Analysis and Predictive Models: Chronic Kidney Disease (CKD)

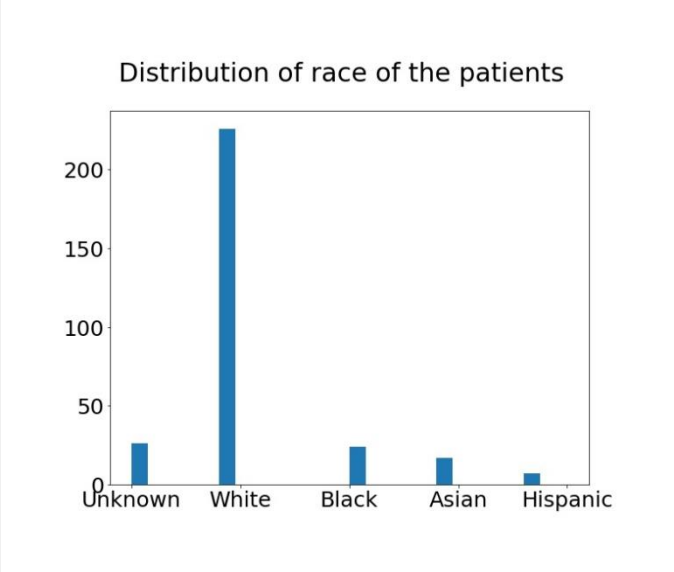
Holmusk Interview Challenge

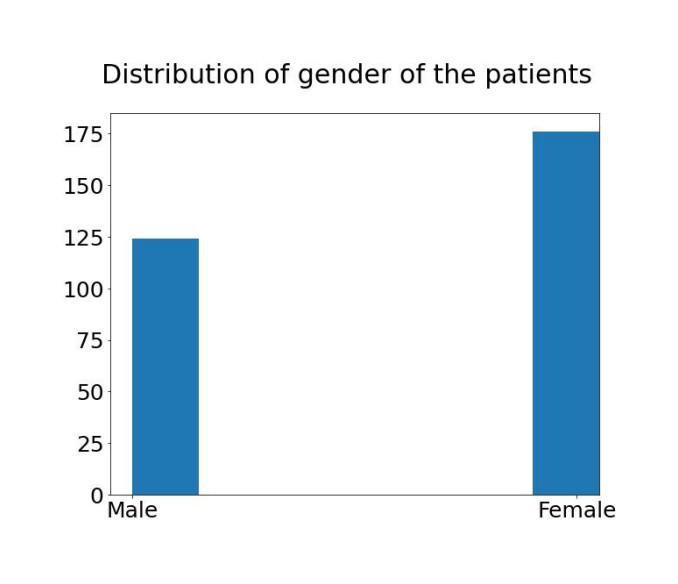
Jetin E Thomas

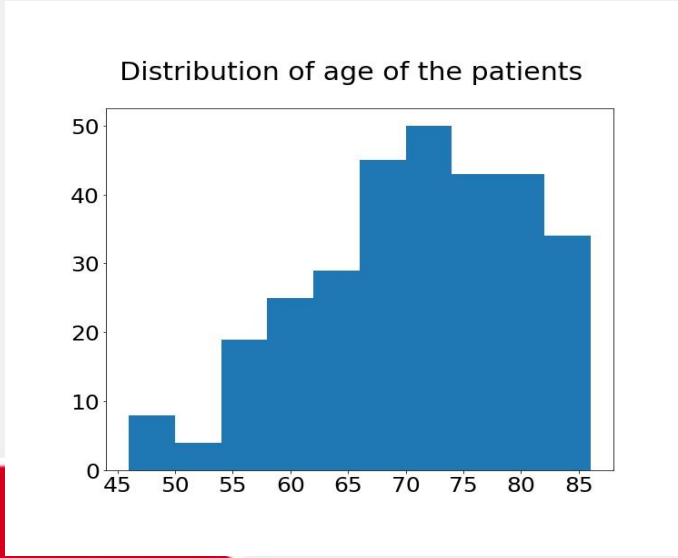
Description of Datasets

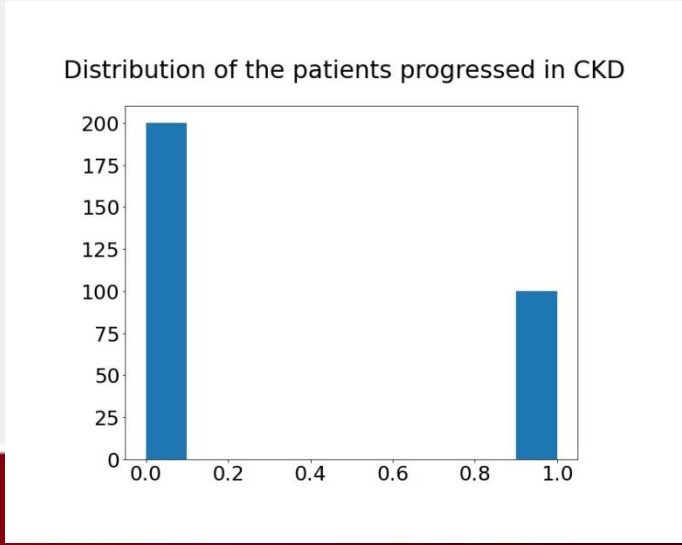
- The Dataset that is analyzed consists of 300 patients who have been diagnosed with chronic kidney disease (CKD). Their demographic information, medications and lab measurements along with their time is given in the dataset.
- The demographic information consists of the race, gender and age of the patients.
- The lab measurements performed on the patients are their serum creatinine count in mg/dl, diastolic blood pressure in mmHg, systolic blood pressure in mmHg, Hemoglobin level in g/dl, glucose level in mmol/l, low density lipoprotein level in mg/dl along with the time of measurement in days.
- The medications given to them along with the daily dosage, the starting and ending day of the prescription.
- Additionally, a datasheet telling these patients have progressed in chronic kidney disease or not.

Analysis: Demographic Information









Insights: Analysis

- The white race is predominant in the dataset.
- The female to male ratio is about 5:7.
- The peak distribution of age is between 70 and 75.
- The patients diagnosed and progressed in CKD is 1:2 in the dataset.

Predictive Model: Lab measurements including time in dataset

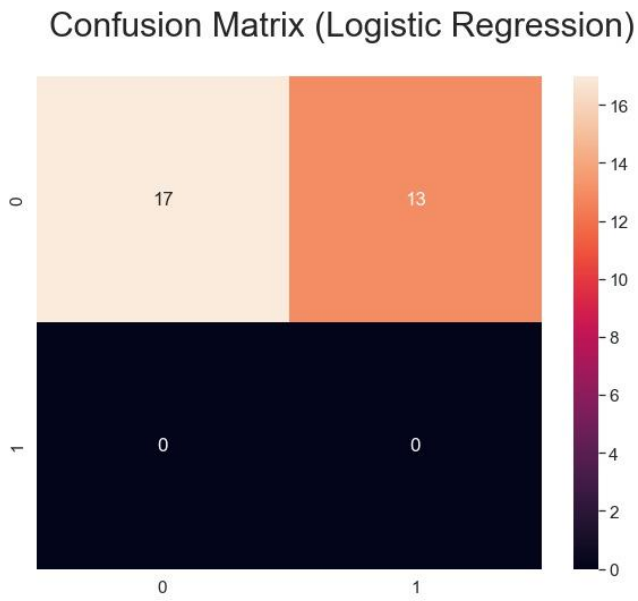
- Neural Network based on Logistic Regression

