File Edit Insert Tools View Help | 27_ BreastcancerTumorPrediction_XGBOOST... — Edited

41 Days Evaluated

Mind Map | Outliner

Topic | Subtopic | Relationship | Summary | Boundary | Insert | ZEN | Pitch | Panel

2 **Dataset** structure

Breast cancer T
prediction us
XGBOOST

dataset.csv - Excel

File | Home | Insert | Page Layout | Formulas | Data | Review | View | Tell me what you want to do... | Sign in | Share

Paste | Calibri | 11 | Wrap Text | General | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | Sort & Filter | Find & Select

Clipboard | Font | Alignment | Number | Styles | Cells | Editing

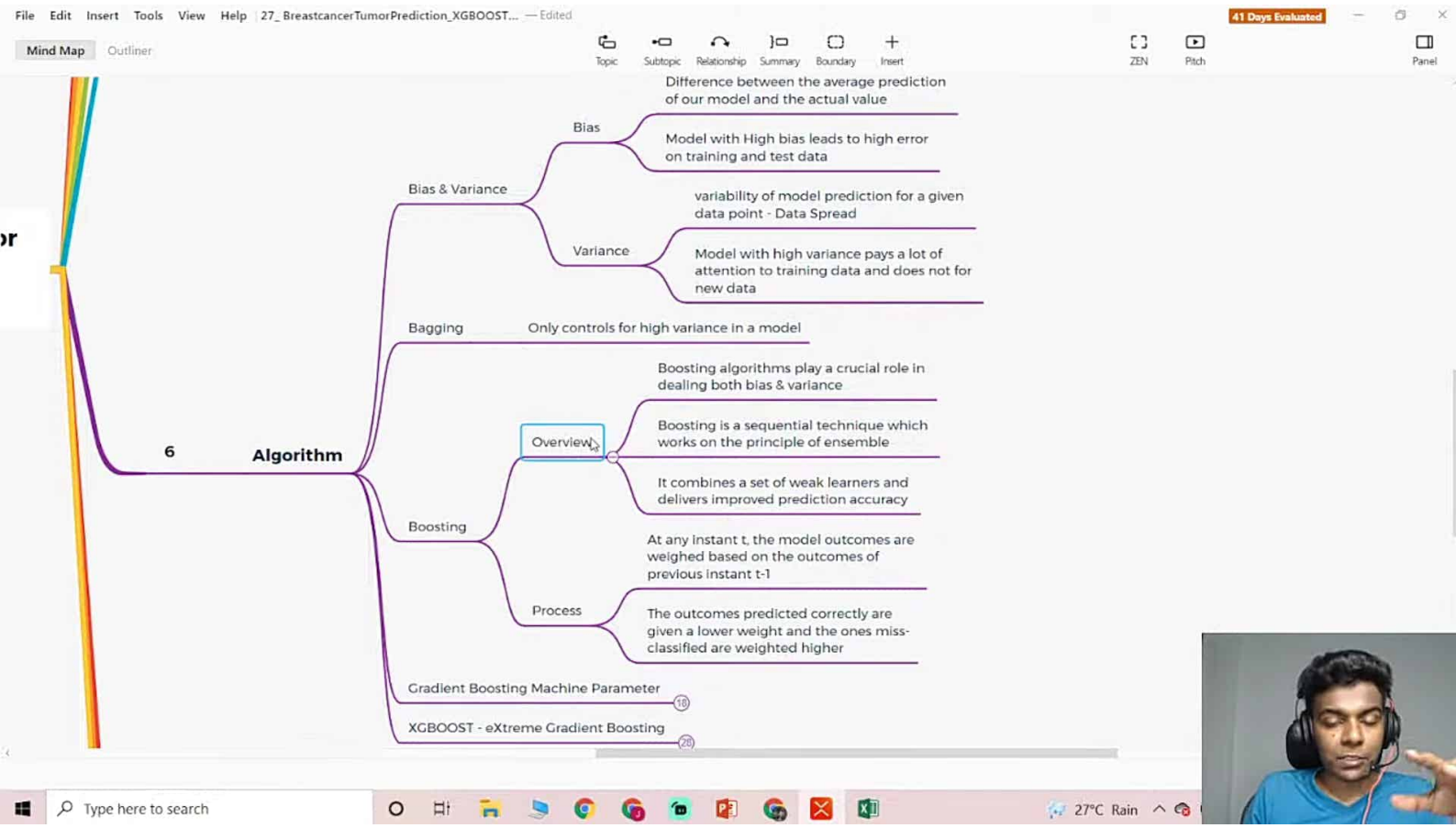
K2 | 2

	B	C	D	E	F	G	H	I	J	K	L
1	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial C	Bare Nuclei	Bland Chrc	Normal Nu	Mitoses	Class	
2	5	1	1	1	2	1	3	1	1	2	
3	5	4	4	5	7	10	3	2	1	2	
4	3	1	1	1	2	2	3	1	1	2	
5	6	8	8	1	3	4	3	7	1	2	
6	4	1	1	3	2	1	3	1	1	2	
7	8	10	10	8	7	10	9	7	1	4	
8	1	1	1	1	2	10	3	1	1	2	
9	2	1	2	1	2	1	3	1	1	2	
10	2	1	1	1	2	1	1	1	5	2	
11	4	2	1	1	2	1	2	1	1	2	
12	1	1	1	1	1	1	3	1	1	2	
13	2	1	1	1	2	1	2	1	1	2	
14	5	3	3	3	2	3	4	4	1	4	
15	1	1	1	1	2	3	3	1	1	2	
16	8	7	5	10	7	9	5	5	4	4	
17	7	4	6	4	6	1	4	3	1	4	
18	4	1	1	1	2	1	2	1	1		

