**QUEENSLAND UNIVERSITY OF TECHNOLOGY**

****

School of Information Systems

Faculty of Science and Engineering

IFN 509 Data Mining and Exploration

**Assessment Item 2 Part 1**

Semester One 2020

Submitted To: Richi Nayak and Thiru Balasubramaniam

Date: 24th-May-2020

Submitted By:

|  |  |
| --- | --- |
| Student’s Full Name | Student ID |
| Drashti Nayak | N10599568 |
| Radha Kundan Shah | N10565302 |
| Sepehr Borooshan | N10602399 |

Task 1: Preparation of data (2 marks)

1. What is the proportion of patients who are classified as COVID-19 positive? Would you require a sampling technique on this data?

Ans—The approach used for this question was pretty straight forward. We first find out the total number of exams result which are positive and negative. Once we got that the formula is simple:

Hence the total proportion of patients who are classified as COVID-19 positive is: 45.366%

Note: We used the function of ‘np. round’ to limit the decimal places up to 3.

Sampling Technique—Yes, we did use the stratified sampling technique to keep data consistent and to make sure that the test and train data set which we created have equal number of positive and negative cases to give us a correct accuracy.

1. Did you have to fix any data quality problems? Detail them.

Ans—Yes, we did fix some data in our preprocessing. this was done to make data more understandable by the model and give better results. The following is the list of the data quality we fixed.

1. Age: In patient quarantine age first the rows with age ‘0’ is replaced and filled with the mean of the ages given in the entire column. Then the quarantine age column is converted into a string data type because the age feature is not to be counted in calculations.
2. Influenza: The empty spaces in rapid test influenza A and rapid test influenza B is filled and replaced with the mode of the values in the entire column. Here we used mode instead of mean because if we look the column there were 3 value ‘NAN’ apart from positive and negative. Another reason of doing mode is that the columns of influenza contains categorical variables which mean it can either be positive or can be negative.
3. Mean Inputs: Same as in for the age column, other columns we choose and had ‘0’ or empty values was replaced and filled with the mean value of that particular column.
4. One-hot encoding: after all the missing values were filled last, we did was one-hot encoding the reason of performing this was to convert all the categorical variables into binary variables. As some modules take only numerical values as their inputs.
5. What variables did you include in the analysis? justify the choice of input selection (input/target).

Ans— As the hospital wanted to know the characteristics of people who are COVID-19 positive hence the SARC-Cov-2 exam result will be the target variable here and as the hospital is interested in knowing the outcome of the blood test characteristics and the influenza results, they are as input variables

|  |  |
| --- | --- |
| Input Variables | Target Variables |
| 'Proteina C reativa', 'Neutrophils', 'Mean platelet volume', 'Monocytes', 'Red blood cell distribution width', 'Red blood Cells', 'Platelets', 'Eosinophils', 'Basophils', 'Leukocytes', 'Mean corpuscular hemoglobin', 'Mean corpuscular volume', 'Mean corpuscular hemoglobin concentration ', 'Lymphocytes', 'Hemoglobin', 'Hematocrit' ’Influenza B rapid test’,‘Influenza A rapid test’ | ‘SARS-Cov-2 exam result’, |

Task 2: Predictive modeling using Decision Tree (4 marks)

1. Build a decision tree using the default setting.

1. The train accuracy we got in default decision tree is: 0.7398373983739838

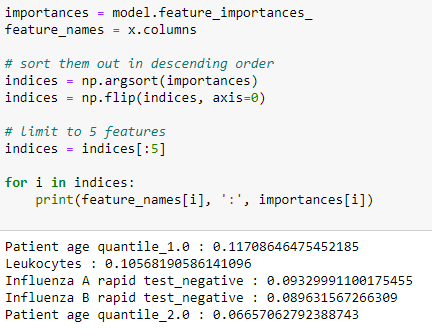
The test accuracy we got in default decision tree is: 0.7032520325203252

1. Total numbers of nodes are 121 and 11 rules used to predict sample 0.
2. Variable in the first split?

Ans- For the first split, patient age quantile\_1 is used where gini = 0.496 quantifies the purity of node and shows that sample in that node belong to different classes. Samples = 984 as data contains 984 relevant samples. Value [538, 446] shows how many samples fall into category of exam result being negative or positive.

