**IFN 509: Data Exploration and Mining**

**Assessment 2**

**Team Name: [ Group12 ]**

**Group No. [ 12 ]**

|  |  |
| --- | --- |
| **Student Name** | **Student Id** |
| **Qiaojing Luo** | **N10635173** |
| **Xian Chen** | **N10159771** |
| **Xiulun An** | **N10159894** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Student 1** | **Student 2** | **Student 3** |
| **Student 1** | **<100%>** | **<100 %>** | **< 100%>** |
| **Student 2** | **<100 %>** | **<100 %>** | **<100 %>** |
| **Student 3** | **<100 %>** | **<100 %>** | **<100 %>** |

**Project (a):**

**Association mining to find hotspots based on a Patient Route Data**

**1. What pre-processing was required on the dataset before building the association mining model? What variables did you include in the analysis? Justify your choice.**

**Solution:** Before building the association mining model, it is required

We selected the ‘location’ and ‘patient\_id’ variables in the analysis, because we need to find the connection between the patient and the location they have been. Other variables are not helpful for our analysis, so we choose to delete global\_num, date, latitude and longitude. Below is the code.



**2. Conduct association mining and answer the following:**

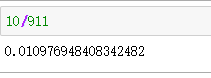
**a.What ‘min\_support’ and `min\_confidence’ thresholds were set for this mining exercise? Rationale why these values were chosen?**

First, we group by patient\_id, then list all routes.



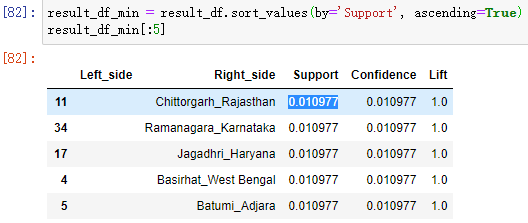
Totally 911 routes.

Then we want routes that at least appear 10 times.

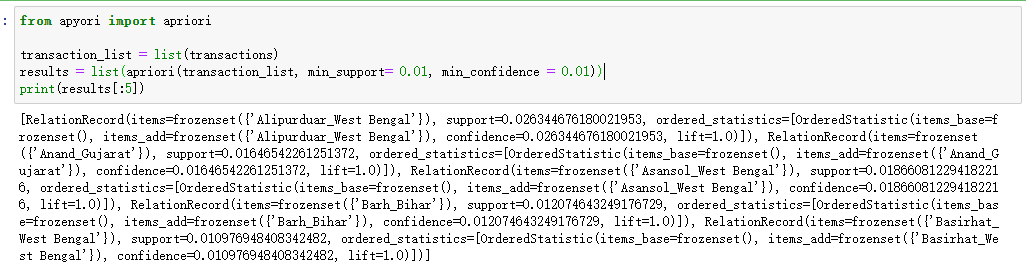


Therefore, we set min\_support as 0.01.

Next, we found the lowest support of all single locations is 0.01977, which means that every location at least appears 10 times in all routes.

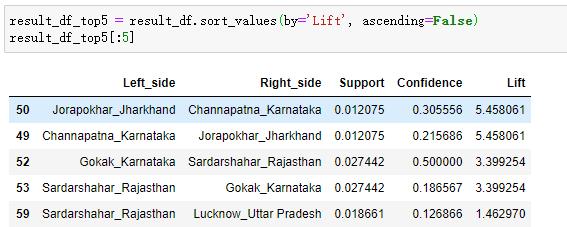
****

We want the min\_confidence higher than all single locations’ support. Therefore, we set the min\_confidence as 0.01 also.

****

**b.Report the top-5 rules and interpret them**

Because we want to find the top 5 highly correlated locations, we sorted the results by “Lift”.

****

From the above tests we foundthree highly correlated locations rules:

1: Town Jorapokhar in Jharkhand state ⇋ Town Channapatna in Karnataka state

30% of patients who visited Jorapokhar also visited Channapatna. This usually happens in 1.2% of all routines.

2: Town Gokak in Karnataka state ⇋ Town Sardarshahar in Rajasthan state

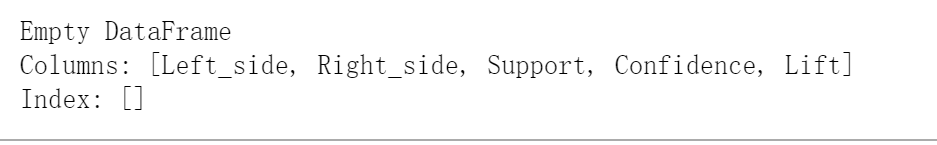
50% of patients who visited Gokak also visited Sardarshahar .This usually happens in 2.7% of all routines.

3: Town Sardarshahar in Rajasthan state ⇋ Town Lucknow in Uttar Pradesh state

12% of patients who visited Sardarshahar also visited Lucknow.This usually happens in 1.8% of all routines.

**3. Identify top-5 common routes that COVID-19 positive patients from the town Ranebennur in Karnataka state have travelled.**

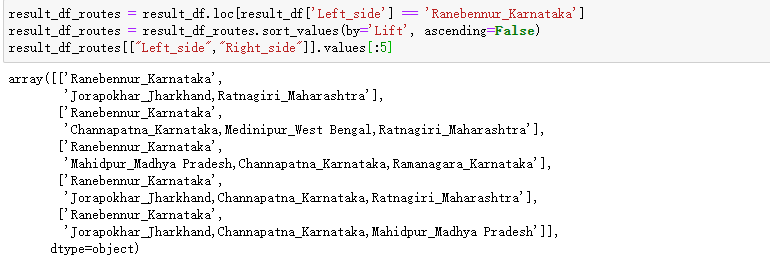
If we set min\_support and min\_confidence to 0.01, it shows the empty dataframe.

****

So we need to use the smaller min\_support to find the top-5 common routes, and we set it as 0.001as shown below.

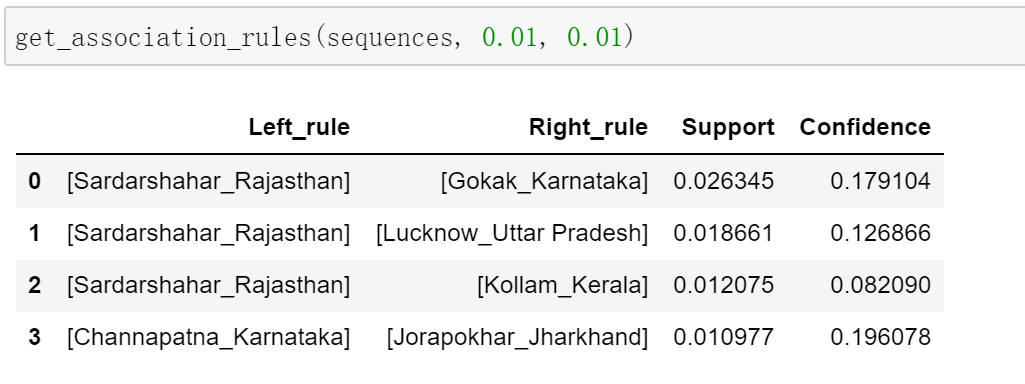


We sort the top-5 common routesby lift.

****

**4. Can you perform sequence analysis on this dataset? If yes, present your results. If not, rationalize why.**

Yes, set min\_sup, min\_confidence as 0.01 and 0.01.

****

**5. How can the outcome of this study be used by the relevant decision-makers?**

From the association test above, we can identify three pair correlated regions:

* Town Jorapokhar in Jharkhand state ⇋ Town Channapatna in Karnataka state
* Town Gokak in Karnataka state ⇋ Town Sardarshahar in Rajasthan state
* Town Sardarshahar in Rajasthan state ⇋ Town Lucknow in Uttar Pradesh state

Therefore, it is advised that the government should share the covid-19 information with the correlated region. This may be extremely helpful for tracking patients’ routines and preventing regional outbreak of infection.

**Project (b):**

**Clustering Diabetes data**

**1. What pre-processing was required on the dataset (D2.csv) before building the clustering model?**

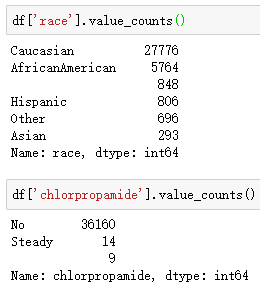
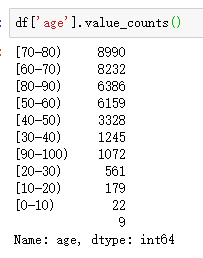
****

From the .info() output, we found the **admission\_type\_id,discharge\_disposition\_id and admission\_source\_id** variable types are set incorrectly. The output lists set these variables as integers, while based on the description, these variables should be categorical variables.

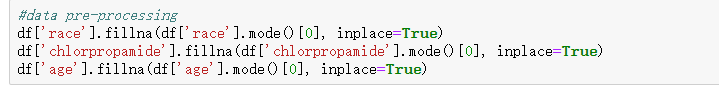
Thus, we changed incorrect data types.



Next we applied .value\_counts() to get more information, and found some empty values exist in **age, race** and **chlorpropamide** columns.



Filling missing values with mode.



**2. Build a clustering model to profile the characteristics of diabetic patients. Answer the followings:**

**a. What clustering algorithm have you used?**

We run the K Prototypes clustering since we have mixed variables types.

Two categorical variables: **A1C result** and **age** as attributes.

Because **A1C results** reflect patients’ average blood sugar level for the past two to three months which can evaluate the blood glucose control in diabetic patients. And **age** is a commonly used categorical variable

Three numerical variables:**'num\_medications', 'number\_diagnoses' and 'time\_in\_hospital'**.

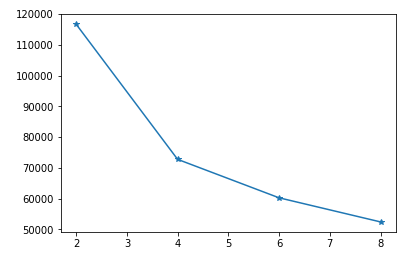
These three variables are useful for assessing the severity of diabetic patients.

**b. List the attributes used in this analysis.**

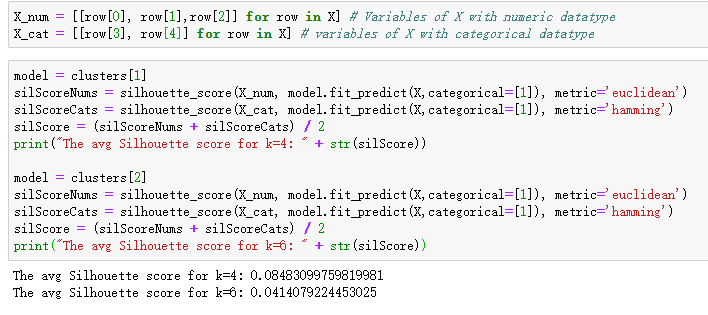
We want to classify patients according to the severity of diabetes. Thus, we used three numerical variables including **'num\_medications', 'number\_diagnoses', 'time\_in\_hospital'**.And two categorical variables including **A1C result** and **age**.

**c. What is the optimal number of clusters identified? How did you reach this optimal number?**

To get the optimal number of clusters we first used the elbow plot to find the turning point.



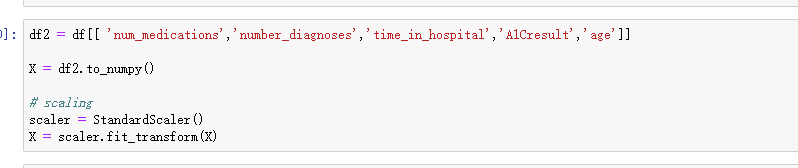
From the above plot, we see that the turning points appear in k = 4 and k = 6. Next, we use the silhouette score to decide which is the best parameter.



The optimal number is 4, the silhouette score of 4 is 0.08, which is higher than the score of 6.