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_5 (Dense)              (None, 1000)              4000
_________________________________________________________________
dense_6 (Dense)              (None, 1000)              1001000
_________________________________________________________________
dense_7 (Dense)              (None, 1000)              1001000
_________________________________________________________________
dense_8 (Dense)              (None, 1)                 1001
=================================================================
Total params: 2,007,001
Trainable params: 2,007,001
Non-trainable params: 0
_________________________________________________________________
```

Results: Demographic Information

- Logistic Regression

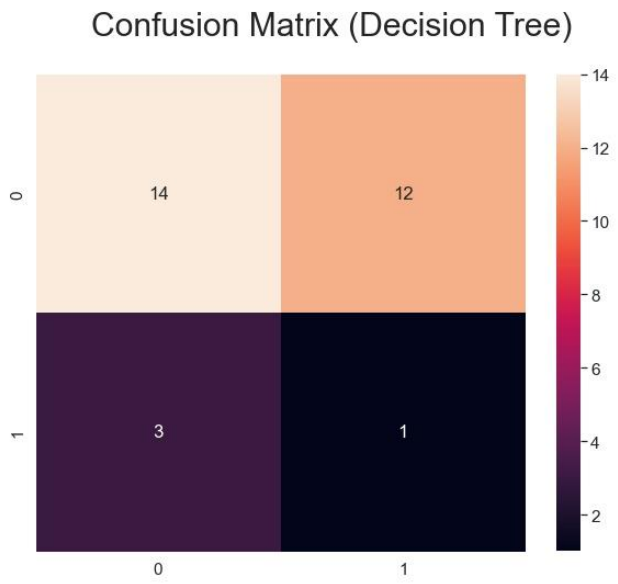
```
-----------------------------------------------------------------
Accuracy of Logistic Regression Classifier : 0.5666666666666667
-----------------------------------------------------------------
Classification report :
              precision    recall  f1-score   support

           1       0.00      0.00      0.00        13
           0       0.57      1.00      0.72        17

    accuracy                           0.57        30
   macro avg       0.28      0.50      0.36        30
weighted avg       0.32      0.57      0.41        30

-----------------------------------------------------------------
```



Results: Demographic Information

- Decision Tree

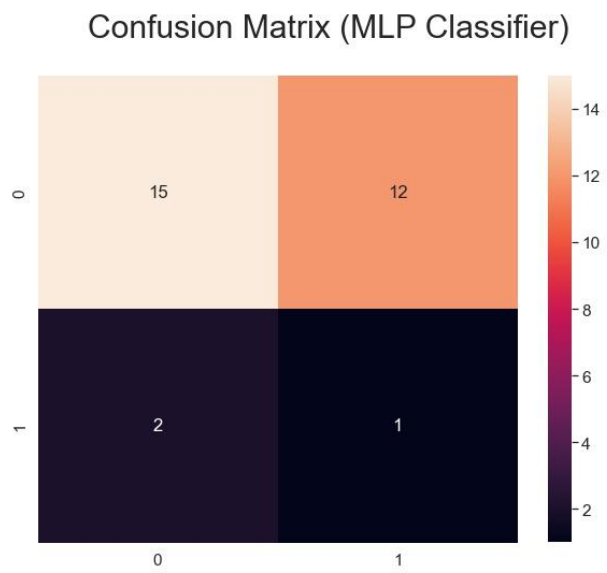
```
-----------------------------------------------------------
Accuracy of Decision Tree Classifier : 0.5
-----------------------------------------------------------
Classification report :
              precision    recall  f1-score   support

           1       0.25      0.08      0.12        13
           0       0.54      0.82      0.65        17

    accuracy                           0.50        30
   macro avg       0.39      0.45      0.38        30
weighted avg       0.41      0.50      0.42        30

-----------------------------------------------------------
```



Results: Demographic Information

- MLP Classifier

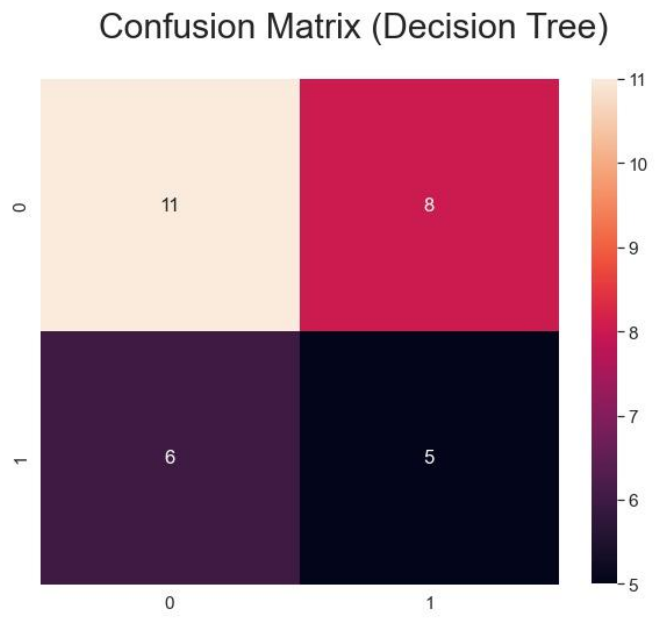
```
-----------------------------------------------------------------
Accuracy of MLPClassifier : 0.5333333333333333
-----------------------------------------------------------------
Classification report :
              precision    recall  f1-score   support

           1       0.33      0.08      0.12        13
           0       0.56      0.88      0.68        17

    accuracy                           0.53        30
   macro avg       0.44      0.48      0.40        30
weighted avg       0.46      0.53      0.44        30

-----------------------------------------------------------------
```



Results: Demographic Information

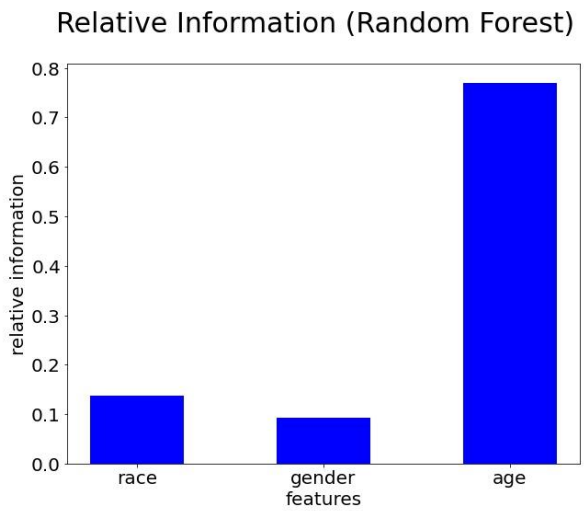
- Random Forest

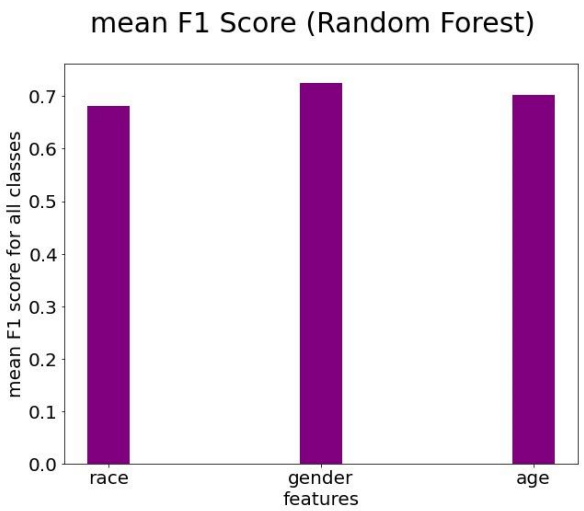
```
-----------------------------------------------------------------
Accuracy of Random Forest Classifier :  0.6333333333333333
-----------------------------------------------------------------
Classification report :
              precision    recall  f1-score   support

           1       0.60      0.46      0.52        13
           0       0.65      0.76      0.70        17

    accuracy                           0.63        30
   macro avg       0.62      0.61      0.61        30
weighted avg       0.63      0.63      0.62        30