1. 5 Important Variables

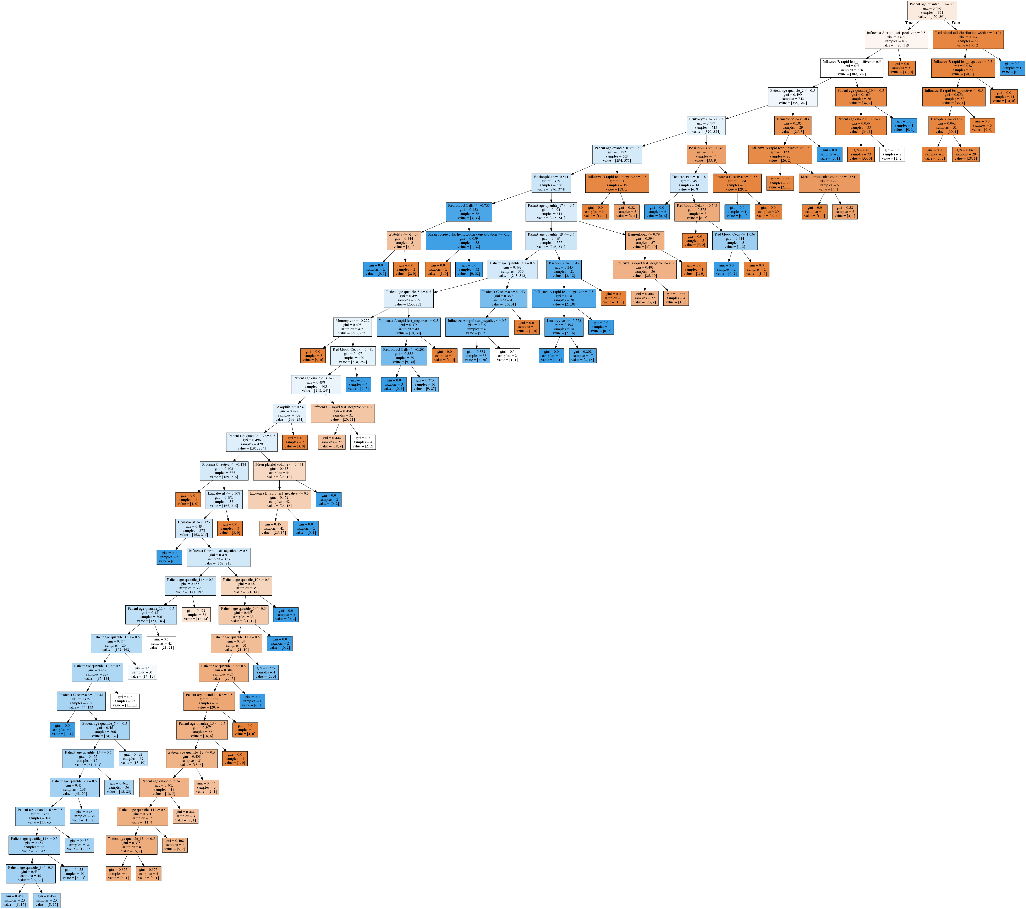
Ans- To find the 5 important variables for the default decision tree we just printed the top 5 features after they were arranged in a descending order with highest value first. The following image shows the 5 important variables. However, we cannot include this code in while making decision tree because this was is to made on default setting so we will remove the indices ‘:5’ and keep it ‘:’ to print all data.



1. Parameters used in making the decision tree

Ans- Since the model was build on the default setting the following are the parameters used:

* max\_depth here is none, max\_depth is for limiting how deeper the model can go
* criterion is gini. Criterion for defining the quality when the data is split into various data sets.
* min\_samples\_leaf is set to 1 in the default. This is to determine the minimum number of leaf nodes in the decision tree
* The splitter used is best and random sample is set to 10 which means it will create 10 different random data sets splits. And some others are as follows
* class\_weight=None, criterion='gini', max\_depth=None, max\_features=None, max\_leaf\_nodes=None, min\_impurity\_decrease=0.0, min\_impurity\_split=None, min\_samples\_leaf=1, min\_samples\_split=2, min\_weight\_fraction\_leaf=0.0, presort=False, random\_state=10, splitter='best'