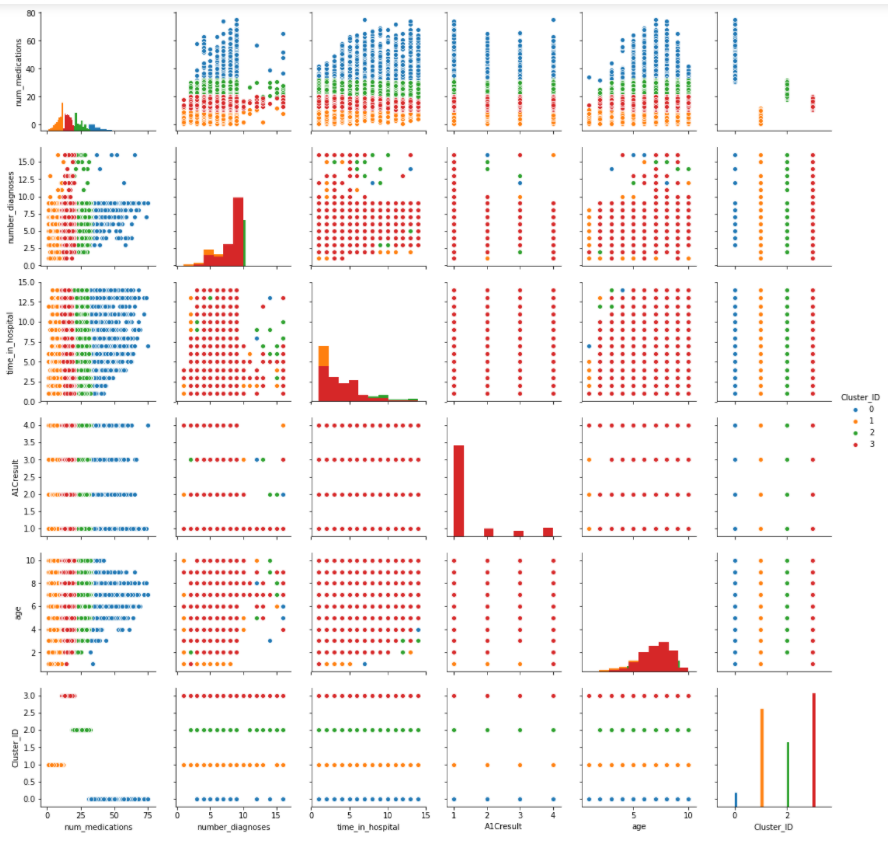
**d. Did you normalise the variables? What was its effect on the model?**

Yes, we use it to adjust all of the values measured on different scales to a notionally common scale.

****

For comparison we also generated clustering without normalization.

1: Not normalization:



2: Normalized



**Does the variable normalization process enable a better clustering solution?**

Yes, the normalization process has a great effect on the model, because clusters are determined by the distance between points in mathematical space, this importance is especially great in cluster analysis. By comparing these above pair plots we found that not normalized clustering is only sensitive in the wider range variables like **'num\_medications'**, where clustering after normalization can interpret all variables well.

**3. For the model with the optimal number of clusters,**

**a. Visualize the clusters using ‘pairplot’ and interpret the visualization.**



The ‘pairplot’ shows us how different cluster members have different value distributions on different variables. Here is how to interpret the plots:

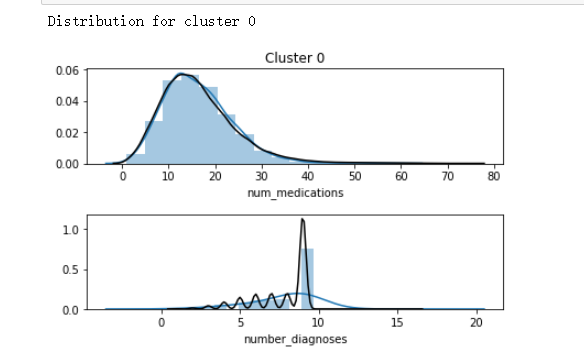
1. Looking at the **num\_medications** and **time\_in\_hospital** (first row, third column), we could see Cluster 2 has a large number of medications and large diagnoses, while cluster 3 took less medications and had less diagnoses .
2. For **age** and **A1C result**  (fifth row, fourth column), pairplot shows cluster 3 is elderly patients with lower blood sugar level, while cluster 0 is young patients with high blood sugar level.

The visualisation helps us to profile the clusters as follow:

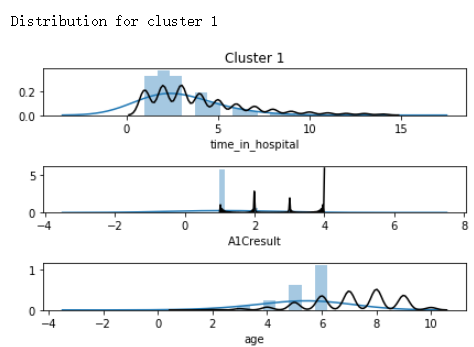
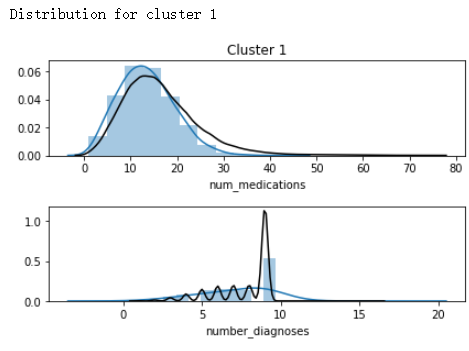
* Cluster0: high blood sugar level and low number of medications taken.
* Cluster1: youth patients with a lower number of medications taken and diagnoses.
* Cluster2: patients stay a long period in hospital.
* Cluster3: elderly patients with low blood sugar level.

**b. Characterize the nature of each cluster by giving it a descriptive label and a brief description. Hint: use cluster distribution.**

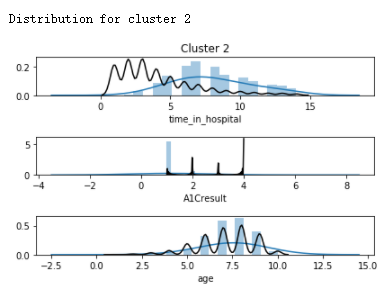
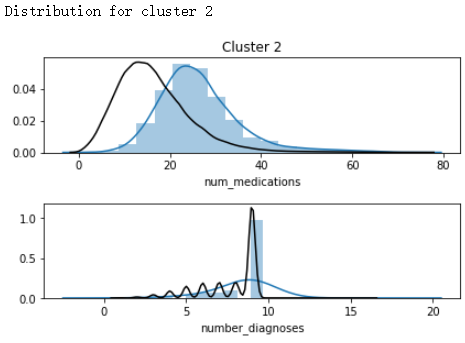
Here, we plot the distributions of cluster 0, cluster 1, cluster 2 and cluster 3 against the distributions from all data. The black lines are the distributions from all records, while light-blue lines are for a specific cluster. These plots show us the key characteristics of the clusters, as follows:

****

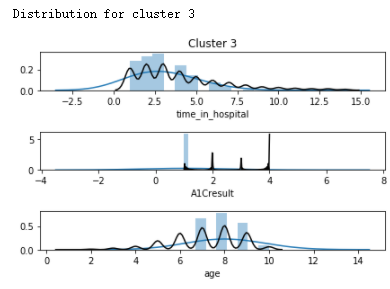
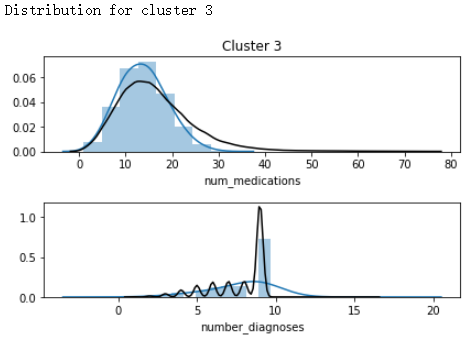
**Cluster0**: All attributes similar to the population.



**Cluster1:** Left leaning **num\_medications**  and **age**. Cluster 1 patients are low blood sugar level,young and taken slightly less medications.

****

**Cluster2:** Right leaning **time\_in\_hospital** and **num\_medications.** Cluster 2 patients are low blood sugar level, stay a long period of time in hospital and take more medications.

****

**Cluster3:** Left leaning **A1C result** and **num\_medications.** Cluster 3 are patients who have low blood sugar level and take less medication.

**4. Build another clustering model using an algorithm that helps to profile the patients of specific races including Asian and Caucasian. Use the best setting (e.g., variable normalisations, optimal K, etc) obtained in the previous model. Answer the following:**

**a. List the attributes used in this analysis.**

In order to profile the patients of Asian and Caucsian, we firstly fiter Asian and Caucasian. Next, we used the same attributes as the previous model including **'num\_medications', 'number\_diagnoses', 'time\_in\_hospital'**. And two categorical variables **A1C result** and **age**.

**b. What difference do you see in this clustering interpretation when compared to the previous one?**

****

We did not see any difference in these two clusterings. This may result from the majority of patients are Caucasian.



Thus, although we filter other races out, we can not get a different result. If we want to profile other races we need more samples.

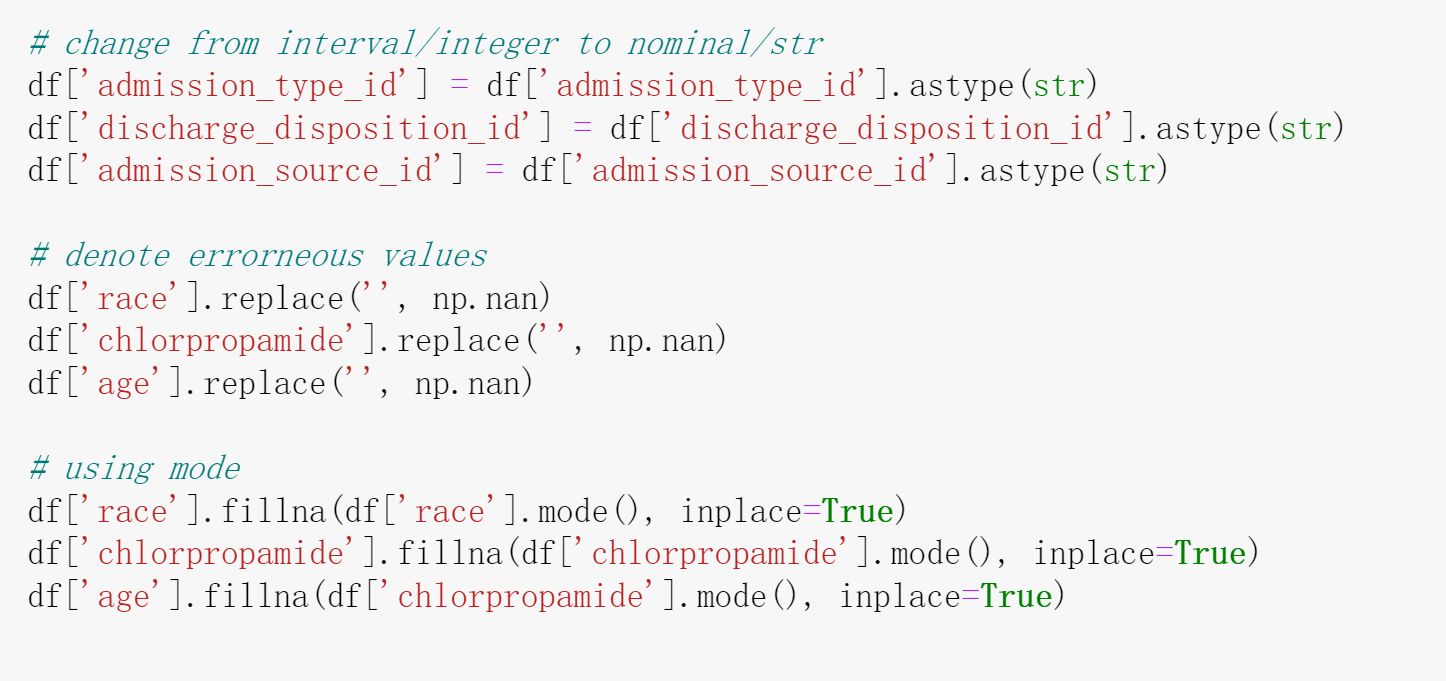
**5. How can the outcome of this study be used by the relevant decision-makers?**

Doctors can make treatment strategies for patients in different clusters based on the clustering results. In addition, hospitals can identify the type of diabetes and make improvements on this basis, which can reduce errors due to individuals.

