

## Project Development Phase Model Performance Test

Date	16 November 2022
Team ID	PNT2022TMID08848
Project Name	Project – Early Detection of Chronic Kidney Disease using Machine Learning
Maximum Marks	10 Marks

### Model Performance Testing:

Project team shall fill the following information in model performance testing template.

S.No.	Parameter	Values	Screenshot
1.	Metrics	<b>Regression Model:</b> MAE - , MSE - , RMSE - , R2 score -  <b>Classification Model:</b> Confusion Matrix - , Accuracy Score- & Classification Report -	See Below
2.	Tune the Model	Hyperparameter Tuning - Validation Method -	See Below

### 1. Metrics

#### Model: Random Forest Classifier

The screenshot displays a Jupyter Notebook interface with a file explorer on the left and a code editor on the right. The file explorer shows a project structure for 'KIDNEY\_CANCER\_PREDICTION\_USING\_MAC...' with folders like '\_pycache\_', 'dataset', 'static', and 'templates', and files like 'models.py', 'README.md', 'requirements.txt', 'server.py', and 'writeCsv.py'. The code editor shows the following Python code:

```
[44] import numpy as np
[45]
[46]
[47] from sklearn.ensemble import RandomForestClassifier
[48]
[49] model = RandomForestClassifier(n_estimators = 20)
[50] model.fit(X_train, y_train)
[51]
[52] RandomForestClassifier(n_estimators=20)
[53]
[54] from sklearn.metrics import confusion_matrix, accuracy_score
[55]
[56] confusion_matrix(y_test, model.predict(X_test))
[57]
[58] array([[23,  0],
[59]        [ 0,  9]])
[60]
[61] print(f"Accuracy is {round(accuracy_score(y_test, model.predict(X_test))*100, 2)}%")
[62]
[63] Accuracy is 100.0%
```

A notification at the bottom right asks: "Do you want to install the recommended extensions for Python?"

## 2. Tune the Model

### Hyperparameter Tuning:

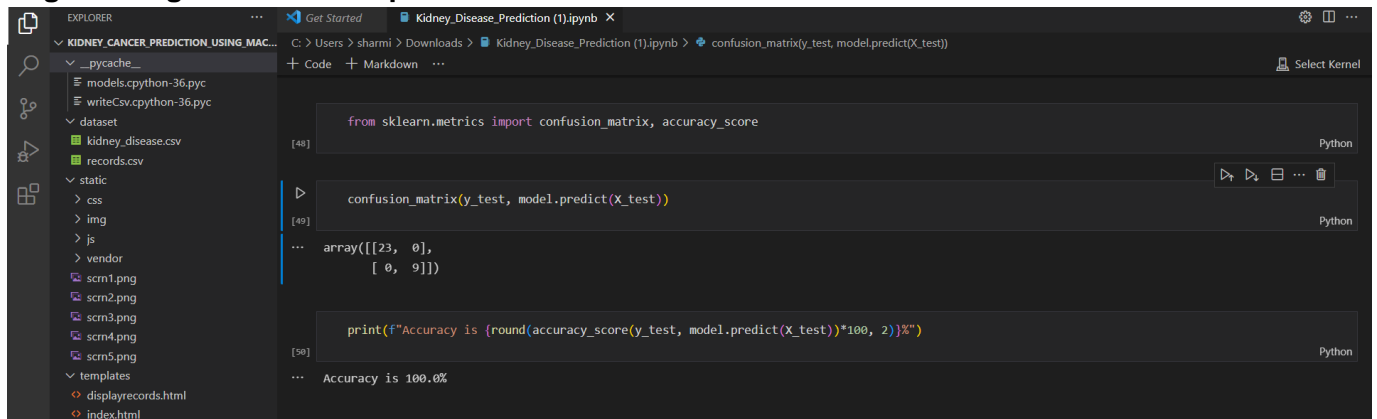
- The number of features is important and should be tuned in random forest classification.
- Initially all parameters in the dataset are taken as independent values to arrive at the dependent decision of Chronic Kidney Disease or No Chronic Kidney Disease.
- But the result was not accurate so used only 8 more correlated values as independent values to arrive at the dependent decision of Chronic Kidney Disease or not.

### Validation Method:

It involves **partitioning the training data set into subsets, where one subset is held out to test the performance of the model.** This data set is called the validation data set.

Cross validation is to use different models and identify the best:

### Logistic Regression Model performance values:



```
from sklearn.metrics import confusion_matrix, accuracy_score

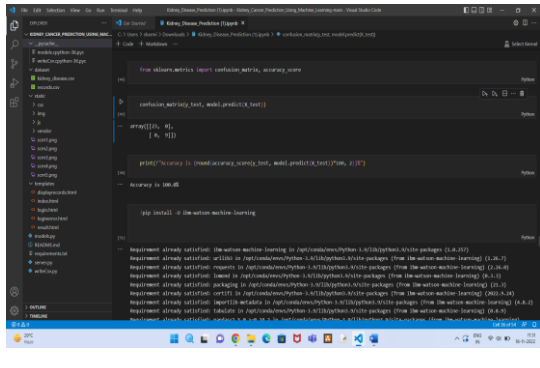
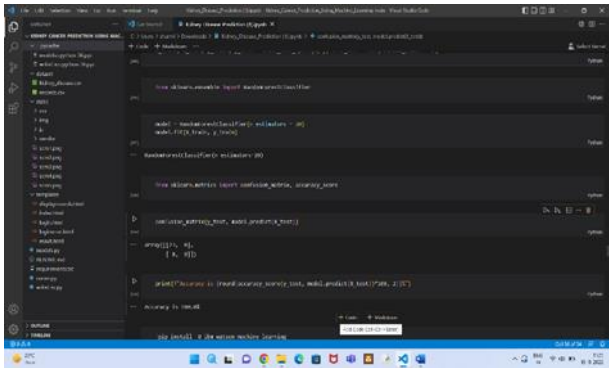
confusion_matrix(y_test, model.predict(X_test))

array([[23,  0],
       [ 0,  9]])

print(f"Accuracy is {round(accuracy_score(y_test, model.predict(X_test))*100, 2)}%")

Accuracy is 100.0%
```

Hence we tested with Logistic regression and Random Forest Classification wherein the accuracy of Random Forest classification is 95% compared with Logistic Regression.

Metric	Logistic Regression	Random Forest Classification
Accuracy	100.0%	100.0%
Other metrics		

The above table shows that Random Forest Classification gives better results over Logistic Regression.