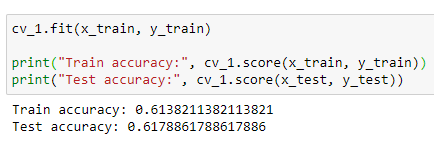
Default Decision Tree

2. Build another decision tree tuned with GridSearchCV

1. The train accuracy we got in default decision tree is: 0.5762195121951219

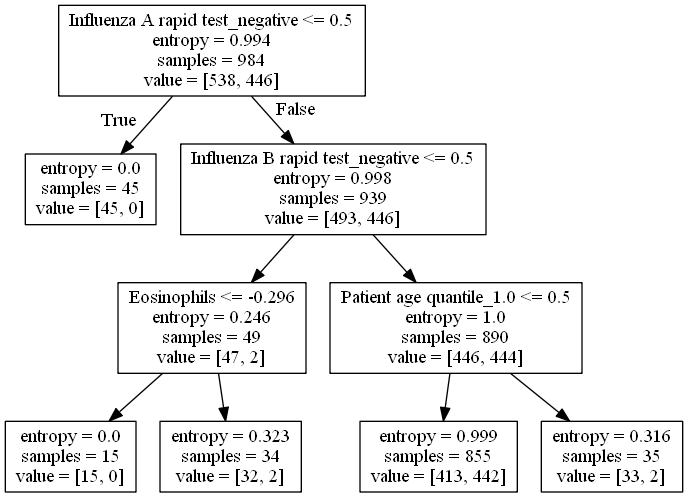
The test accuracy we got in default decision tree is: 0.5894308943089431

1. Yes, as the previous model was overfitting. With tuning of hyperparameters the size of the grid search cv is definitely smaller than that of default decision tree and this is because while forming the grid search cv we limited the depth/complexity of the tree to 4 which means max\_depth=3 and leaf node 20 and so total number of nodes are 9. As a result of which we can see the decision tree using grid search cv is less complex compared to decision tree based on default setting
2. For the first split, Influenza A rapid test negative is used where entropy = 0.994 quantifies the impurity of node and shows that sample in that node belong to different classes. Entropy is higher; hence variance is also must be higher. Samples = 984 as data contains 984 relevant samples. Value [538, 446] shows how many samples fall into category of exam result being negative or positive.
3. 5 Variables Used:
   1. Influenza A rapid test\_negative: 0.3990042303549945
   2. Influenza B rapid test\_negative: 0.347424494071442
   3. Patient age quantile\_1.0: 0.2429134177521585
   4. Eosinophils: 0.010657857821405043
   5. Mean corpuscular hemoglobin: 0.0
4. As per the accuracy, it is found out that the model shows evidence of underfitting while default model showed evidence of overfitting. Hence, it can be said that while tuning the hyperparameters the perfect model is somewhere in between the set range value of both models.



1. In grid search cv the following are the parameters used

* max\_depth here is a range of 1-4, max\_depth is for limiting how deeper the model can go. So here the model will stop after the depth of 4
* criterion is gini and entropy. Criterion for defining the quality when the data is split into various data sets.
* min\_samples\_leaf is set to 0-25 and at a step of 5. This is to determine the minimum number of leaf nodes in the decision tree
* random sample is set to 10 which means it will create 10 different random data sets splits. And splitter to be used is the best



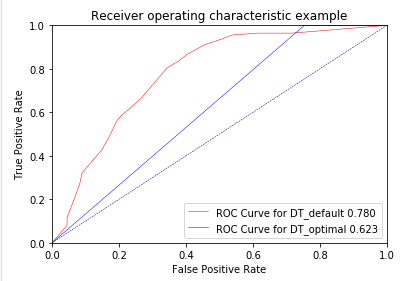
Grid search cv

3. What is the difference do you see between these two decision tree models?

Ans- The test accuracy for default setting is:0.7032520325203252

The test accuracy for grid\_cv is: 0.5894308943089431

The following is the roc curve for both the models



Which means the default setting gives us more accuracy for the data sets compare to that of grid search cv. Also, the default setting gives largest curve area hence the default setting decision tree is the best performing model on this data.