**Project (c): Predictive modelling using Decision Tree**

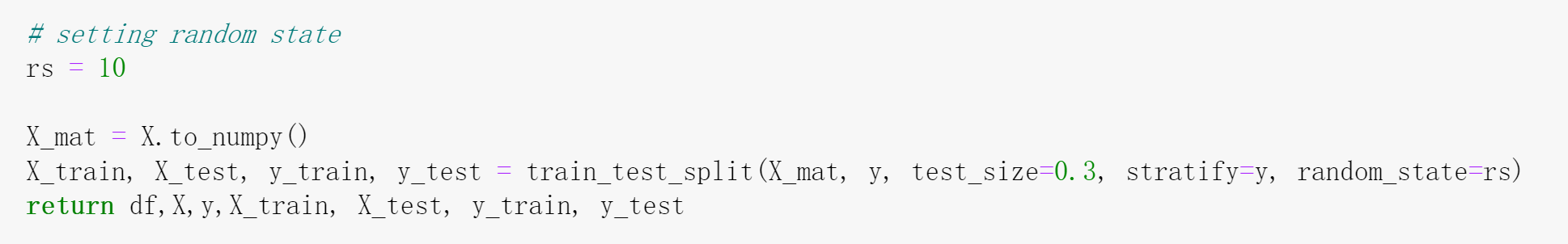
**1. What pre-processing was required on the dataset (D3.csv) before decision tree modelling?**

Firstly, we change three variables from interval/integer to nominal/str. And then we filled the missing values with mode.



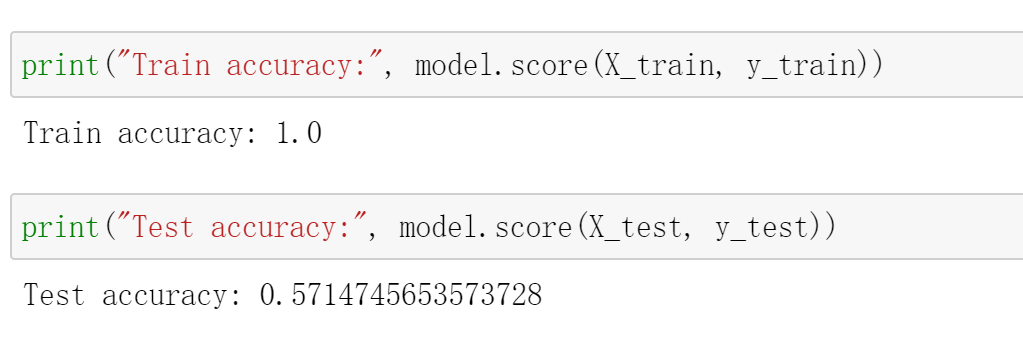
**What distribution split between training and test datasets have you used?**

We set test set as 30% and train set as 70%

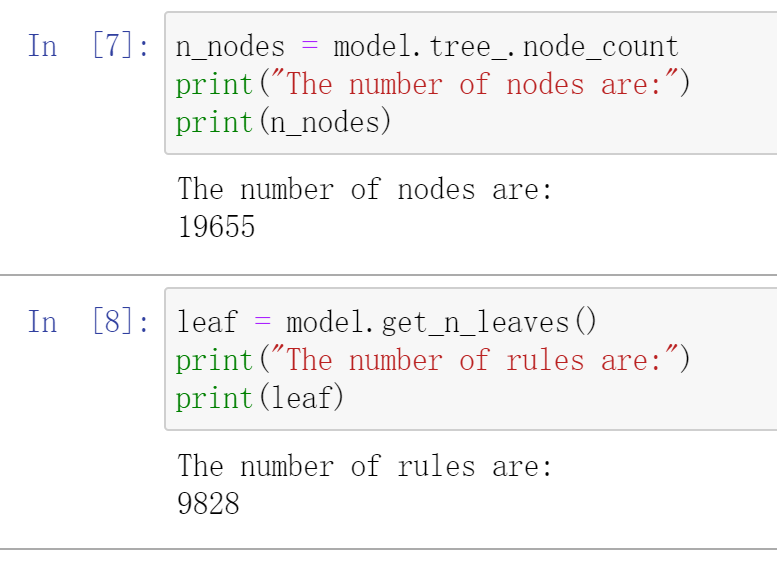


**2. Build a decision tree using the default setting. Answer the followings:**

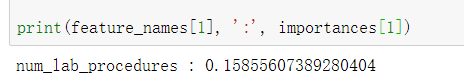
**a. What is the classification accuracy of training and test datasets?**

****

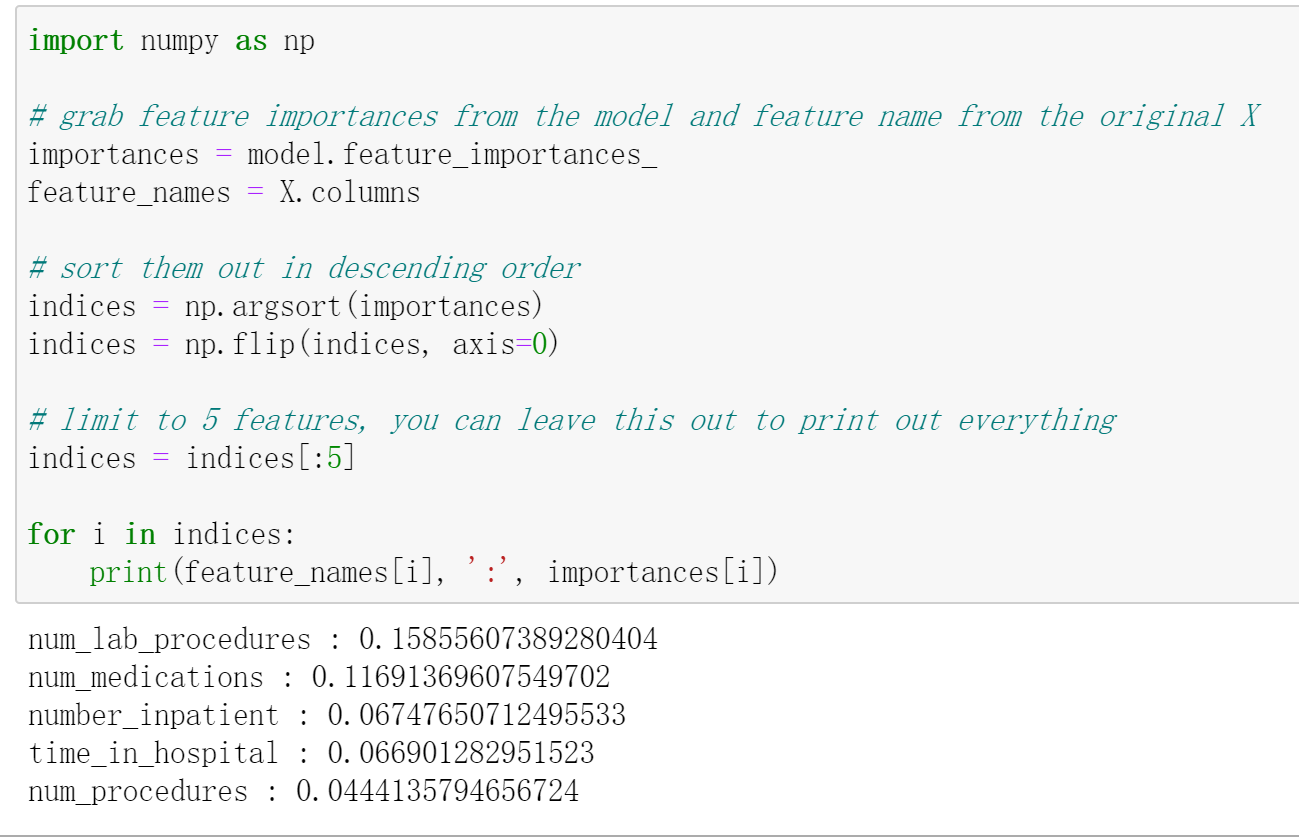
**b. What is the size of the tree (number of nodes and rules)?**

****

**c. Which variable is used for the first split?**



**d. What are the 5 important variables (in the order) in building the tree?**

****

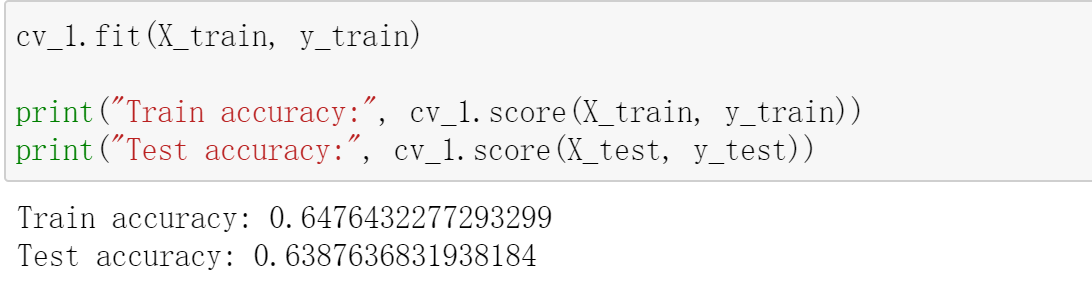
**e. What parameters have been used in building the tree? Detail them.**

The above tree is using the default setting but we can improve the accuracy by three parameters:

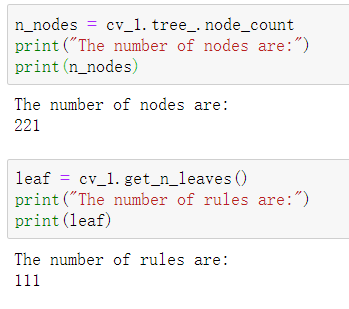
1. **Criterion:** An algorithm for determining a split's quality. This model used the gini method.
2. **Max depth:** The tree's maximum depth. Deeper models are more complicated and contain a greater number of nodes. This model has no depth restrictions, which means it can perfectly fit the data ( resulting in overfitting)
3. **Min samples leaf:** The minimum number of samples that must be present at a leaf node in order to limit the leaf node's minimum size. There is essentially no limitation on a node leaf in this model because it has a minimum sample leaf of one.