dataset

Ready

9

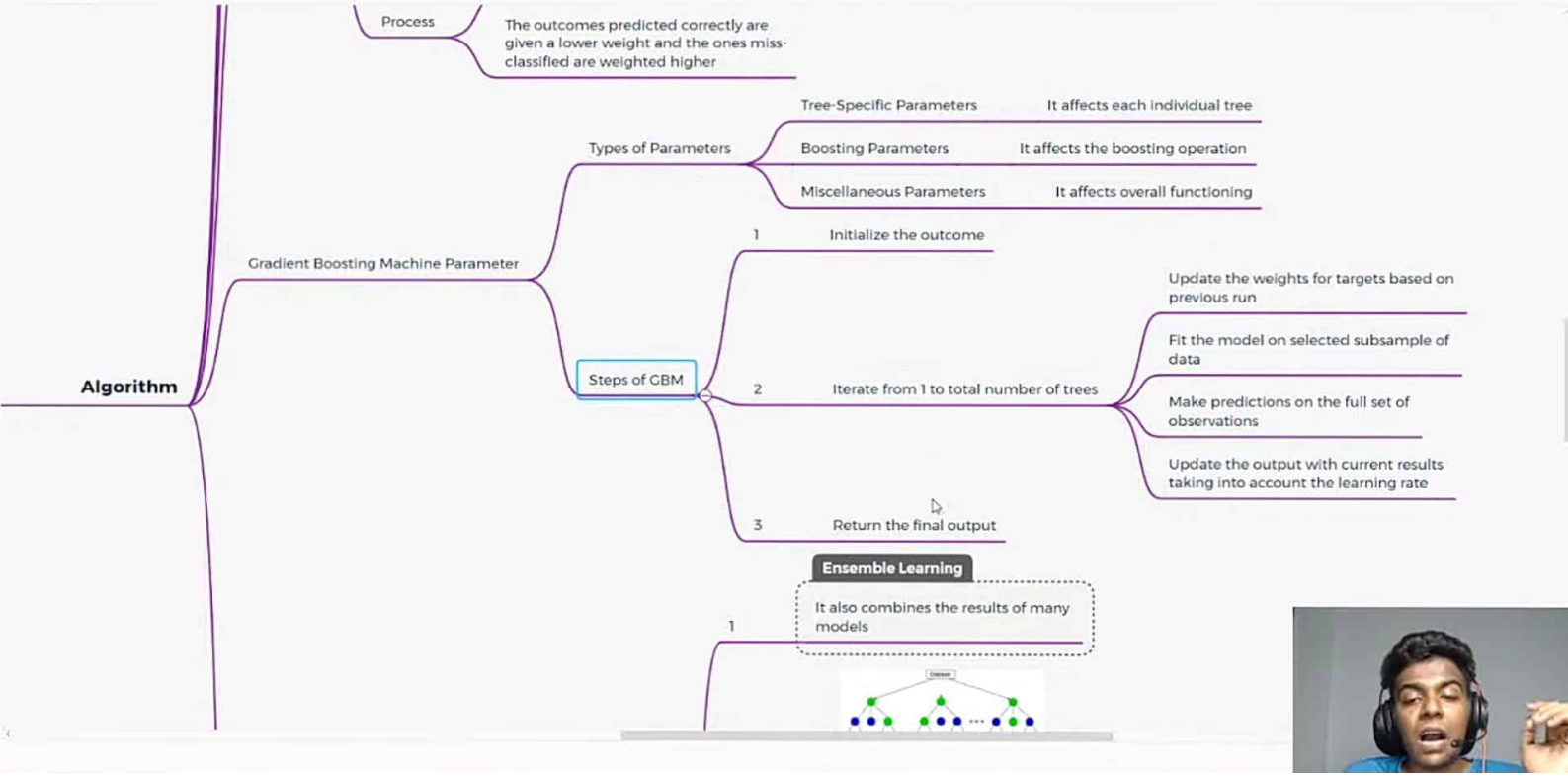

File Edit Insert Tools View Help 27_ BreastcancerTumorPrediction_XGBOOST... — Edited
41 Days Evaluated
Mind Map
Outliner
Topic
Subtopic
Relationship
Summary
Boundary
Insert
ZEN
Pitch
Panel
6
Algorithm
Bias & Variance
Bias
Difference between the average prediction of our model and the actual value
Model with High bias leads to high error on training and test data
Variance
variability of model prediction for a given data point - Data Spread
Model with high variance pays a lot of attention to training data and does not for new data
Bagging
Only controls for high variance in a model
Boosting
Overview
