Linear Kernel | 0.5858 | 0.0000 | 0.5840 | 0.5819 | 0.5775 | 0.1706 | 0.1725 | 0.0460 |
| **lr** | Logistic Regression | 0.5856 | 0.6230 | 0.6245 | 0.5815 | 0.6003 | 0.1713 | 0.1734 | 1.1410 |
| **lda** | Linear Discriminant Analysis | 0.5801 | 0.6209 | 0.6190 | 0.5780 | 0.5958 | 0.1602 | 0.1617 | 0.0450 |
| **nb** | Naive Bayes | 0.5689 | 0.6161 | 0.8598 | 0.5432 | 0.6639 | 0.1388 | 0.1779 | 0.0480 |
| **dummy** | Dummy Classifier | 0.5014 | 0.5000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0420 |

Out[65]:
```
ExtraTreesClassifier

ExtraTreesClassifier(bootstrap=False, ccp_alpha=0.0, class_weight=None,
                     criterion='gini', max_depth=None, max_features='sqrt',
                     max_leaf_nodes=None, max_samples=None,
                     min_impurity_decrease=0.0, min_samples_leaf=1,
                     min_samples_split=2, min_weight_fraction_leaf=0.0,
                     n_estimators=100, n_jobs=-1, oob_score=False,
                     random_state=8963, verbose=0, warm_start=False)
```

In [66]:
```

```

Out[66]: (510, 11)

In [73]:
```
etc=create_model('et')
```

| Fold | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| 0 | 0.9444 | 0.9599 | 0.9444 | 0.9444 | 0.9444 | 0.8889 | 0.8889 |
| 1 | 0.9167 | 0.9691 | 0.8889 | 0.9412 | 0.9143 | 0.8333 | 0.8346 |
| 2 | 0.8333 | 0.8889 | 0.8333 | 0.8333 | 0.8333 | 0.6667 | 0.6667 |
| 3 | 0.9444 | 1.0000 | 1.0000 | 0.9000 | 0.9474 | 0.8889 | 0.8944 |
| 4 | 0.9722 | 0.9583 | 0.9444 | 1.0000 | 0.9714 | 0.9444 | 0.9459 |
| 5 | 0.9722 | 1.0000 | 1.0000 | 0.9474 | 0.9730 | 0.9444 | 0.9459 |
| 6 | 0.9444 | 0.9676 | 0.9444 | 0.9444 | 0.9444 | 0.8889 | 0.8889 |
| 7 | 0.9714 | 1.0000 | 1.0000 | 0.9474 | 0.9730 | 0.9427 | 0.9443 |
| 8 | 0.9429 | 1.0000 | 1.0000 | 0.8947 | 0.9444 | 0.8860 | 0.8918 |
| 9 | 0.9143 | 0.9330 | 0.8235 | 1.0000 | 0.9032 | 0.8276 | 0.8402 |
| Mean | 0.9356 | 0.9677 | 0.9379 | 0.9353 | 0.9349 | 0.8712 | 0.8742 |
| Std | 0.0394 | 0.0343 | 0.0648 | 0.0471 | 0.0405 | 0.0788 | 0.0788 |

In [74]:
```
predict_model(etc,data=df2)
```

| | Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|---|---|---|---|---|---|---|---|---|
| 0 | Extra Trees Classifier | 0.9804 | 0.9944 | 0.9725 | 0.9880 | 0.9802 | 0.9608 | 0.9609 |

Out[74]:

| | Age | Gender | Protein1 | Protein2 | Protein3 | Protein4 | Tumour_Stage | Histology | HER2 status | Surgery_type | Patient_Status | predic |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 36 | 0 | 0.080353 | 0.42638 | 0.547150 | 0.273680 | 2 | 0 | 0 | 1 | 0 | |
| 1 | 43 | 0 | -0.420320 | 0.57807 | 0.614470 | -0.031505 | 1 | 2 | 0 | 0 | 1 | |
| 2 | 69 | 0 | 0.213980 | 1.31140 | -0.327470 | -0.234260 | 2 | 0 | 0 | 2 | 0 | |
| 3 | 56 | 0 | 0.345090 | -0.21147 | -0.193040 | 0.124270 | 1 | 0 | 0 | 1 | 0 | |
| 4 | 56 | 0 | 0.221550 | 1.90680 | 0.520450 | -0.311990 | 1 | 0 | 0 | 2 | 1 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 505 | 64 | 0 | -0.969950 | -0.76926 | 0.556800 | -0.720150 | 0 | 0 | 0 | 1 | 1 | |
| 506 | 56 | 0 | 0.326000 | 1.86020 | -1.077100 | 0.336640 | 2 | 0 | 0 | 1 | 1 | |
| 507 | 61 | 0 | -0.719470 | 2.54850 | -0.150240 | 0.339680 | 1 | 0 | 0 | 0 | 1 | |
| 508 | 47 | 0 | 0.121070 | 0.78513 | -0.197620 | 0.352450 | 1 | 1 | 0 | 2 | 1 | |
| 509 | 60 | 0 | 0.292200 | 1.77530 | -0.093631 | 0.567040 | 1 | 0 | 0 | 3 | 1 | |

510 rows × 13 columns

In [75]:
```
save_model(etc,'Extra_Trees_Classifier')
```

```
Transformation Pipeline and Model Successfully Saved
```

Out[75]:
```
(Pipeline(memory=Memory(location=None),
          steps=[('numerical_imputer',
                  TransformerWrapper(exclude=None,
                                     include=['Age', 'Gender', 'Protein1',
                                              'Protein2', 'Protein3', 'Protein4',
                                              'Tumour_Stage', 'Histology',
                                              'HER2 status', 'Surgery_type'],
                                     transformer=SimpleImputer(add_indicator=False,
                                                               copy=True,
                                                               fill_value=None,
                                                               keep_empty_features=False,
                                                               missing_values=nan,
                                                               strategy='m...
                  ExtraTreesClassifier(bootstrap=False, ccp_alpha=0.0,
                                       class_weight=None, criterion='gini',
                                       max_depth=None, max_features='sqrt',
                                       max_leaf_nodes=None, max_samples=None,
                                       min_impurity_decrease=0.0,
                                       min_samples_leaf=1, min_samples_split=2,
                                       min_weight_fraction_leaf=0.0,
                                       n_estimators=100, n_jobs=-1,
                                       oob_score=False, random_state=8963,
                                       verbose=0, warm_start=False))],
          verbose=False),
 'Extra_Trees_Classifier.pkl')
```

In [ ]:
```

```