**3. Build another decision tree tuned with GridSearchCV. Answer the followings:**

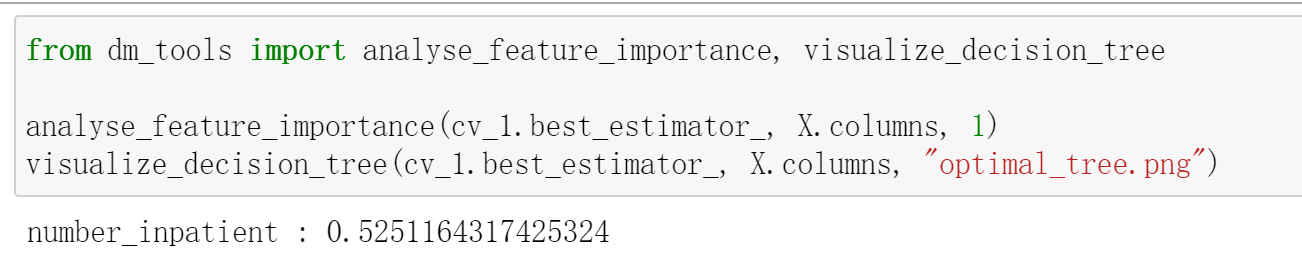
**a. What is the classification accuracy of training and test datasets?**

****

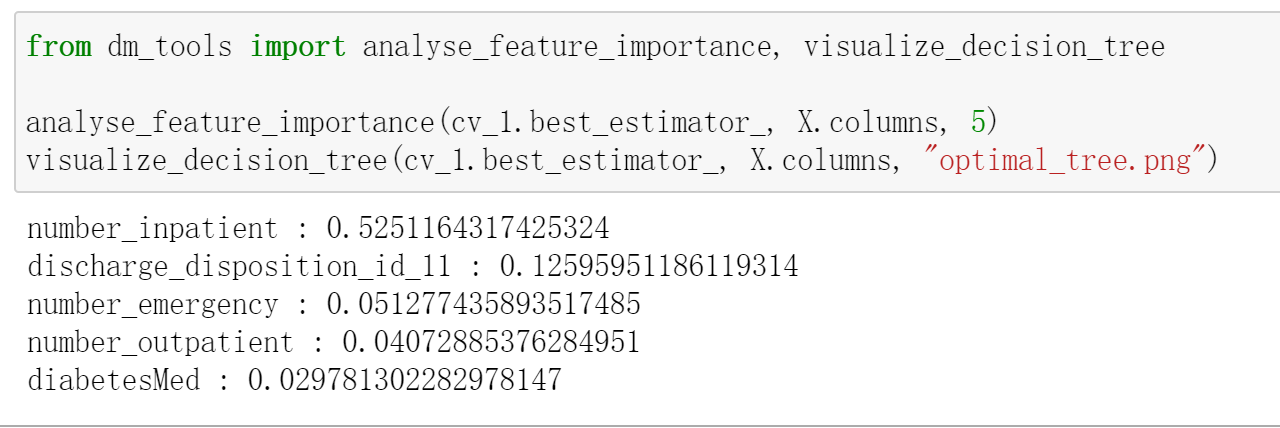
**b. What is the size of the tree (i.e. number of nodes and rules)?**

****

**c. Which variable is used for the first split?**

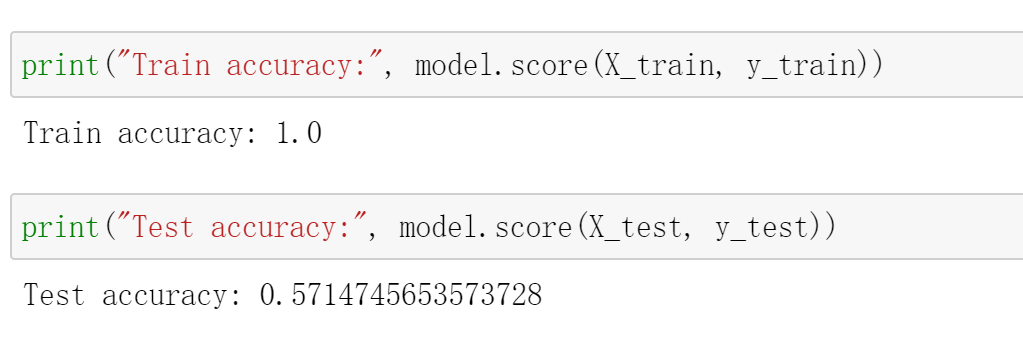
****

**d. What are the 5 important variables (in the order) in building the tree?**

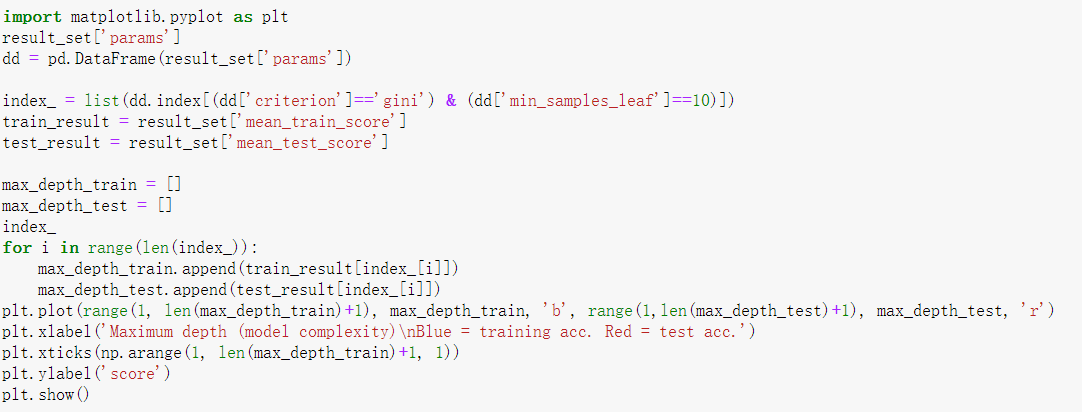
****

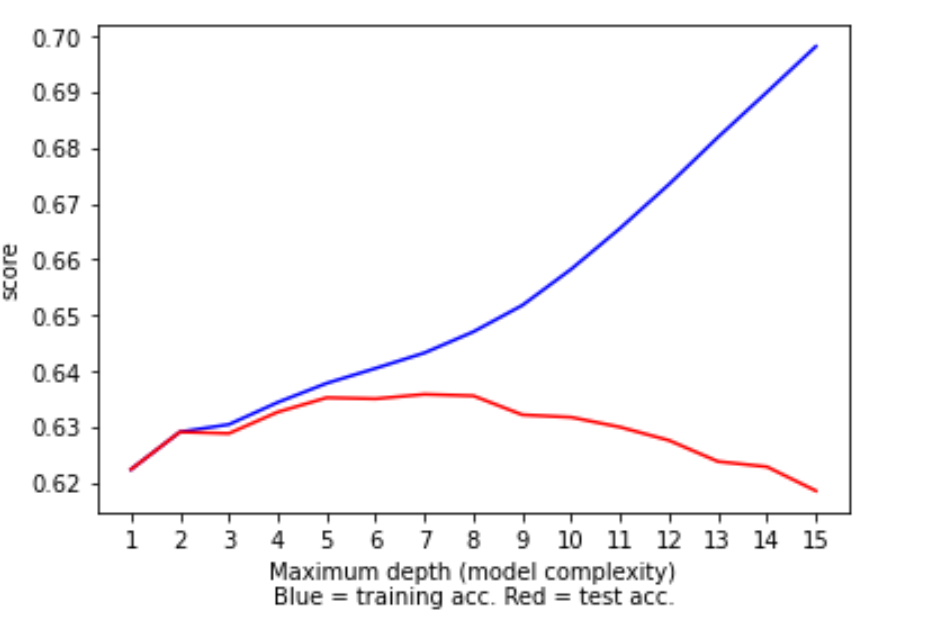
**e. Report if you see any evidence of model overfitting.**

As we did not set the max depth for the first tree we can find a 100% accuracy in the train set but lower accuracy in the test set.

****

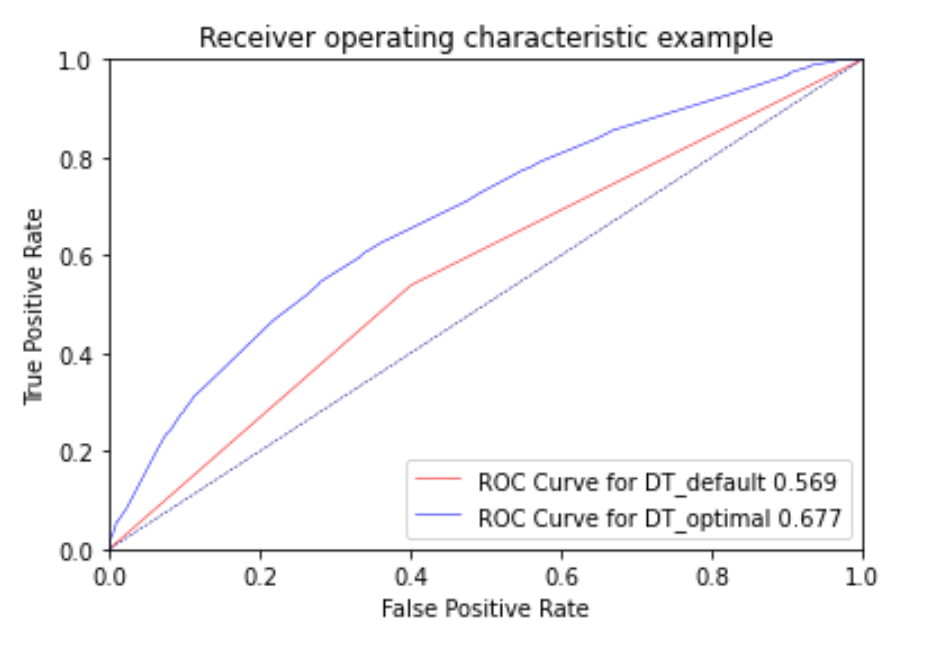
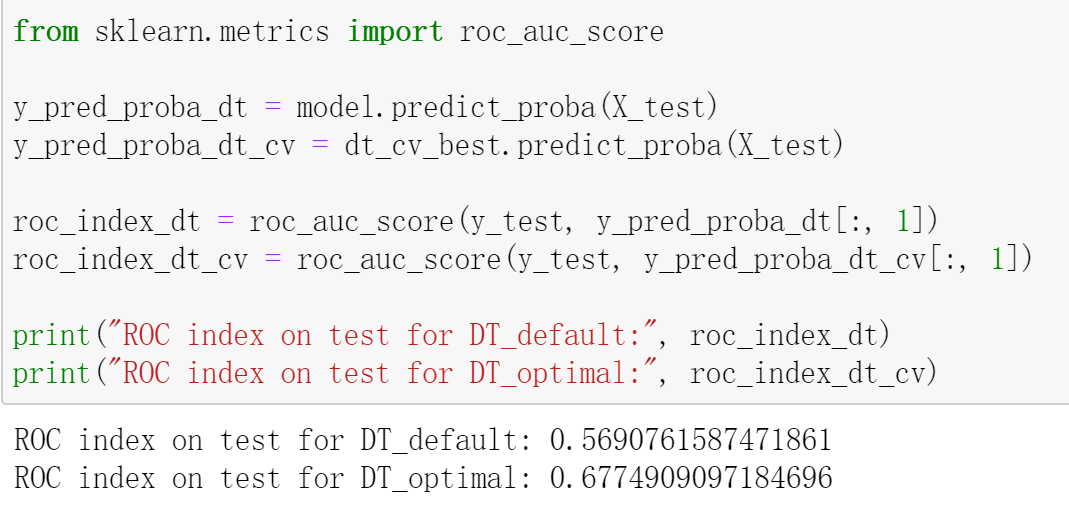
Next, we focus only on the max\_depth to understand the bias and variance for the fixed criterion = gini and max\_sample\_leaf = 10.



****

From the above plot we found overfitting appears after depth = 8.