Boosting algorithms play a crucial role in dealing both bias & variance
Boosting is a sequential technique which works on the principle of ensemble
It combines a set of weak learners and delivers improved prediction accuracy
Process
At any instant t, the model outcomes are weighed based on the outcomes of previous instant t-1
The outcomes predicted correctly are given a lower weight and the ones miss-classified are weighted higher
Gradient Boosting Machine Parameter
18
XGBOOST - eXtreme Gradient Boosting
28
Type here to search
27°C Rain

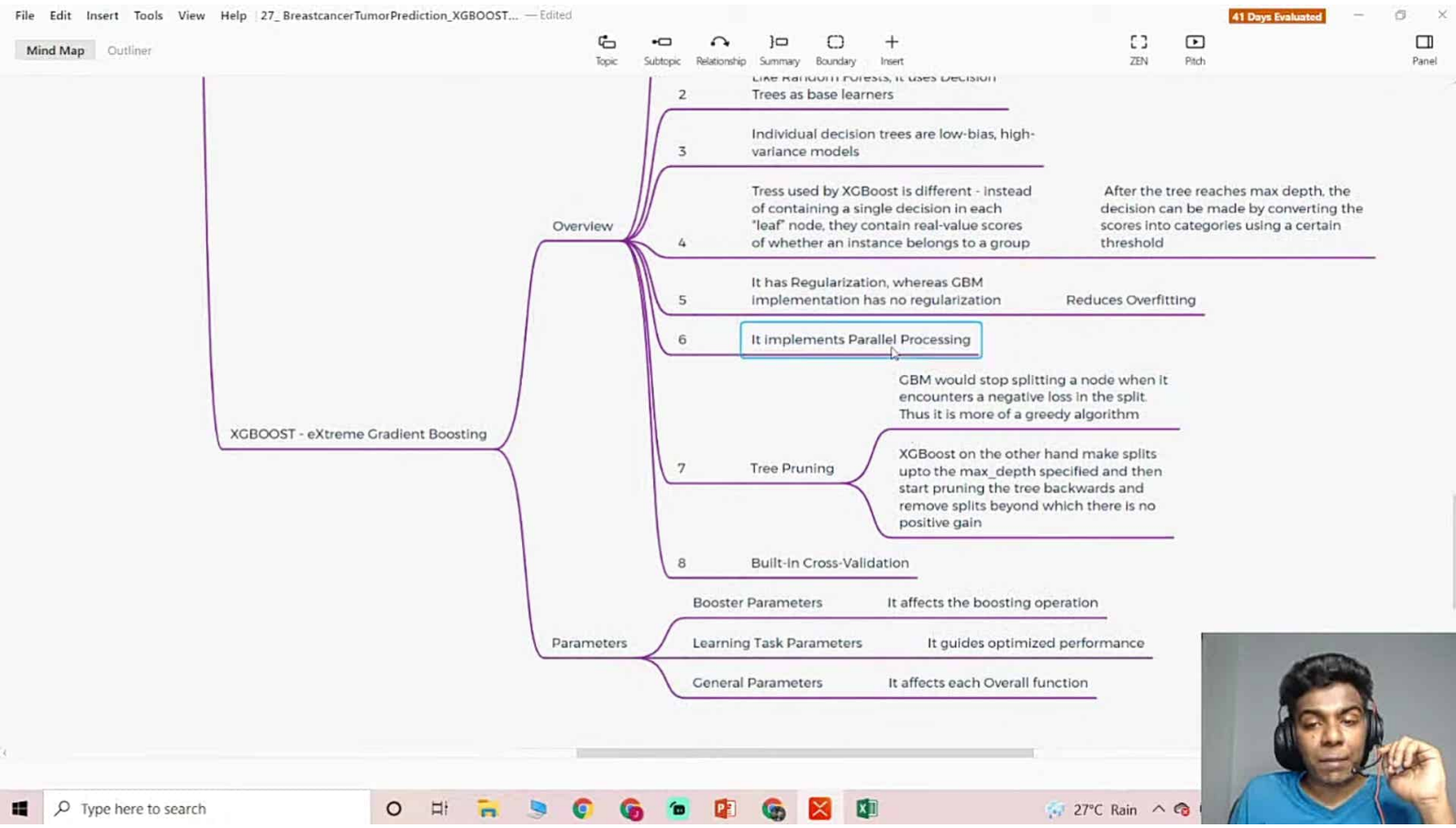
File Edit Insert Tools View Help | 27_ BreastcancerTumorPrediction_XGBOOST... — Edited

