

PROJECT DEVELOPMENT PHASE

PERFORMANCE TEST

Date	17 Nov 2022
Team ID	PNT2022TMID12917
Project Name	Early Detection of Chronic Kidney Disease Using Machine Learning
Maximum Marks	10 Marks

Model Performance Testing :

S.no	parameters	Values	screenshot
1	Metrics	Regression Model: MAE - , MSE - , RMSE - , R2 score - Classification Model: Confusion Matrix - , Accuracy Score- & Classification Report -	See below
2	Tune the model	Hyperparameter Tuning - Validation Method -	See below

1. Metrics

Model: LOGISTIC REGRESSION MODEL

Accuracy = 91.25%

Confusion matrix is displayed below.

```
: accuracy_score(y_test,y_pred)
: 0.9125

: confusion_mat = confusion_matrix(y_test,y_pred)
: confusion_mat
: array([[47,  7],
:        [ 0, 26]], dtype=int64)
```

Mean absolute error = 0.0875 = 8.75%

```
from sklearn.metrics import mean_absolute_error

mean_absolute_error(y_test,y_pred)

0.0875
```

Classification Report:

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	1.00	0.87	0.93	54
1	0.79	1.00	0.88	26
accuracy			0.91	80
macro avg	0.89	0.94	0.91	80
weighted avg	0.93	0.91	0.91	80

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.
- This caused the results to be less accurate. So, only 8 columns in the dataset were taken to be independent values and were used to train the ML model.

Validation Method:

Validation method involves partitioning the training data set into subsets, where one subset is used to test the performance of the ML model. This data set is called the validation data set.

The Linear Regression model is compared with the Gradient Boosting Classifier.

GRADIENT BOOSTING CLASSIFIER:

```
from sklearn.ensemble import GradientBoostingClassifier

gb_clf = GradientBoostingClassifier(n_estimators=20, learning_rate=1, max_features=2, max_depth=2, random_state=0)
gb_clf.fit(x_train, y_train)

G:\Anaconda\lib\site-packages\sklearn\utils\validation.py:63: DataConversionWarning: A column-vector y was passed when
ay was expected. Please change the shape of y to (n_samples, ), for example using ravel().
    return f(*args, **kwargs)

GradientBoostingClassifier(max_depth=2, max_features=2, n_estimators=20,
                           random_state=0)

y_pred=gb_clf.predict(x_test)
accuracy_score(y_test,y_pred)

0.909375
```

The accuracy of the gradient boosting classifier = 0.90375 = 90.375%

The accuracy of Logistic regression ML model is slightly greater than Gradient Boosting Classifier and hence is used in Early Detection of Chronic Kidney Disease.

The Logistic Regression ML model that we have used here has better performance in speed and accuracy compared to other models. We have compared the performance metrics of 2 models and selected this as the best for the application. The model performed well for all the test cases.

The accuracy score, confusion matrix and error values are attached above to support the above statement.

The ML model trained is found to have good accuracy with very minimal error. So, the developed ML model for Early Detection of Chronic Kidney Disease is highly accurate.