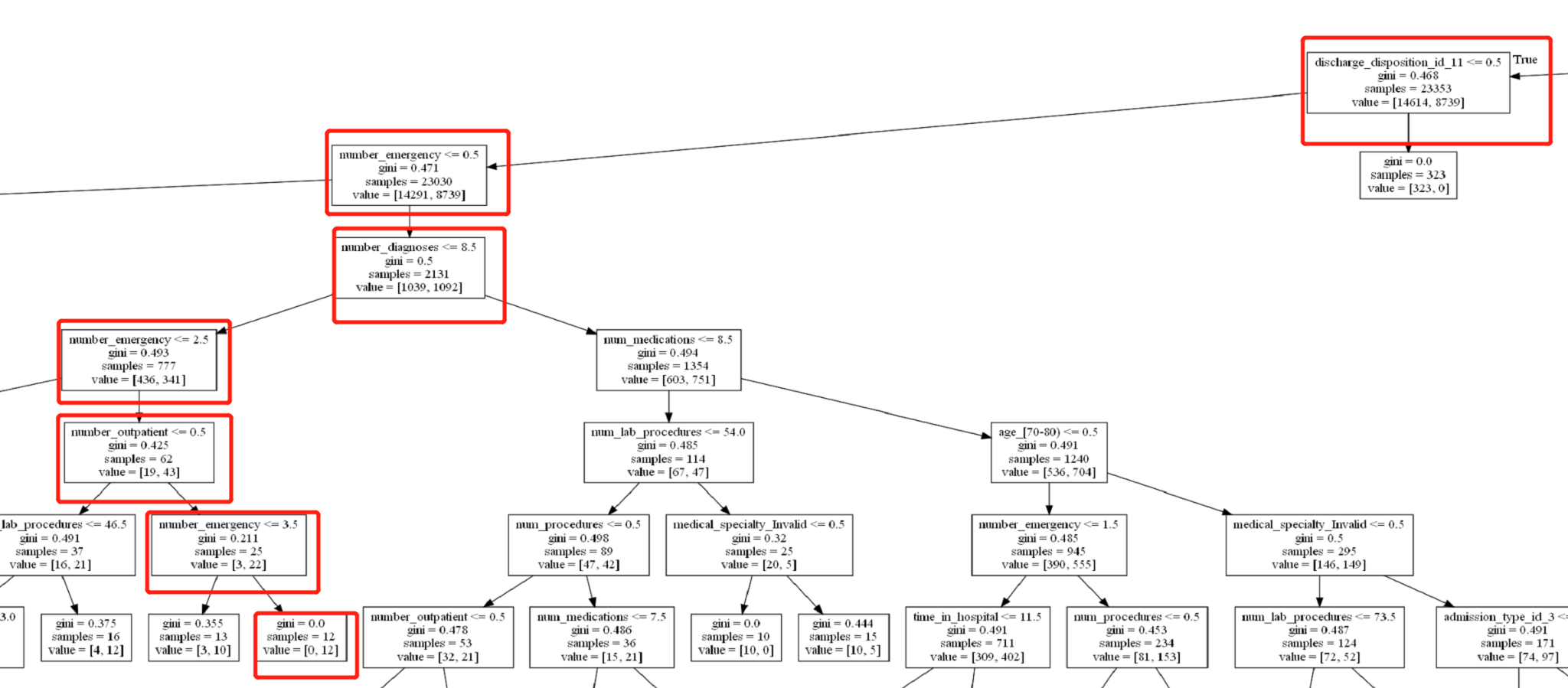
**4. What differences do you observe between these two decision tree models (with and without fine-tuning)? How do they compare performance-wise? Produce the ROC curve for both DTs. Explain why those changes may have happened.**

****

From the ROC score we can see the score of DT\_CV with fine-tuning is the highest (0.677). And the DT without fine-tuning gets a low score (0.569).

In addition, we also need to see the performance of different model’s curves. The closer the curve is to the top and left corner, the better the model is.The DT with GridSearchCV (DT\_CV) shows the largest curve area compared to the DT with default settings (DT). Thus DT\_CV is the best performing model.

**5. From the better model, can you identify which patients could potentially be "readmitted"? Can you provide general characteristics of those patients?**

****

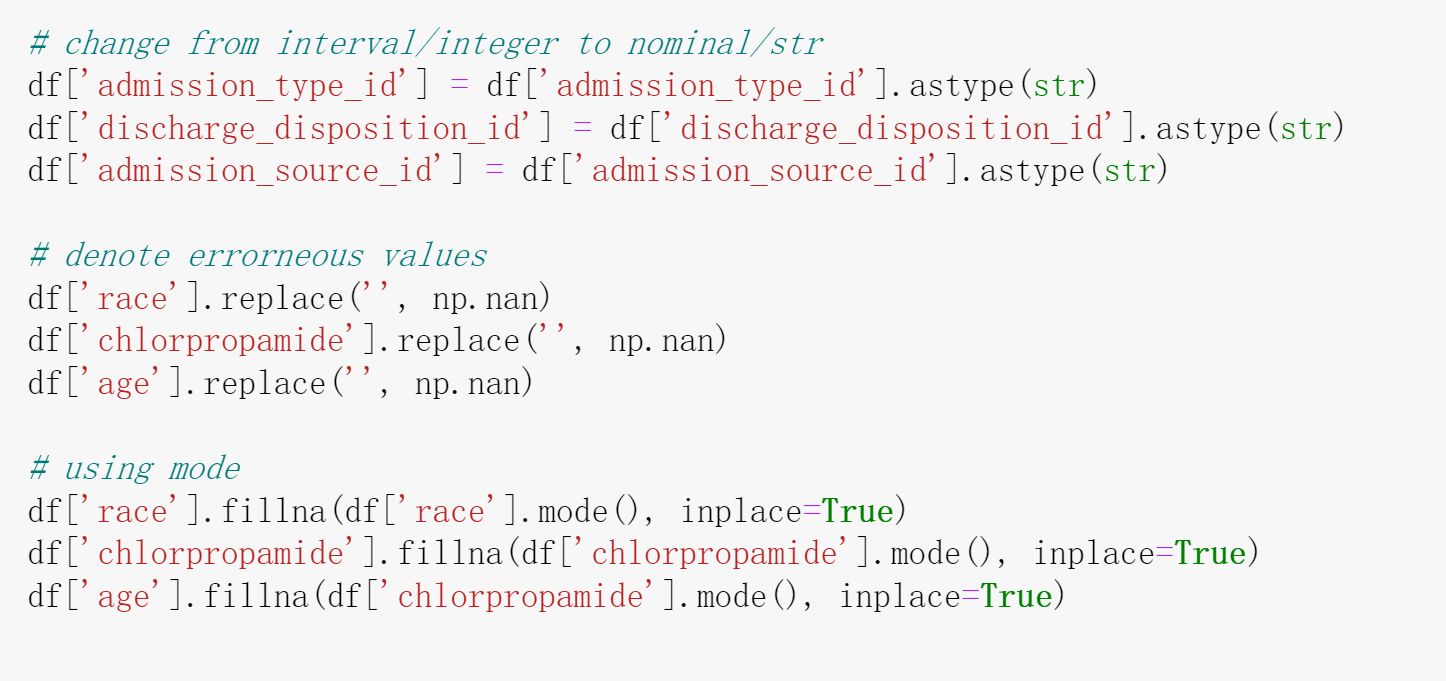
From the optimal decision tree, we can identify patients whose discharge\_disposition id is not 11(Expired), emergency visits of the patient in the year preceding the encounter high than 4, outpatient visits higher than 0 of the patient, and diagnoses entered to the system lower than 9 are more likely to be "readmitted".

* Discharge\_disposition is not “Expired”
* emergency visits >= 4
* outpatient visits > 0
* Diagnoses < 9

**Predictive modelling using Regression**

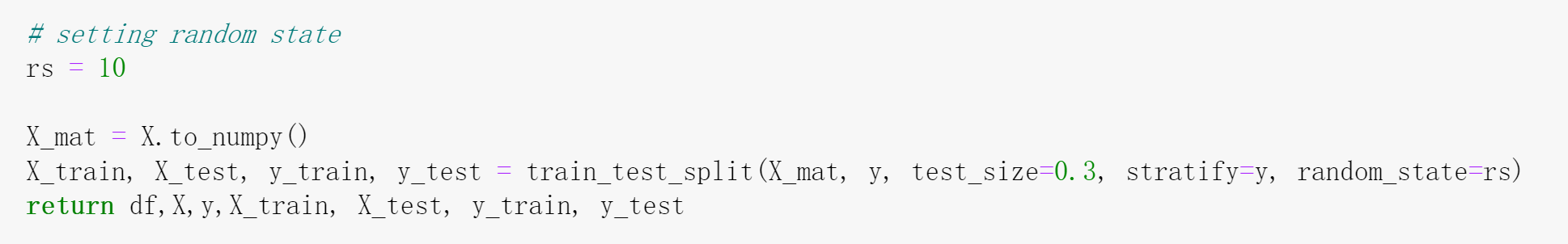
**1. What pre-processing was required on the dataset before regression modelling?**

Firstly, we change three variables from interval/integer to nominal/str. And then we filled the missing values with mode.



**What distribution split between training and test datasets have you used?**

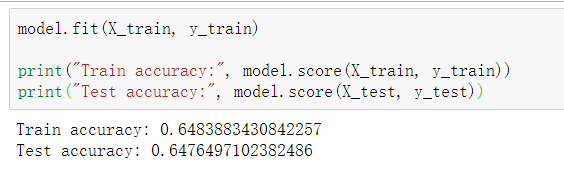
We set test set as 30% and train set as 70%



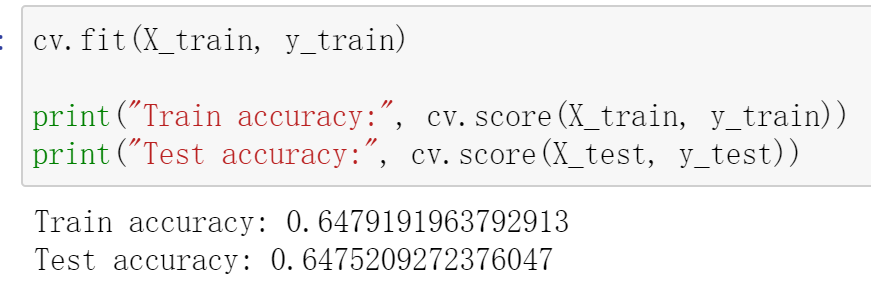
**2. Build a regression model using the default regression method with all inputs. Build another regression model tuned with GridSearchCV. Now, choose a better model to answer the followings:**

**a. Explain why you chose that model.**

The score of train data and test data in default model:



The score of train data and test data in the model tuned with GridSearchCV:

****

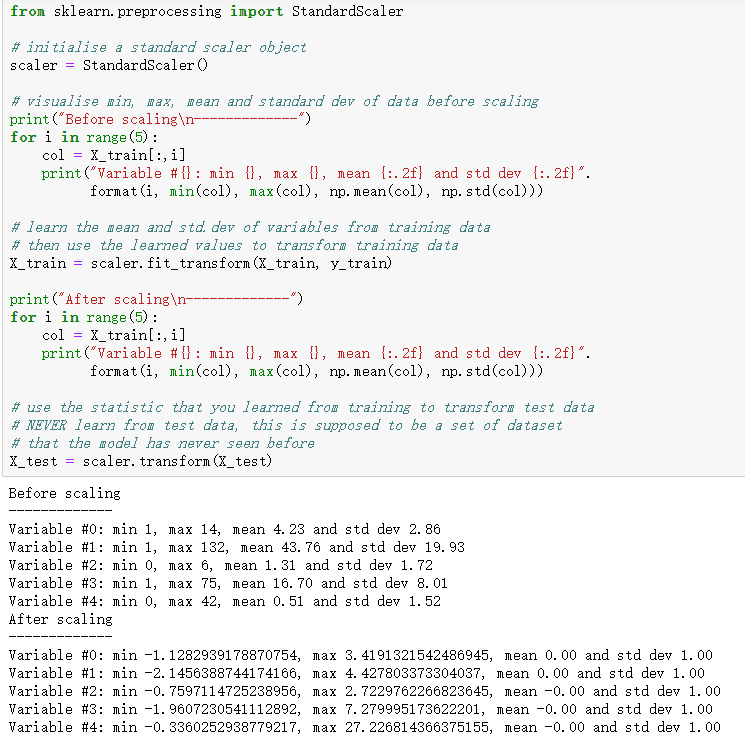
Firstly, we notice that the default model's train accuracy is slightly higher than the test accuracy, which could indicate slight overfitting, but this needs to be investigated further. We used GridSearchCV to fine-tune this logistic regression model to improve its accuracy. However, by comparing the accuracy of two models we choose the default model as it provides a slightly better accuracy in the test set.

**b. Name the regression function used.**

Logistic regression

As we are dealing with classification predictive mining, therefore, the logistic regression model is the best regression function.

**c. Did you apply standardization of variables? Why would you normalise the variables for regression mining?**

****

Yes. Because input variables on different scales make comparison between data points difficult.

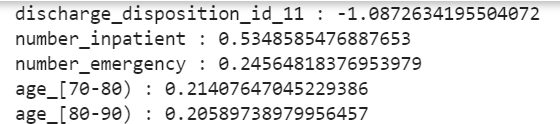
**d. Report the variables included in the regression model.**

The detailed variables are: race, gender, age, admission\_type\_id, discharge\_disposition\_id, admission\_source\_id, time\_in\_hospital, medical\_specialty, num\_lab\_procedures, num\_procedures, num\_medications, number\_outpatient, number\_emergency, number\_inpatient, number\_diagnoses, max\_glu\_serum, A1C result, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, insulin, change, diabetesMed.