41 Days Evaluated

Mind Map | Outliner

Topic | Subtopic | Relationship | Summary | Boundary | Insert | ZEN | Pitch | Panel




File Edit Insert Tools View Help 27_ BreastcancerTumorPrediction_XGBOOST... — Edited
41 Days Evaluated
Mind Map
Outliner
Topic
Subtopic
Relationship
Summary
Boundary
Insert
ZEN
Pitch
Panel
XGBOOST - eXtreme Gradient Boosting
Overview
2
Trees as base learners
3
Individual decision trees are low-bias, high-variance models
4
Tress used by XGBoost is different - instead of containing a single decision in each "leaf" node, they contain real-value scores of whether an instance belongs to a group
After the tree reaches max depth, the decision can be made by converting the scores into categories using a certain threshold
5
It has Regularization, whereas GBM implementation has no regularization
Reduces Overfitting
6
It implements Parallel Processing
7
Tree Pruning
GBM would stop splitting a node when it encounters a negative loss in the split. Thus it is more of a greedy algorithm
XGBoost on the other hand make splits upto the max_depth specified and then start pruning the tree backwards and remove splits beyond which there is no positive gain
8
Built-in Cross-Validation
Parameters
Booster Parameters
It affects the boosting operation
Learning Task Parameters
It guides optimized performance
General Parameters
It affects each Overall function
Type here to search
27°C Rain

27_ BreastcancerTumorPrediction_XGBOOST_ML.ipy...

Comment Share

File Edit View Insert Runtime Tools Help

+ Code + Text

RAM Disk

Editing

Importing the basic libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
```

Load Dataset from Local Directory

```
[ ]  from google.colab import files
```

0s completed at 19:23

27_ BreastcancerTumorPrediction_XGBOOST_ML.ipy...

Comment Share

File Edit View Insert Runtime Tools Help

+ Code + Text

RAM Disk

Editing

- **dataset.csv**(application/vnd.ms-excel) - 19635 bytes, last modified: 5/9/2020 - 100% done

```
Saving dataset.csv to dataset.csv
```

Importing the dataset

```
dataset = pd.read_csv('dataset.csv')
print(dataset.shape)
print(dataset.head(5))
```

```
(683, 11)
   Sample code number  Clump Thickness  ...  Mitoses  Class
0             1000025                5  ...        1      2
1             1002945                5  ...        1      2
2             1015425                3  ...        1      2
```

7s completed at 19:24

27_ BreastcancerTumorPrediction_XGBOOST_ML.ipy...

Comment Share

File Edit View Insert Runtime Tools Help

+ Code + Text

RAM Disk

Editing

```
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values
```

0s

Splitting Dataset into Train & Test

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

Training with XGBoost

```
[ ]  from xgboost import XGBClassifier
```

0s completed at 19:24

27_ BreastcancerTumorPrediction_XGBOOST_ML.ipy...

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

Editing

Training with XGBoost

```
from xgboost import XGBClassifier
model = XGBClassifier()
model.fit(X_train, y_train)
```

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0,
              learning_rate=0.1, max_delta_step=0, max_depth=3,
              min_child_weight=1, missing=None, n_estimators=100, n_jobs=1,
              nthread=None, objective='binary:logistic', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=1, verbosity=1)
```

0s completed at 19:24

27_ BreastcancerTumorPrediction_XGBOOST_ML.ipy...