-----------------------------------------------------------------
```



Further Results: Demographic Information

- Random Forest

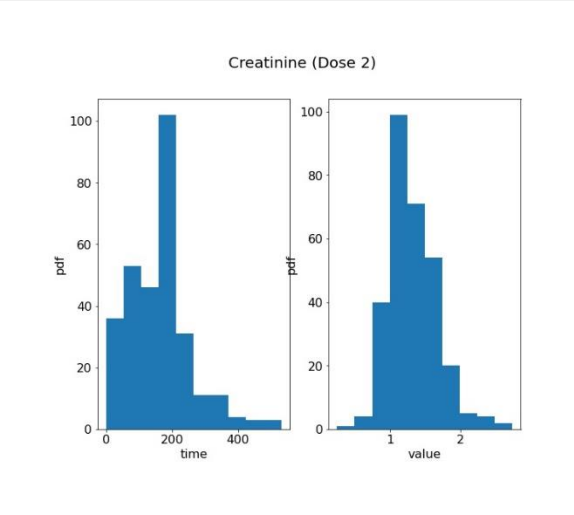


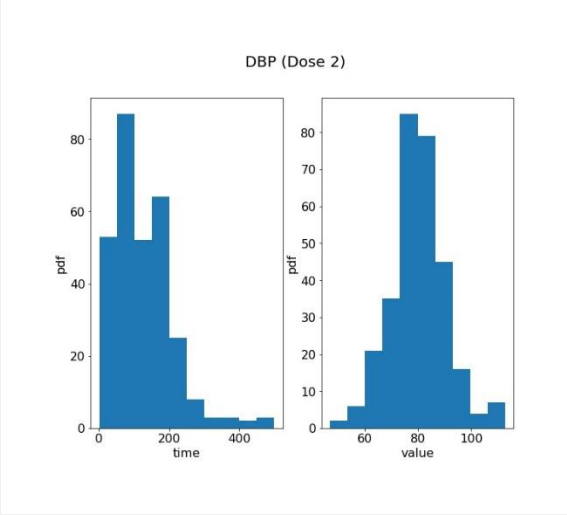


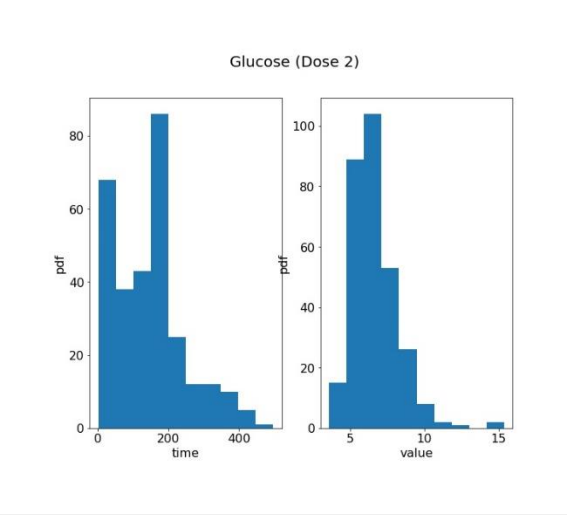
Insights: Demographic Information

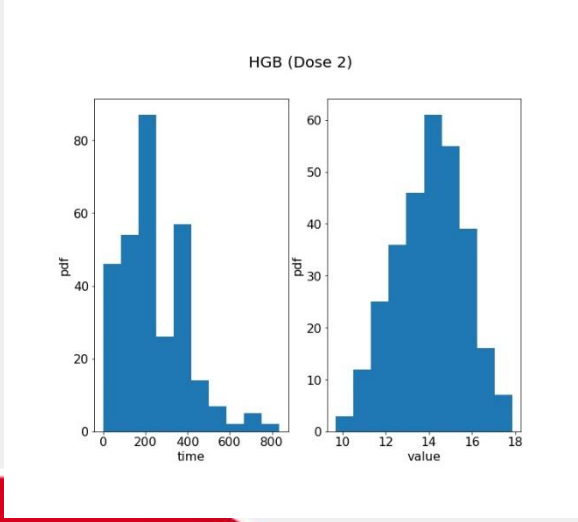
- There are four different types of neural networks used in making the predictive model. They are neural networks based on logistic regression, decision tree, MLP classifer and Random Forest.
- The neural network based on Random Forest performed the best.
- The neural network based on Logistic Regression performed the worst with no predictions of patients progressed into CKD in the test set.
- However, the overall accuracy of all the neural networks is poor and could be made better.
- The relative information of the feature age is the highest and could be useful if somebody would like to identify the important features and incorporate them in neural network.

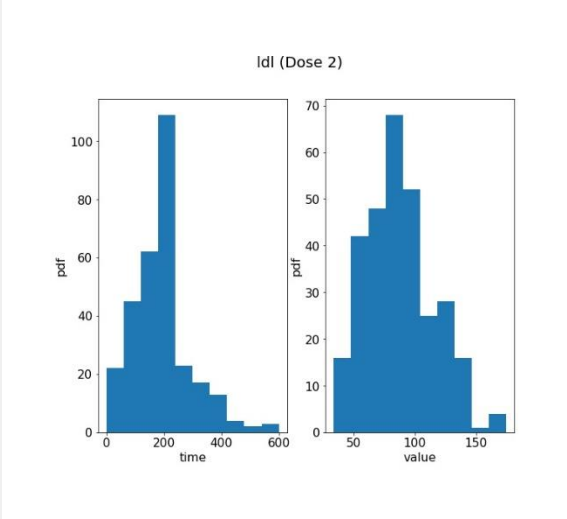
Analysis: Lab measurements including time in dataset

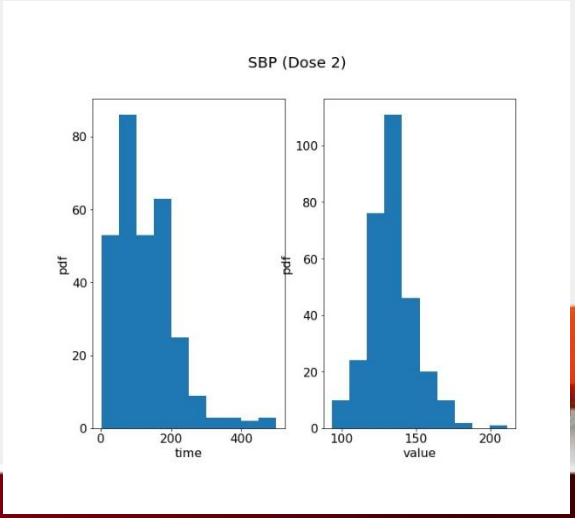




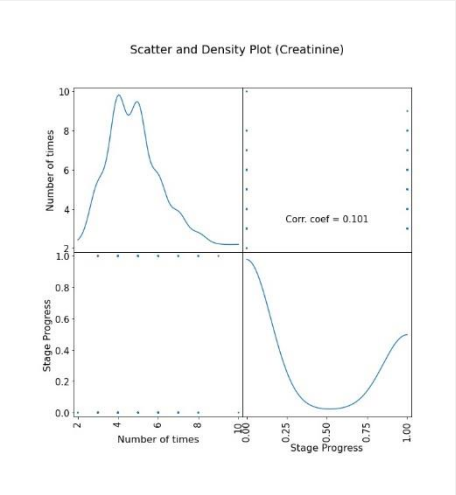


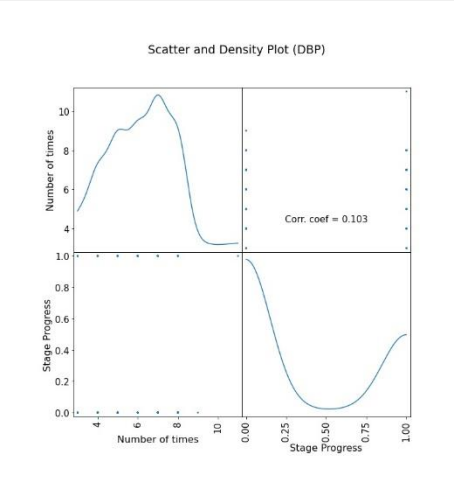


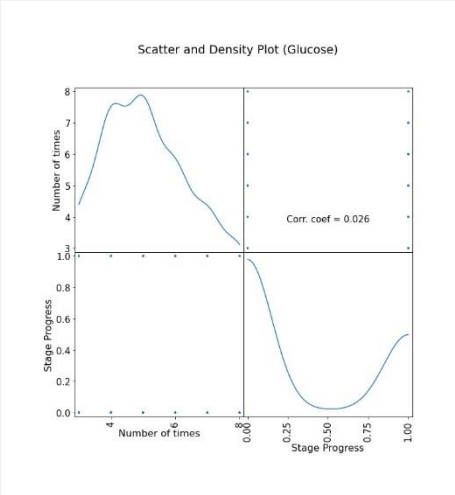


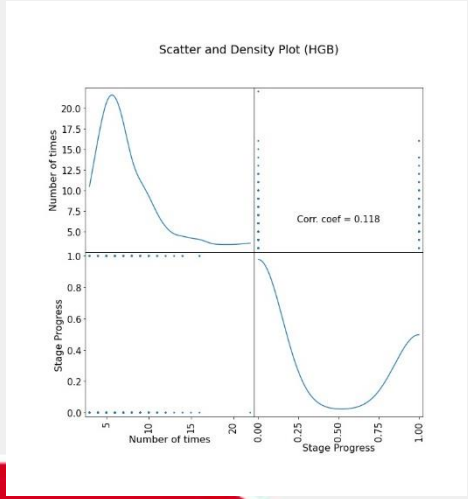


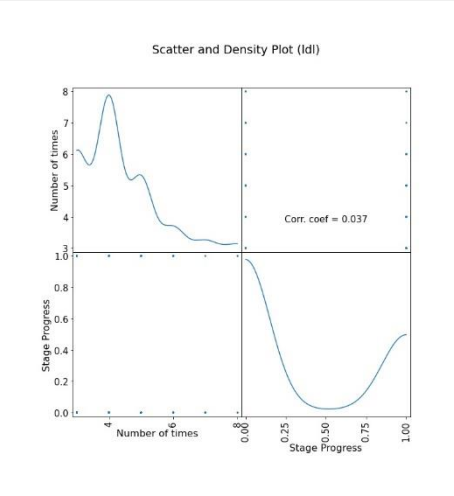
Analysis: Lab measurements including time in dataset

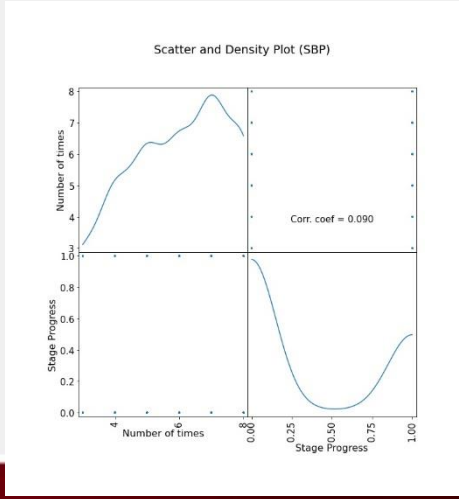












Insights: Analysis

- The distribution of values of lab measurements are pretty gaussian. The peaks for time values is shifted to the left of the mid-value of the range of time in the dataset. These observations are for the second dose given to the patients.
- The number of times a measurement done on a patient is fairly uncorrelated with his/her progress in CKD.
- The self correlation first increases and then decreases for all the lab measurements besides SBP where it increases for a fair number of doses. The implication of this observation isn't clear.
- The correlation coefficient is small for all the lab measurements. They are in the range between 0.0 to 0.15.
- There are some patients on whom large number of times lab measurements have been performed and they have either progressed in CKD or not. For example, patients with more performance of DBP has progressed in CKD and patients with more performance of HGB has not progressed in CKD. These observations maybe useful in understanding the cure of CKD.

Predictive Model: Lab measurements including time in dataset

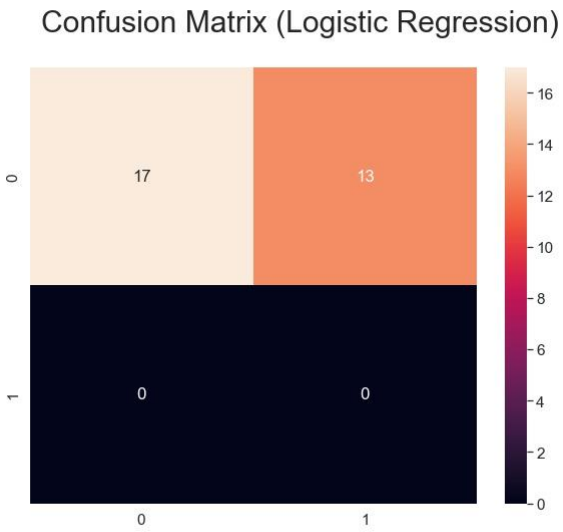
- Neural Network based on Logistic Regression

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_5 (Dense)              (None, 1000)              135000
_________________________________________________________________
dense_6 (Dense)              (None, 1000)              1001000
_________________________________________________________________
dense_7 (Dense)              (None, 1000)              1001000
_________________________________________________________________
dense_8 (Dense)              (None, 1)                 1001
=================================================================
Total params: 2,138,001
Trainable params: 2,138,001
Non-trainable params: 0
_________________________________________________________________
```