Target value is readmitted.

**e. Report the top-5 important variables (in the order) in the model.**

The top-5 important variables are as follows:

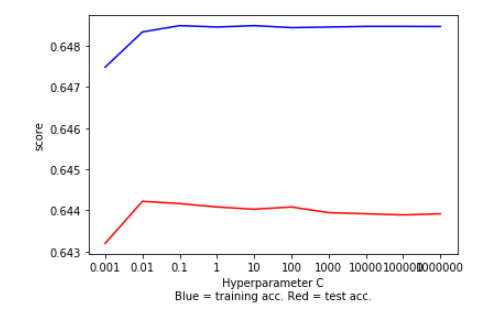
****

* Discharge\_disposition\_id\_11: Negative coefficient
* Number\_inpatient: Positive coefficient
* Number\_emergency: Positive coefficient
* Age\_[70 - 80): Positive coefficient
* Age\_[80 - 90): Positive coefficient

**f. What is the classification accuracy on training and test datasets?**

****

**g. Report any sign of overfitting in this model.**

****

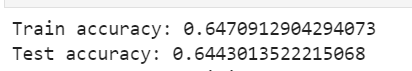
From the plot above, it shows that the model performs best when C = 0.01, and overfitting starts at this point. The gap between test accuracy and training accuracy gradually increases after 0.01.

**3. Build another regression model on the reduced variables set. Perform dimensionality reduction with Recursive feature elimination. Tune the model with GridSearchCV to find the best parameter setting. Answer the followings:**

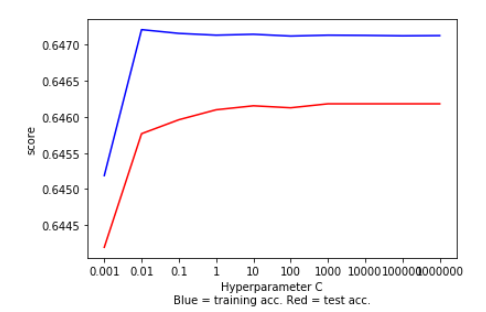
**a. Was dimensionality reduction useful to identify a good feature set for building the accurate model?**

RFE works by first training the model on all features. Each feature is assigned a weight. Features with small weights (less important) are eliminated, making a smaller feature set. This process is repeated a number of times until reaching the optimal performance. Therefore, it is useful to identify a good feature set for building an accurate model.

**b. What is the classification accuracy on training and test datasets?**

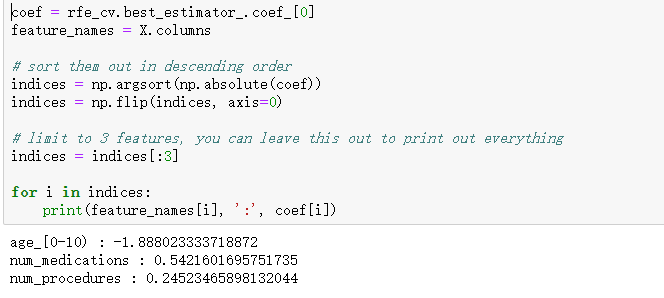
****

**c. Report any sign of overfitting.**

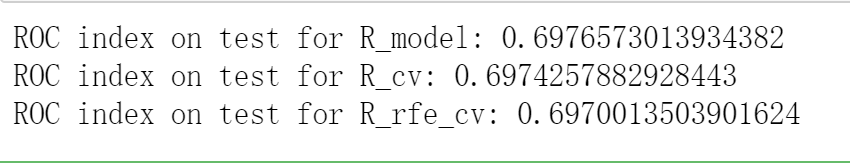
****

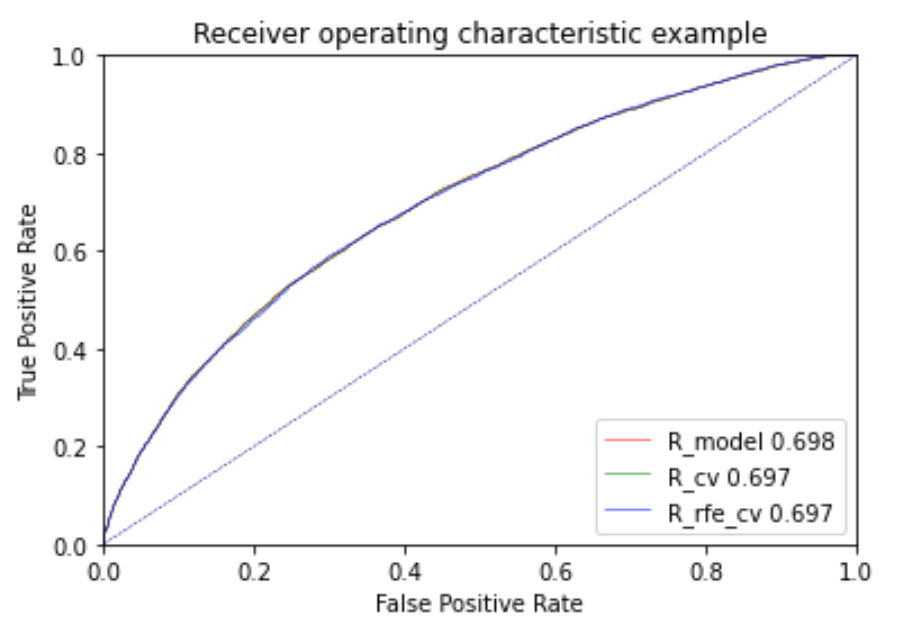
From the above plot, we could not find any sign of overfitting. The model is trained well after c=0.1.

**d. Report the top-3 important variables (in the order) in the model.**

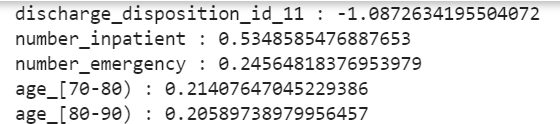
****

**4. Produce the ROC curve for all different regression models. Using the best regression model, can you identify which patients could potentially be “readmitted”? Can you provide general characteristics of those patients?**

****

****

From the ROC result, we choose the default regression model.

****

By the top 5 features shown above, we know patients who have elder age and a large number of inpatient visits and emergency visits will probably be readmitted.

**Predictive modelling using Neural Networks**

**1. What pre-processing was required on the dataset before neural network modelling? What distribution split between training and test datasets have you used?**

We reuse the data preparation function we builded in the previous model.

****

**2. Build a Neural Network model using the default setting. Answer the following:**

**a. Explain the parameters used in building this model, e.g., network architecture, iterations, activation function, etc.**

In the first MLPClassifier, we build the model without any additional parameter except the random state.

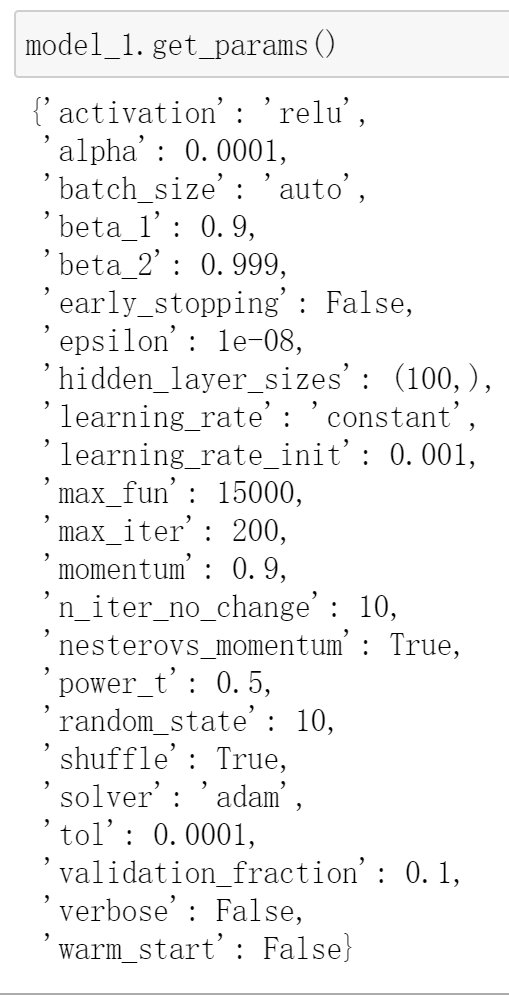
**network architecture**: it is formed in three layers, called the input layer, hidden layer, and output layer.

**iterations**：describes the number of times a batch of data passed through the algorithm.

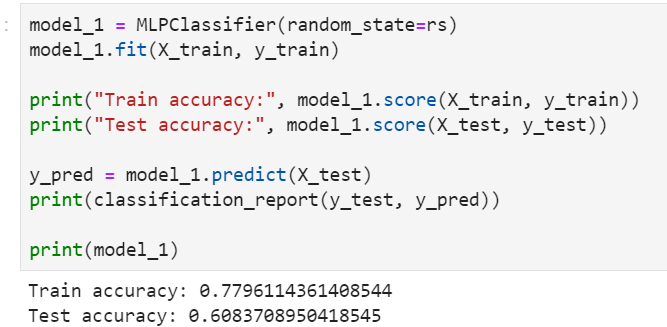
**activation function**：it is a function that converts all negative numbers to zero.

**Solver**: the solver for weight optimization.

The parameters of the default setting as below:



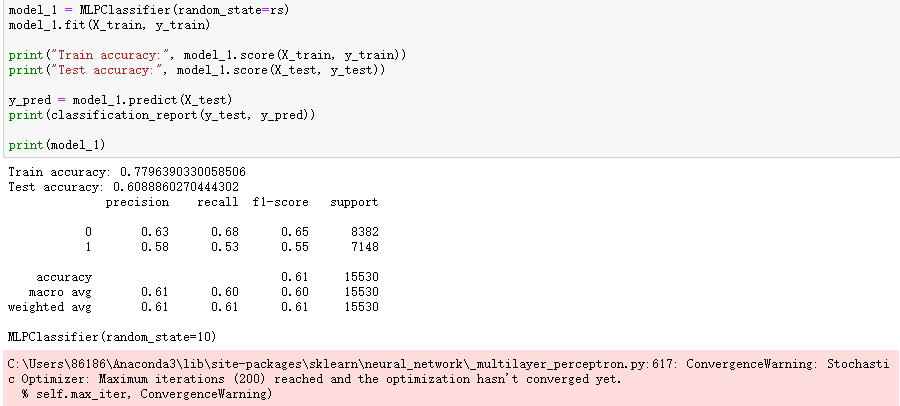
**b. What is the classification accuracy on training and test datasets?**

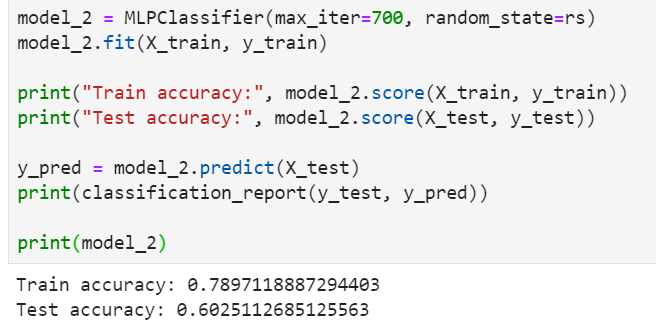
****

**c. Did the training process converge and result in the best model?**

No, the training process did not converge and result in the best model. In sklearn, if a neural network does not achieve convergence before maximum iteration, it will raise a "convergence is not reached" warning message，like the below picture. And then, the `max\_iter` hyperparameter of the neural network should be increased. We set the `max\_iter` as 700.

The training dataset shows the accuracy of 0.7796 while the testing dataset shows 0.60837 (lower), leading to overfitting to the training data. It is necessary to notice a convergence warning. Also, the neural network after setting the iteration of 700 performed with a little higher accuracy on the training dataset(0.7897), but test accuracy remains the same. This also indicates that overfitting to the training data.