Comment Share

File Edit View Insert Runtime Tools Help

+ Code + Text

RAM Disk

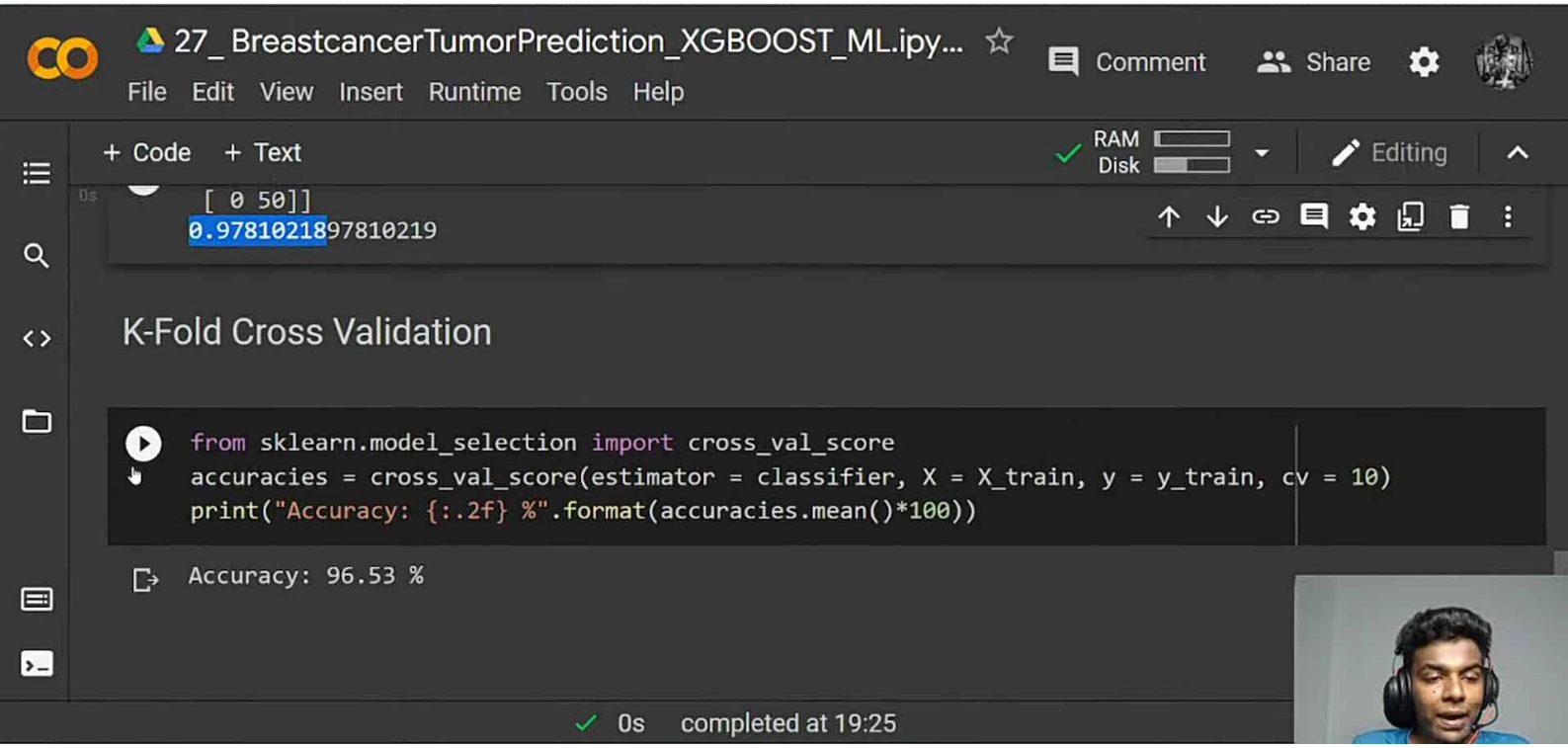
Editing

```
from sklearn.metrics import confusion_matrix, accuracy_score
y_pred = model.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print(cm)
accuracy_score(y_test, y_pred)
```

```
[[84  3]
 [ 0 50]]
0.9781021897810219
```

K-Fold Cross Validation

0s completed at 19:24


27_ BreastcancerTumorPrediction_XGBOOST_ML.ipy...
Comment
Share
File Edit View Insert Runtime Tools Help
+ Code + Text
RAM
Disk
Editing
[0 50]]
0.9781021897810219
K-Fold Cross Validation
from sklearn.model_selection import cross_val_score
accuracies = cross_val_score(estimator = classifier, X = X_train, y = y_train, cv = 10)
print("Accuracy: {:.2f} %".format(accuracies.mean()*100))
Accuracy: 96.53 %
0s completed at 19:25