Results: Lab measurements including time in dataset

- Logistic Regression

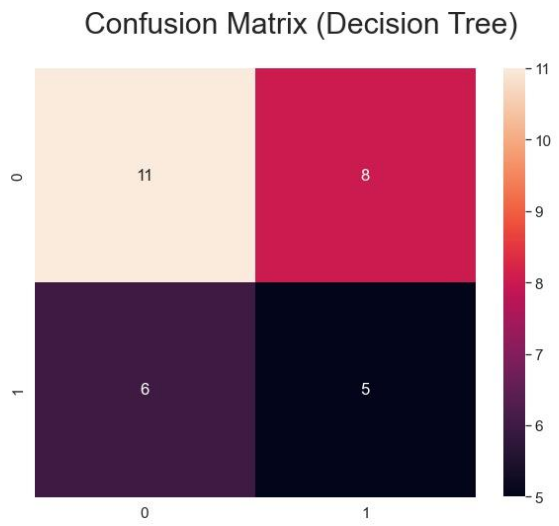
```
-----------------------------------------------------------------------
Accuracy of Logistic Regression Classifier : 0.5666666666666667
-----------------------------------------------------------------------
Classification report :
               precision    recall  f1-score   support

           1       0.00      0.00      0.00        13
           0       0.57      1.00      0.72        17

    accuracy                           0.57        30
   macro avg       0.28      0.50      0.36        30
weighted avg       0.32      0.57      0.41        30

-----------------------------------------------------------------------
```



Results: Lab measurements including time in dataset

- Decision Tree

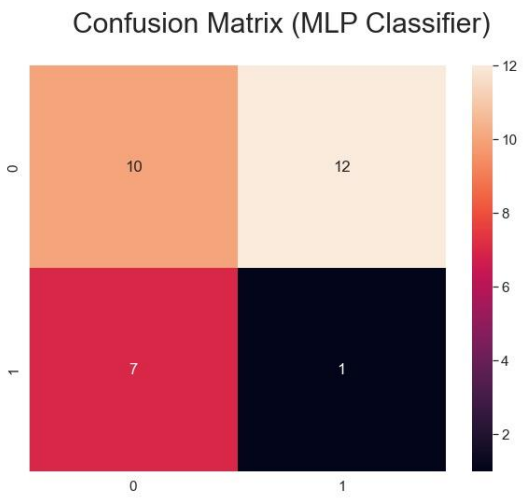
```
-----------------------------------------------------------
Accuracy of Decision Tree Classifier : 0.5333333333333333
-----------------------------------------------------------
Classification report :
               precision    recall  f1-score   support

           1       0.45      0.38      0.42        13
           0       0.58      0.65      0.61        17

    accuracy                           0.53        30
   macro avg       0.52      0.52      0.51        30
weighted avg       0.53      0.53      0.53        30

-----------------------------------------------------------
```