****



**3. Refine this network by tuning it with GridSearchCV. Report the trained model.**

**a. Explain the parameters used in building this model, e.g., network architecture, iterations, activation function, etc.**

**network architecture**: it is formed in three layers, called the input layer, hidden layer, and output layer.

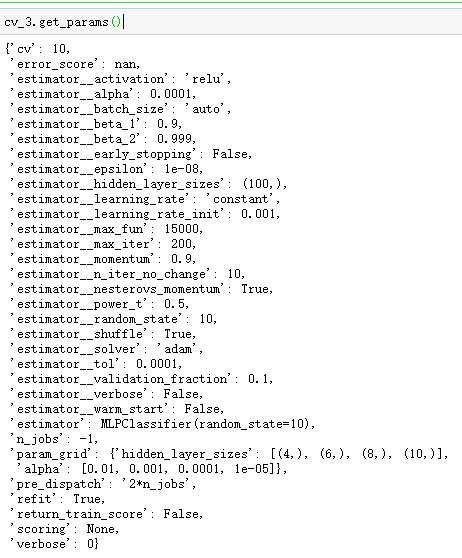
**Iterations**: describes the number of times a batch of data passed through the algorithm.

**activation function**: it is a function that converts all negative numbers to zero.

**Solver**: the solver for weight optimization.

**param\_grid**: is the dictionary with parameters names (str) as keys and lists of parameter settings to try as values.

The parameters of this model setting as below：

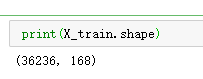


We used **GridSearchCV** to find the best hidden layers size and alpha rate.

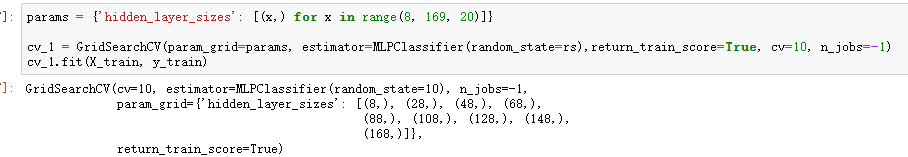
Hidden layers size: the number of neurons exist in the hidden layers.

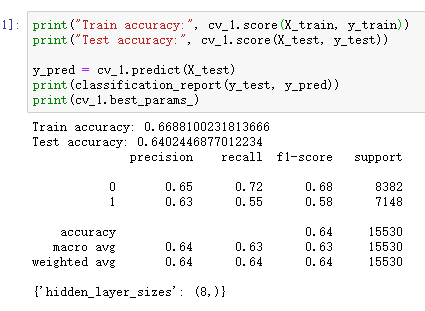
Alpha rate: A momentum item used to update weights during training. The momentum is intended to move the weight change in a consistent direction.

We can see there are 168 input features.

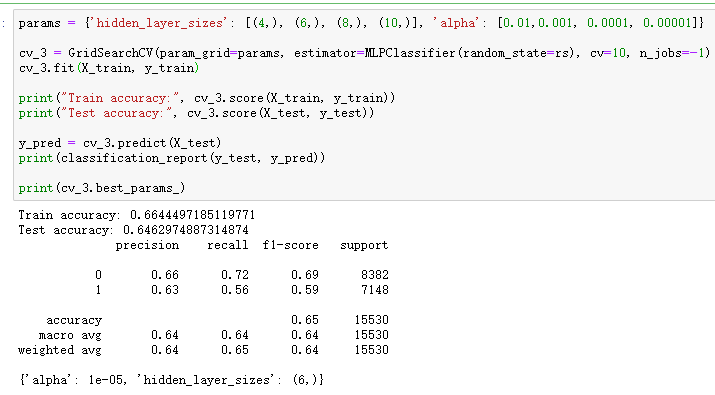


Therefore,we will start tuning with one hidden layer of 8 to 168 neurons, increment of 20.



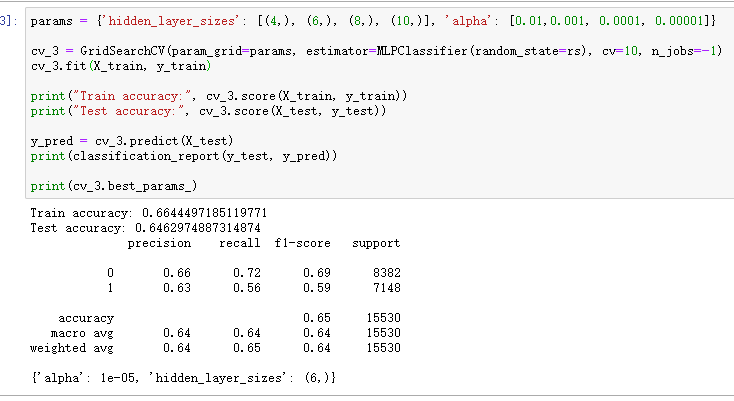


We found the optimal number of neurons is 8. Next, we attempted to tune the model with the lower number of neurons in the hidden layer and try to find the best alpha rate.



Finally, we found the best alpha rate is 1e-05.

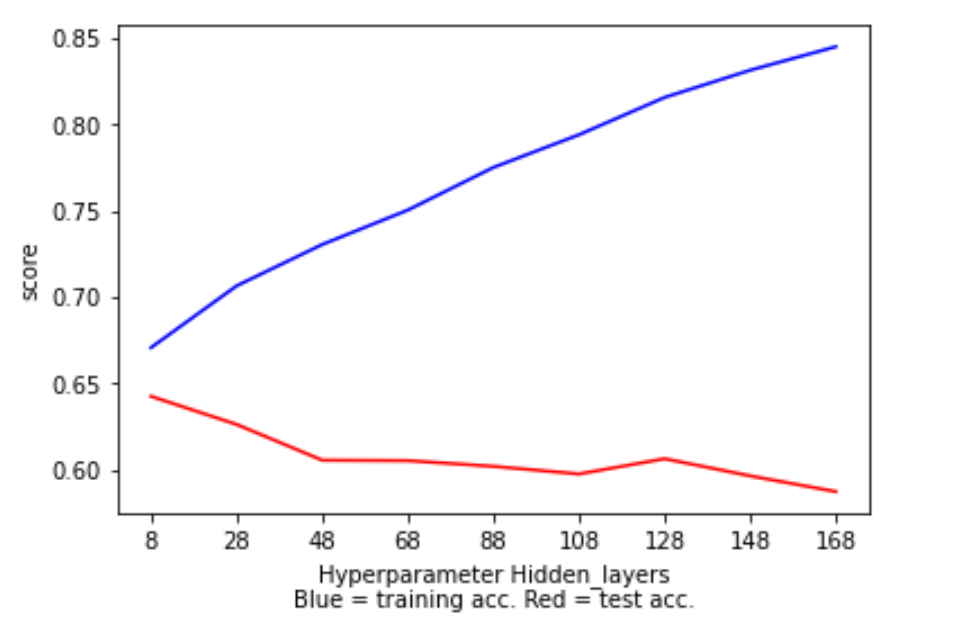
**b. What is the classification accuracy on training and test datasets?**

****

**c. Did the training process converge and result in the best model?**

Yes, the training process converged and resulted in the best model. In sklearn, if a neural network does not achieve convergence before maximum iteration, it will raise a "convergence is not reached" warning message. We don’t receive the warning message “Maximum iterations (200) reached and the optimization hasn't converged yet.” It means the training process converged.

**d. Do you see any sign of over-fitting?**

****

It is obvious that the plot above shows the sign of over-fitting from the hidden\_layer\_sizes as 8.

**4. Let us see if feature selection helps in improving the model? Build another Neural Network model with reduced features set. Perform dimensionality reduction by selecting variables with a decision tree (use the best decision tree model that you have built in the previous modelling task). Tune the model with GridSearchCV to find the best parameters setting.**

**Answer the following:**

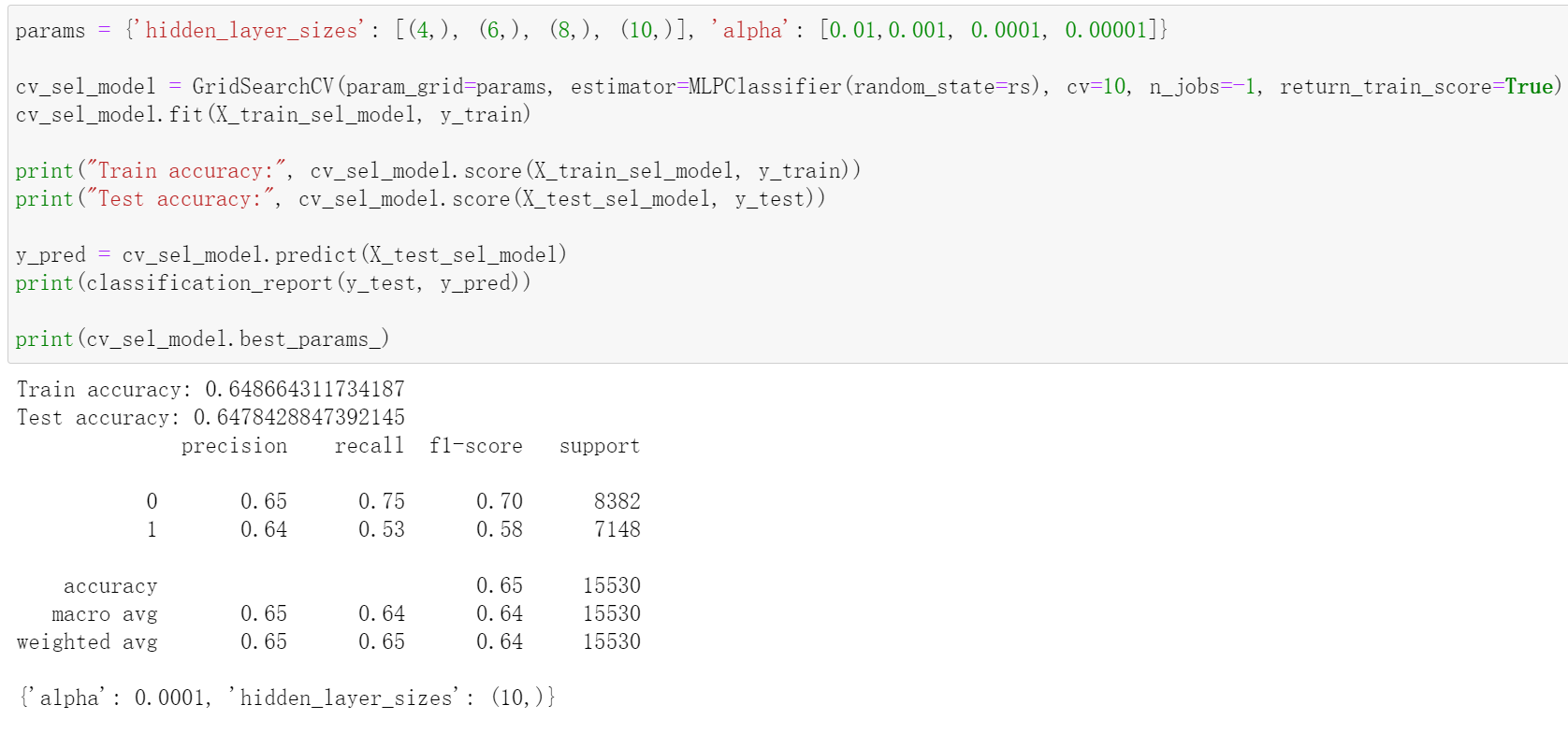
**a. Did feature selection favour the outcome? Any change in network architecture? What inputs are being used as the network input?**

Yes, the Neural Network model trained with decision tree selected variables manage to improve model performance.