Results: Lab measurements including time in dataset

- MLP Classifier

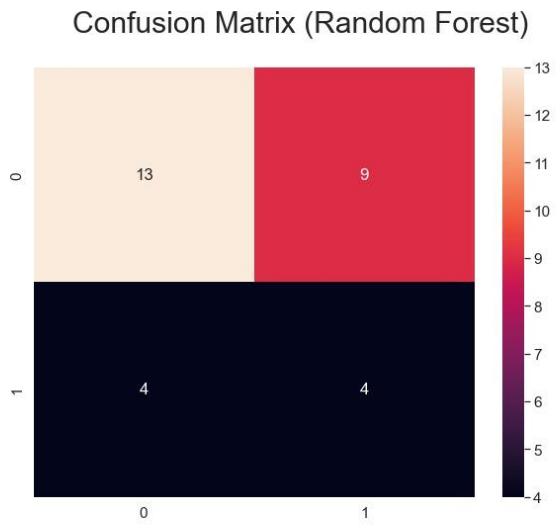
```
-------------------------------------------------------------
Accuracy of MLPClassifier : 0.36666666666666664
-------------------------------------------------------------
Classification report :
              precision    recall  f1-score   support

           1       0.12      0.08      0.10        13
           0       0.45      0.59      0.51        17

    accuracy                           0.37        30
   macro avg       0.29      0.33      0.30        30
weighted avg       0.31      0.37      0.33        30

-------------------------------------------------------------
```



Results: Lab measurements including time in dataset

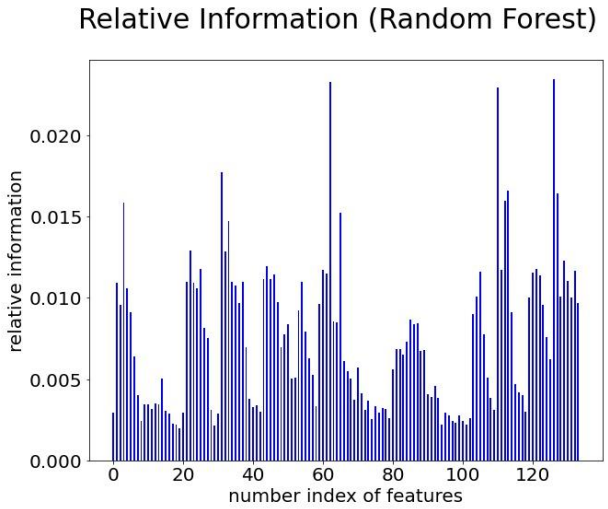
- Random Forest

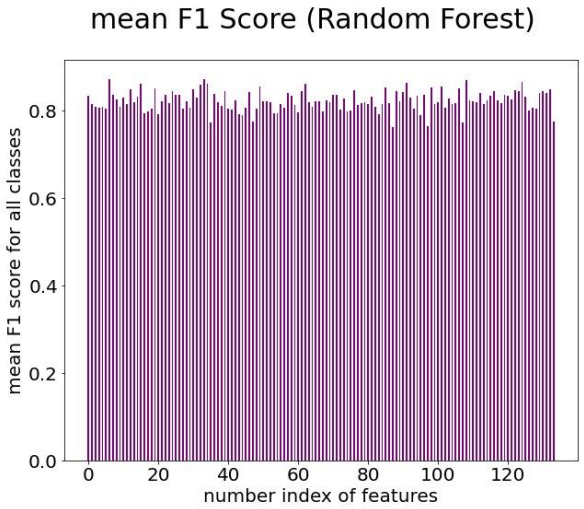
```
-----------------------------------------------------------
Accuracy of Random Forest Classifier :  0.5666666666666667
-----------------------------------------------------------
Classification report :
               precision    recall  f1-score   support

           1       0.50      0.31      0.38        13
           0       0.59      0.76      0.67        17

    accuracy                           0.57        30
   macro avg       0.55      0.54      0.52        30
weighted avg       0.55      0.57      0.54        30

-----------------------------------------------------------
```



Further Results: Lab measurements including time in dataset

- Random Forest





Insights: Lab measurements including time in dataset

- There are four different types of neural networks used in making the predictive model. They are neural networks based on logistic regression, decision tree, MLP classifer and Random Forest.
- The neural network based on Decision Tree performed the best.
- The neural network based on Logistic Regression performed the worst with no predictions of patients progressed into CKD in the test set.
- However, the overall accuracy of all the neural networks is poor and could be made better.
- The relative information of the features isn’t able to capture the important features so to incorporate them in neural network.

Predictive Model: Using RNN and LSTM

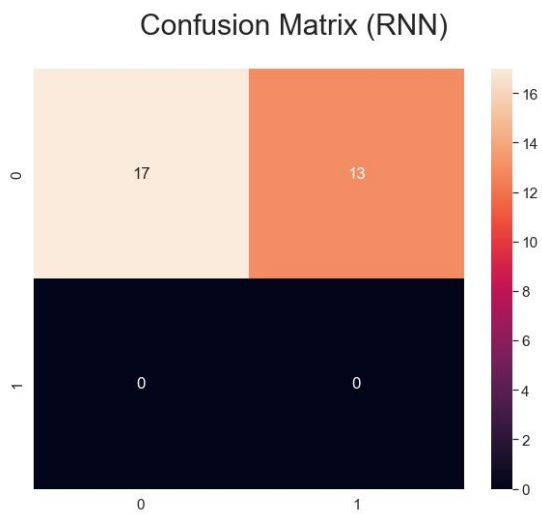
- Neural Network based on RNN and LSTM

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
simple_rnn_3 (SimpleRNN)     (None, 134, 1)            3
_________________________________________________________________
simple_rnn_4 (SimpleRNN)     (None, 134, 1)            3
_________________________________________________________________
lstm_2 (LSTM)                (None, 100)               40800
_________________________________________________________________
dense_3 (Dense)              (None, 1000)              101000
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 1001
=================================================================
Total params: 142,807
Trainable params: 142,807
Non-trainable params: 0
_________________________________________________________________
```

Results: Using RNN and LSTM

- RNN and LSTM

```
-----------------------------------------------------------------
Accuracy of RNN Classifier : 0.5666666666666667
-----------------------------------------------------------------
Classification report :
              precision    recall  f1-score   support

           1       0.00      0.00      0.00        13
           0       0.57      1.00      0.72        17

    accuracy                           0.57        30
   macro avg       0.28      0.50      0.36        30
weighted avg       0.32      0.57      0.41        30