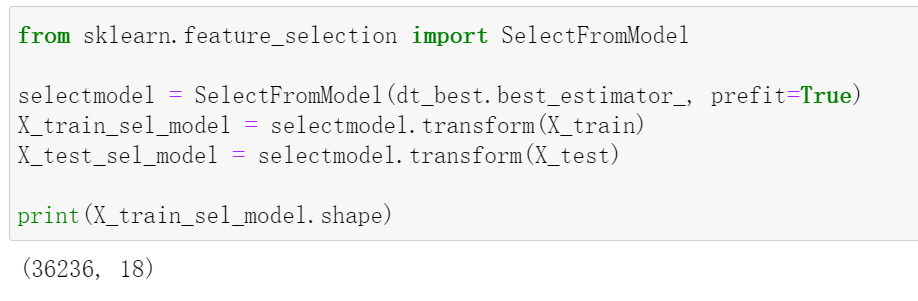
The change in network architecture is the parameter ‘hidden\_layer\_sizes’ changed range from (8, 169) to (4, 6, 8, 10). We tune the parameter of the default model and find the best parameter(hidden layer size: 10).



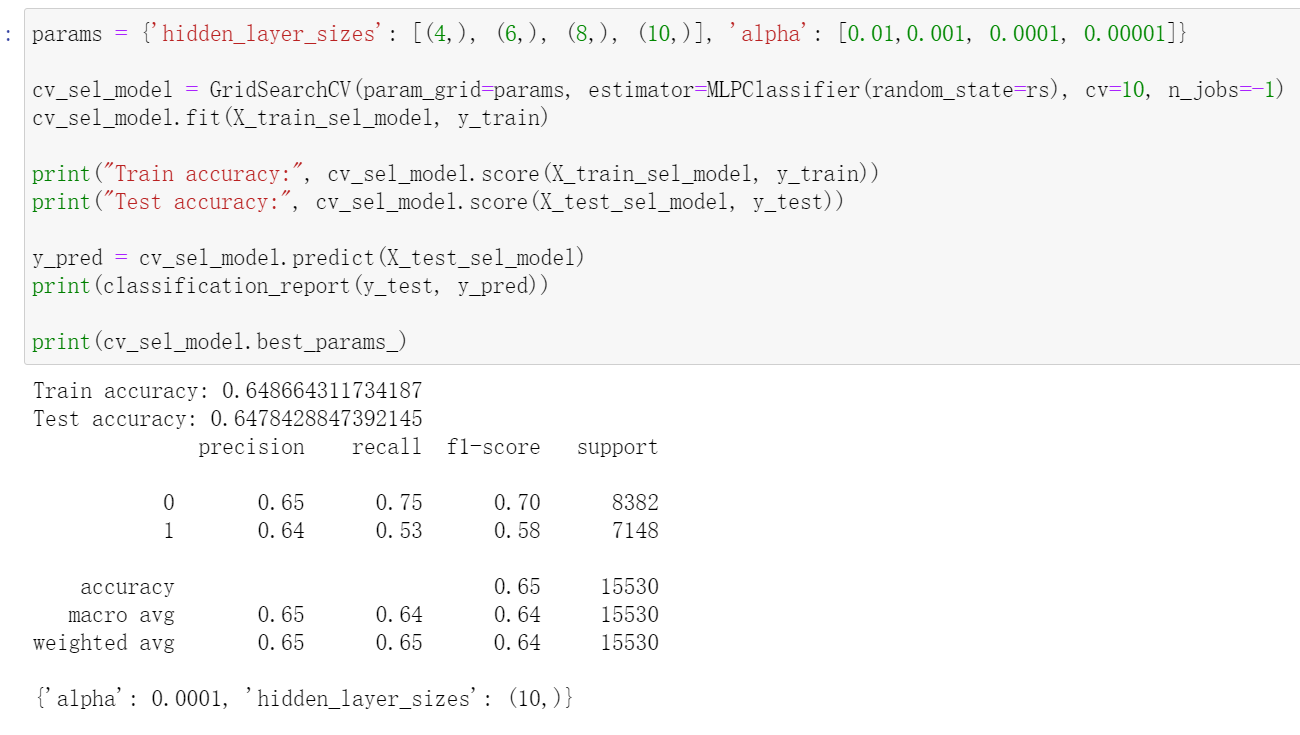




This model identifies the set of 18 variables as the important features.



**b. What is the classification accuracy on training and test datasets?**

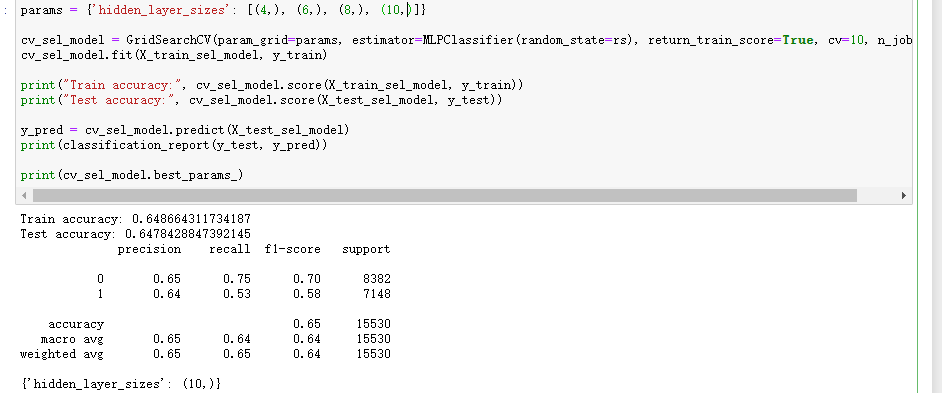
****

**c. How many iterations are now needed to train this network?**

200 or less than 200. We use the default set of the max iteration, and we don’t receive the warning message “Maximum iterations (200) reached and the optimization hasn't converged yet.”



**d. Do you see any sign of over-fitting? Did the training process converge and result in the best model?**

****

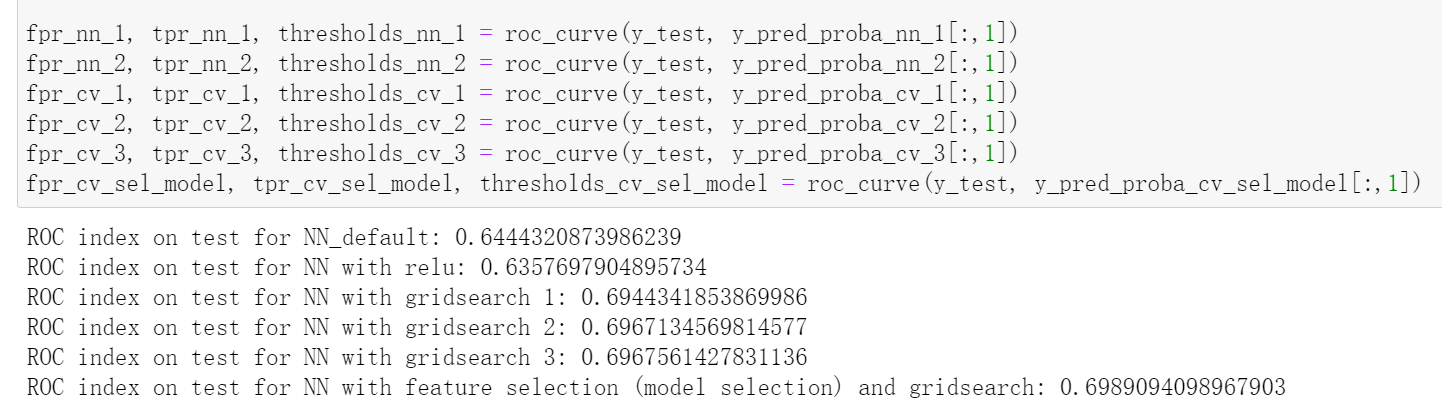
**We used GridSearchCV to find the best hidden\_layers size is 10(4-10).**

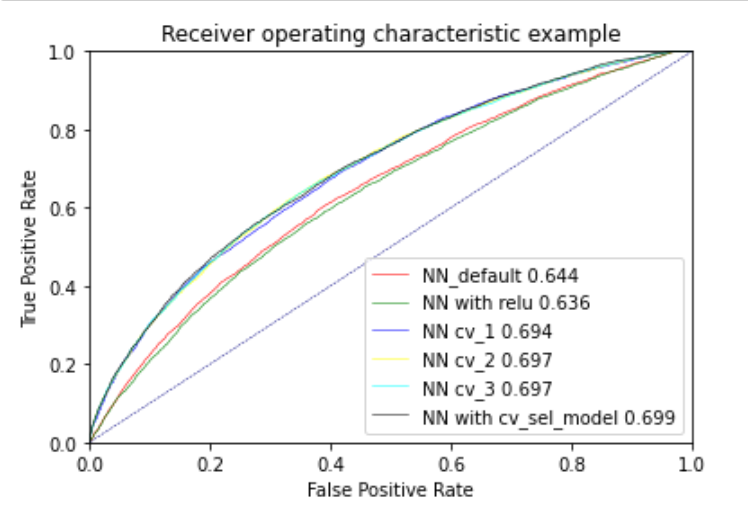
****

**From the above plot, we did not find any sign of over-fitting.**

**5. Produce the ROC curve for all different NNs. Now, using the best neural network model, can you provide general characteristics of the patients identified by the model? If it is difficult (or even infeasible) to comprehend, discuss why?**

According to the ROC curve below, the neural network with feature selection(a decision tree) and grid search is the best model(ROC: 0.699).

****

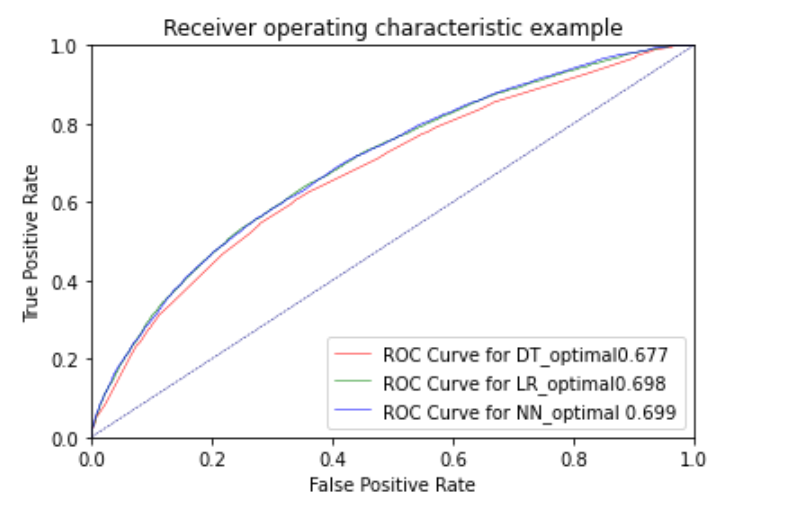
****

No, because neural networks can only predict readmission based on the features of patients, it is difficult to directly get general features of such patients from the neural network. There is no straightforward relationship between the weights and the estimated function.

**Final remarks: Decision making**

**1. Finally, based on all models and analysis, is there a model you will use in decision making? Justify your choice. Draw a ROC chart and Accuracy Table to support your findings.**

We will use the Neural Network model in decision making. From the ROC chart and accuracy table below, the Neural NetWork has the highest ROC score(0.699) and accuracy.

****

|  |  |  |
| --- | --- | --- |
|  | **Train accuracy** | **Test accuracy** |
| **Decision Tree** | **0.6476432277293299** | **0.6387636831938184** |
| **Logistic Regression** | **0.6483883430842257** | **0.6476497102382486** |
| **Neural Network** | **0.648664311734187** | **0.6478428847392145** |

**2. Can you summarise the positives and negatives of each predictive modelling method based on this analysis?**

|  |  |  |
| --- | --- | --- |
|  | **Positives** | **Negatives** |
| **Decision tree** | It takes less time to run the code(less amount of computation). | Easy to be overfitting. |
| **Logistic Regression** | Easy to display each feature’s impact (coefficients) to the target variable.  Less likely to be overfitting | The missing values and noise need to be processed in the dataset. |
| **Neural Network** | Can build a more accurate model compared to other methods. | Large amounts of computation (take a long time to run the code) for training.  Easy to be overfitting  Hard to know the details of variables in the models (Black-box). |