-----------------------------------------------------------------
```



Predictive Model: Using CWRNN

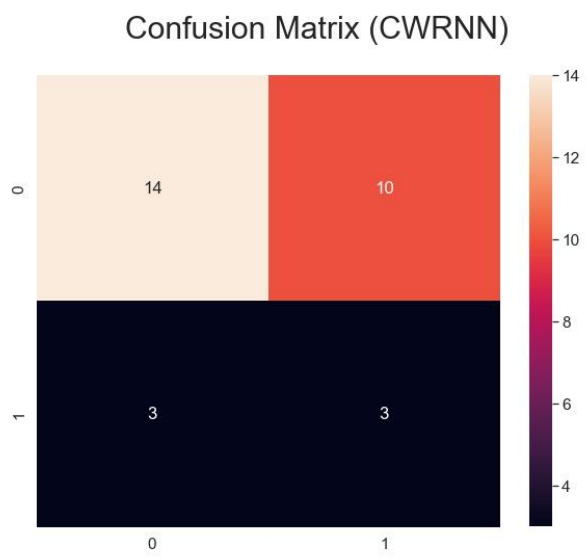
- Neural Network based on CWRNN

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
clockwork_simple_rnn_1 (Cloc (None, 134)               4814
_________________________________________________________________
dense_2 (Dense)              (None, 1000)              135000
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 1001
=================================================================
Total params: 140,815
Trainable params: 140,815
Non-trainable params: 0
_________________________________________________________________
```

Results: Using CWRNN

- CWRNN

```
-----------------------------------------------------------------
Accuracy of CWRNN Classifier : 0.5666666666666667
-----------------------------------------------------------------
Classification report :
               precision    recall  f1-score   support

           1       0.50      0.23      0.32        13
           0       0.58      0.82      0.68        17

    accuracy                           0.57        30
   macro avg       0.54      0.53      0.50        30
weighted avg       0.55      0.57      0.52        30

-----------------------------------------------------------------
```



Predictive Model: Lab Measurement placed at the timeth entry

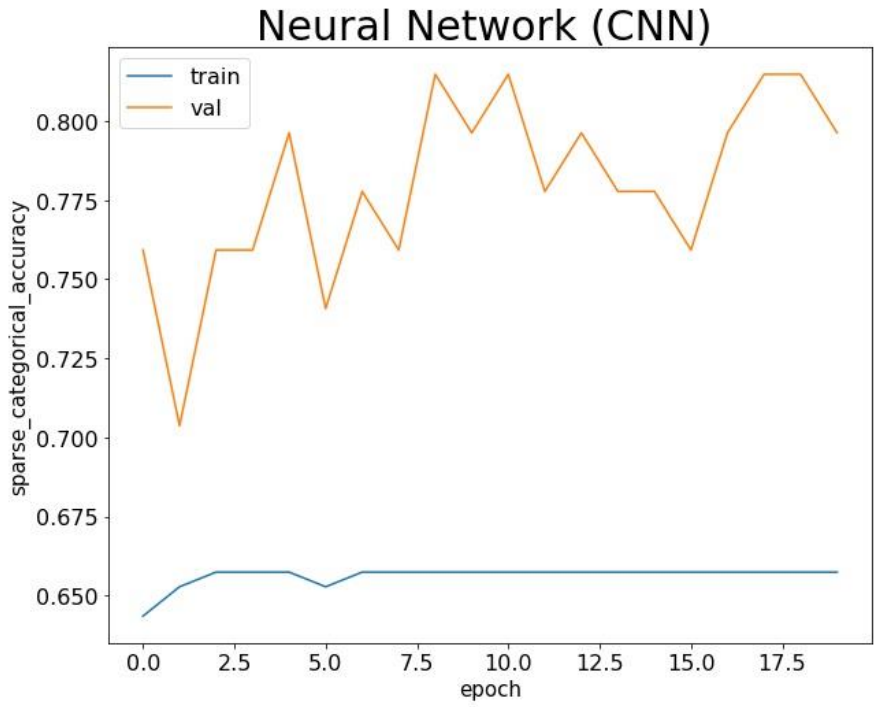
- Neural Network based on CNN

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         (None, 3632, 1)           0
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 3632, 64)          256
_________________________________________________________________
batch_normalization_4 (Batch (None, 3632, 64)          256
_________________________________________________________________
re_lu_4 (ReLU)               (None, 3632, 64)          0
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 3632, 64)          12352
_________________________________________________________________
batch_normalization_5 (Batch (None, 3632, 64)          256
_________________________________________________________________
re_lu_5 (ReLU)               (None, 3632, 64)          0
_________________________________________________________________
conv1d_6 (Conv1D)            (None, 3632, 64)          12352
_________________________________________________________________
batch_normalization_6 (Batch (None, 3632, 64)          256
_________________________________________________________________
re_lu_6 (ReLU)               (None, 3632, 64)          0
_________________________________________________________________
global_average_pooling1d_2 ( (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 130
=================================================================
Total params: 25,858
Trainable params: 25,474
Non-trainable params: 384
_________________________________________________________________
```

Results: Lab Measurement placed at the timeth entry

- CNN

```
-----------------------------------------------------------
Accuracy from CNN Classifier :  0.4333333373069763
-----------------------------------------------------------
```



Insights: Lab Measurements

- There are three different types of neural networks used in making the predictive model by using the time information of the lab measurements taken. They are neural networks based on RNN and LSTM, CWRNN, CNN.
- For RNN and LSTM and CWRNN, the time is added as additional variables in the input along with the lab measurements whereas for CNN, the position of lab measurements is decided by the time it got conducted in the matrix.
- The neural network based on CWRNN performed the best.
- The neural network based on RNN and LSTM performed the worst with no predictions of patients progressed into CKD in the test set.
- However, the overall accuracy of all the neural networks is poor and could be made better.
- Moreover, the validation accuracy is greater than training accuracy for CNN network implying no sign of overfitting.

Predictive Model: Medications, daily dosage, prescriptions start and end day

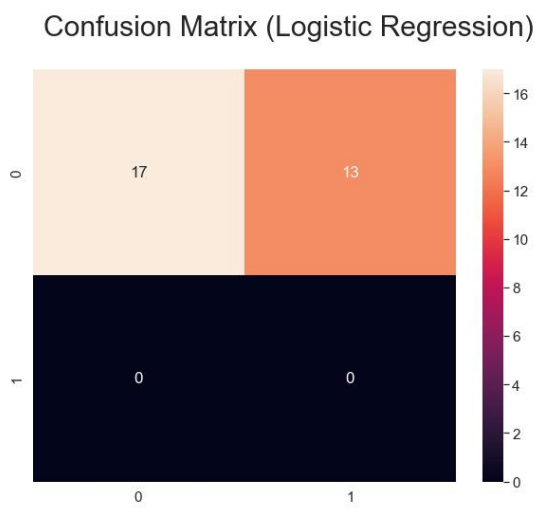
- Neural Network based on Logistic Regression

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 1000)              649000
_________________________________________________________________
dense_2 (Dense)              (None, 1000)              1001000
_________________________________________________________________
dense_3 (Dense)              (None, 1000)              1001000
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 1001
=================================================================
Total params: 2,652,001
Trainable params: 2,652,001
Non-trainable params: 0
_________________________________________________________________
```

Results: Medications, daily dosage, prescriptions start and end day

- Logistic Regression

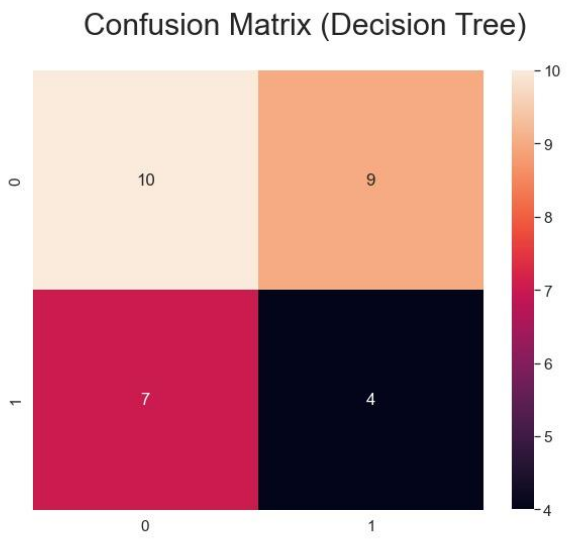
```
-----------------------------------------------------------------------
Accuracy of Logistic Regression Classifier : 0.5666666666666667
-----------------------------------------------------------------------
Classification report :
              precision    recall  f1-score   support

           1       0.00      0.00      0.00        13
           0       0.57      1.00      0.72        17

    accuracy                           0.57        30
   macro avg       0.28      0.50      0.36        30
weighted avg       0.32      0.57      0.41        30

-----------------------------------------------------------------------
```



Results: Medications, daily dosage, prescriptions start and end day

- Decision Tree

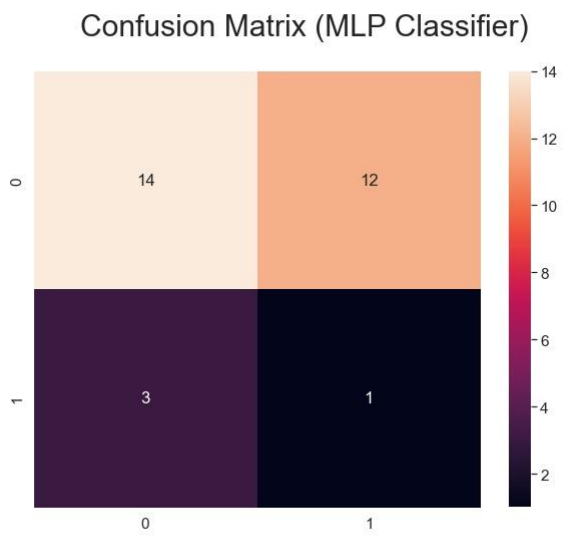
```
-----------------------------------------------------------
Accuracy of Decision Tree Classifier : 0.4666666666666667
-----------------------------------------------------------
Classification report :
               precision    recall  f1-score   support

           1       0.36      0.31      0.33        13
           0       0.53      0.59      0.56        17

    accuracy                           0.47        30
   macro avg       0.44      0.45      0.44        30
weighted avg       0.46      0.47      0.46        30

-----------------------------------------------------------
```



Results: Medications, daily dosage, prescriptions start and end day

- MLP Classifier

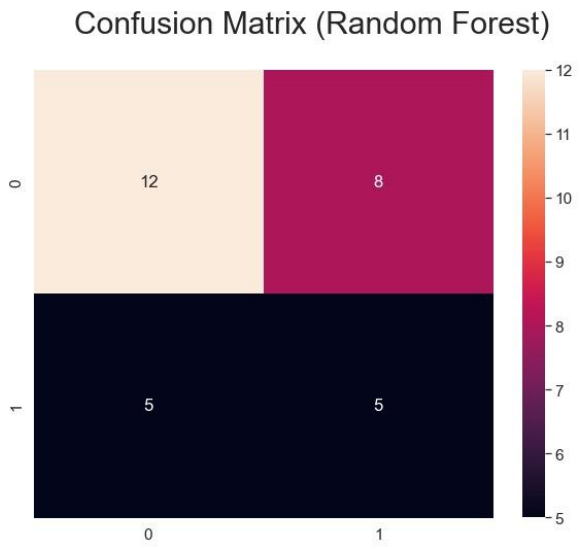
```
-------------------------------------------------------------
Accuracy of MLPClassifier : 0.5
-------------------------------------------------------------
Classification report :
              precision    recall  f1-score   support

           1       0.25      0.08      0.12        13
           0       0.54      0.82      0.65        17

    accuracy                           0.50        30
   macro avg       0.39      0.45      0.38        30
weighted avg       0.41      0.50      0.42        30

-------------------------------------------------------------
```



Results: Medications, daily dosage, prescriptions start and end day

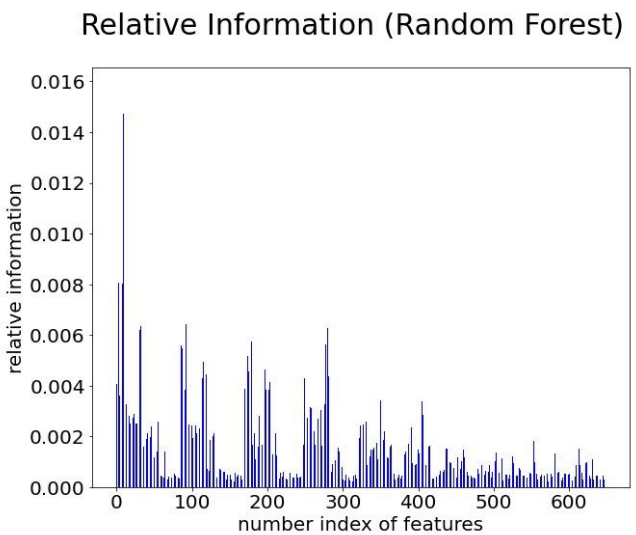
- Random Forest

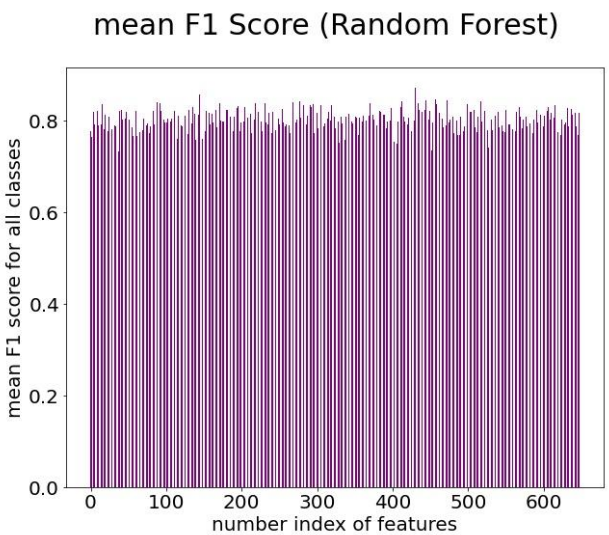
```
-----------------------------------------------------------
Accuracy of Random Forest Classifier :  0.5666666666666667
-----------------------------------------------------------
Classification report :
              precision    recall  f1-score   support

           1       0.50      0.38      0.43        13
           0       0.60      0.71      0.65        17

    accuracy                           0.57        30
   macro avg       0.55      0.55      0.54        30
weighted avg       0.56      0.57      0.56        30

-----------------------------------------------------------
```



Further Results: Medications, daily dosage, prescriptions start and end day

- Random Forest





Insights: Medications, daily dosage, prescriptions start and end day

- There are four different types of neural networks used in making the predictive model for the medication dataset. They are neural networks based on logistic regression, decision tree, MLP classifer and Random Forest.
- The neural network based on Random Forest performed the best.
- The neural network based on Logistic Regression performed the worst with no predictions of patients progressed into CKD in the test set.
- However, the overall accuracy of all the neural networks is poor and could be made better.
- The relative information of the features isn't able to capture the important features so to incorporate them in neural network. However, it does show periodic increase and fall of relative importances with index of the features. The reason for this observation isn't clear.

Predictive Model: Using RNN and LSTM (Meds)

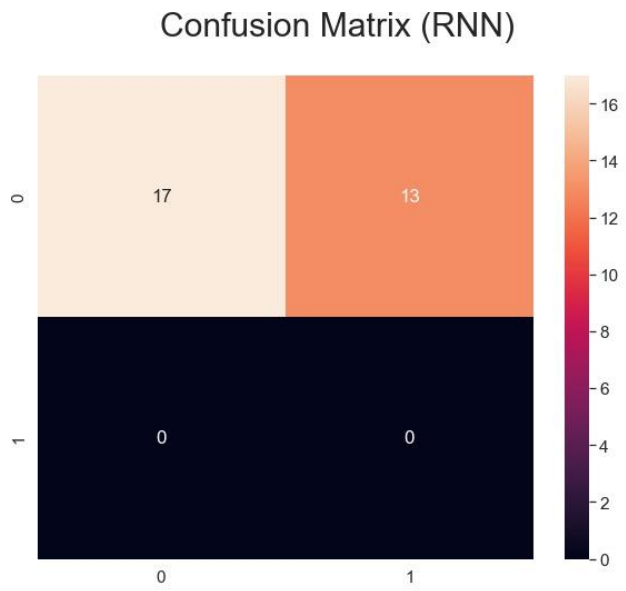
- Neural Network based on RNN and LSTM

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
simple_rnn_1 (SimpleRNN)     (None, 648, 1)            3
_________________________________________________________________
simple_rnn_2 (SimpleRNN)     (None, 648, 1)            3
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)               40800
_________________________________________________________________
dense_1 (Dense)              (None, 1000)              101000
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 1001
=================================================================
Total params: 142,807
Trainable params: 142,807
Non-trainable params: 0
_________________________________________________________________
```

Results: Using RNN and LSTM (Meds)

- RNN and LSTM

```
-----------------------------------------------------------------
Accuracy of RNN Classifier : 0.5666666666666667
-----------------------------------------------------------------
Classification report :
               precision    recall  f1-score   support

           1       0.00      0.00      0.00        13
           0       0.57      1.00      0.72        17

    accuracy                           0.57        30
   macro avg       0.28      0.50      0.36        30
weighted avg       0.32      0.57      0.41        30

-----------------------------------------------------------------
```



Predictive Model: Using CWRNN (Meds)

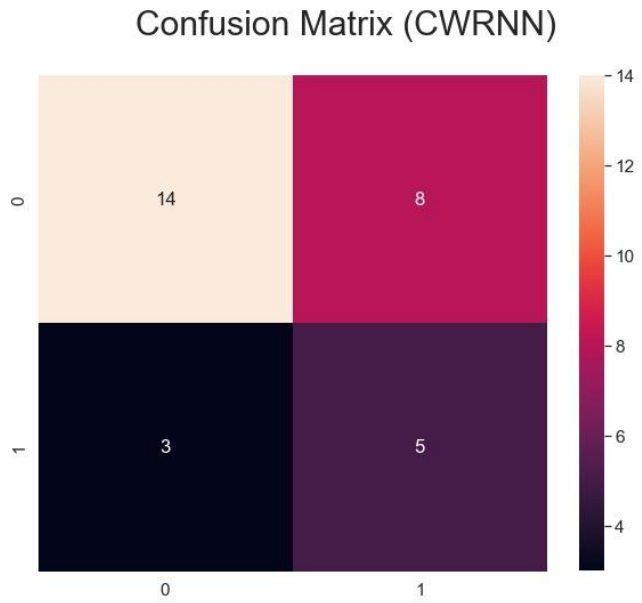
- Neural Network based on CWRNN

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
clockwork_simple_rnn_2 (Cloc (None, 648)               20748
_________________________________________________________________
dense_7 (Dense)              (None, 1000)              649000
_________________________________________________________________
dense_8 (Dense)              (None, 1)                 1001
=================================================================
Total params: 670,749
Trainable params: 670,749
Non-trainable params: 0
_________________________________________________________________
```

Results: Using CWRNN (Meds)

- CWRNN

```
-----------------------------------------------------------------
Accuracy of CWRNN Classifier : 0.6333333333333333
-----------------------------------------------------------------
Classification report :
              precision    recall  f1-score   support

           1       0.62      0.38      0.48        13
           0       0.64      0.82      0.72        17

    accuracy                           0.63        30
   macro avg       0.63      0.60      0.60        30
weighted avg       0.63      0.63      0.61        30

-----------------------------------------------------------------
```



Predictive Model: Using CNN (Meds)

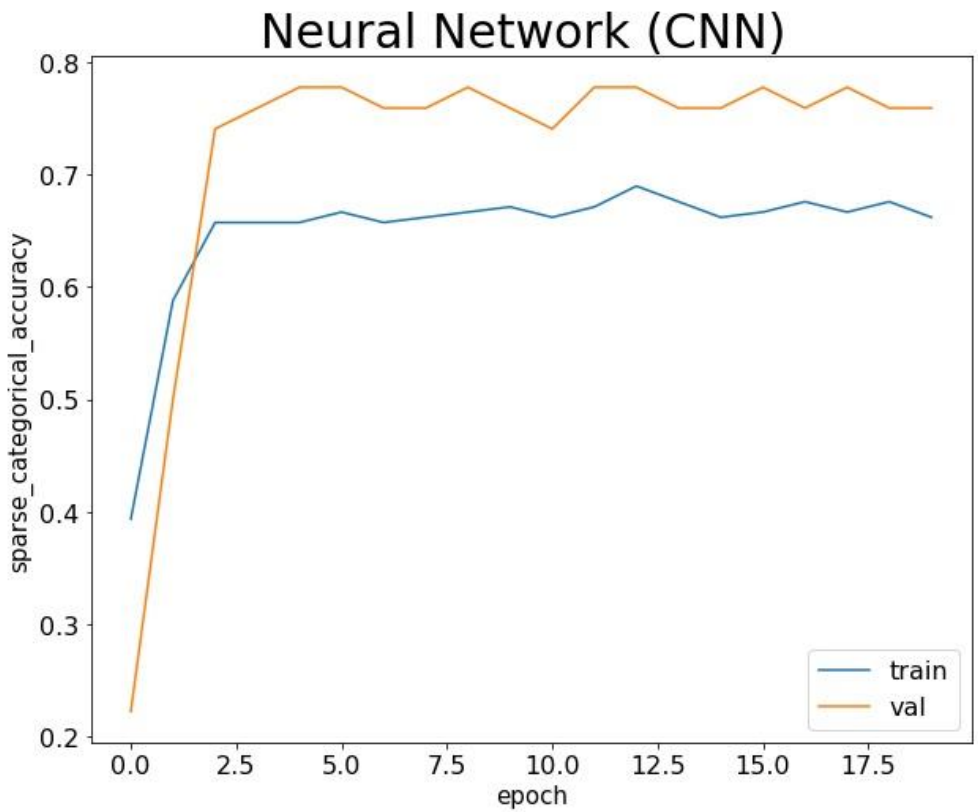
- Neural Network based on CNN

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         (None, 648, 1)            0
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 648, 64)           256
_________________________________________________________________
batch_normalization_4 (Batch (None, 648, 64)           256
_________________________________________________________________
re_lu_4 (ReLU)               (None, 648, 64)           0
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 648, 64)           12352
_________________________________________________________________
batch_normalization_5 (Batch (None, 648, 64)           256
_________________________________________________________________
re_lu_5 (ReLU)               (None, 648, 64)           0
_________________________________________________________________
conv1d_6 (Conv1D)            (None, 648, 64)           12352
_________________________________________________________________
batch_normalization_6 (Batch (None, 648, 64)           256
_________________________________________________________________
re_lu_6 (ReLU)               (None, 648, 64)           0
_________________________________________________________________
global_average_pooling1d_2 ( (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 130
=================================================================
Total params: 25,858
Trainable params: 25,474
Non-trainable params: 384
_________________________________________________________________
```

Results: Using CNN (Meds)

- CNN

```
-----------------------------------------------------------
Accuracy from CNN Classifier :  0.6000000238418579
-----------------------------------------------------------
```



Insights: Medications

- There are three different types of neural networks used in making the predictive model by using the time information of the medications taken. They are neural networks based on RNN and LSTM, CWRNN, CNN.
- For all of them, the time of the starting and ending day is added as additional variables in the input along with the medications and its daily dosage.
- The neural network based on CWRNN and CNN performed the best.
- The neural network based on RNN and LSTM performed the worst with no predictions of patients progressed into CKD in the test set.
- However, the overall accuracy of all the neural networks is poor and could be made better.
- Moreover, the validation accuracy is greater than training accuracy for CNN network implying no sign of overfitting.