* These changes happened because the model build with grid cv is limiting the complexity of the model in terms of depth hence it stops taking into consideration few columns compared to the default setting which take all the columns in the data.
* Also, the results achieved is so can be because when we are limiting the leaf nodes of the decision tree in grid cv by reducing the max\_depth the accuracy is also changing while testing the ranges of max depth we saw at the range 1-7 the test accuracy score of grid cv model was same to that of default setting. At 1-8 and so on the accuracy of grid cv keeps increasing than that of grid. Hence that data set needs to be pruned in so that no matter we limit the leaf nodes the accuracy remains the same.
* Other reason why default setting Is better can be because the sample size of the dataset is not enough to give us proper results.
* One reason we fell is that as many columns originally had missing values or values which were ‘0’ and was not acceptable for that particular data and we had to fill them with mean value which is uncertain if that value can actually be true or not for that patient

1. Identify which patients could potentially be "COVID-19 Positive?

* Since the model using default setting here is our best model, we can say all the characteristics from Default decision tree image.

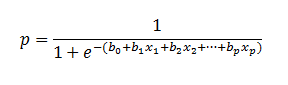
Task 3: Predictive modeling using Regression (5.5 marks)

1. Build a regression model using the default regression method with all inputs. Build another model tuned with GridSearchCV.
2. Explain why did you choose that model

Using logistic regression model because our target variable is binary and input variables are numeric and categorical. Biggest advantage of using logistic regression model is that it is less prone to over-fitting and does fast matrix operations. However, it is feasible with larger relevant dataset. But in our case, positive results are quite less than 50%. Hence prediction will be carried out on class 0: ‘negative’

1. Regression function used

Logistic regression function

****

1. Did you apply standardization of variables? Why would you normalize the variables for regression mining?

* Yes, standardization is necessary to rescale input variables to have mean of 0 and standard deviation of 1 to ensure all are on same scale to compute as with different scale the output is not accurate or preferred.

1. Report on which variables are included in the regression model

* Variables that have positive correlation with the regression model are all the features of regular blood test('Hemoglobin', 'Hematocrit', 'Lymphocytes', 'Leukocytes', 'Basophils', 'Eosinophils', 'Monocytes', 'Neutrophils', 'Proteina C reativa', 'Platelets', 'Mean platelet volume', 'Red blood Cells', 'Red blood cell distribution width', 'Mean corpuscular volume', 'Mean corpuscular hemoglobin', 'Mean corpuscular hemoglobin concentration') and influenza rapid test(A positive and negative, B positive and negative) with considering different age quantiles from 1 to 19.

1. Based on feature importance the five top most important variables are as follows:
   1. Neutrophils: -0.609010429142019
   2. Influenza A rapid test\_negative: 0.5957110216528831
   3. Influenza A rapid test\_positive: -0.5957110216528769
   4. Patient age quantile\_1.0: -0.5830631859063355
   5. Eosinophils: -0.5717910475396188
2. The Train accuracy is: 0.7103658536585366

The Test accuracy is: 0.7357723577235772

1. As the test accuracy is higher than train accuracy there is sign of underfitting than overfitting. Hence, it can be said that model might not capture the required logic of the data.
2. Build another regression model on the reduced variables set. Perform dimensionality reduction with Recursive feature elimination. Tune the model with GridSearchCV
3. Was dimensionality reduction

Yes, dimensionality reduction helps model with less variables gain higher accuracy as variance among classes for samples reduces.

1. Classification accuracy on training and test datasets

The Train accuracy is: 0.5762195121951219

The Test accuracy is: 0.5894308943089431

1. Report any sign of overfitting

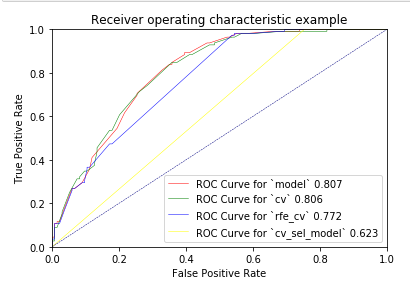
No, there is no sign of overfitting, rather reduction model is slight underfit.

1. Report the top-5 important variables

After reduction of dimensionality the remaining variables are left only 4 features important as per as follows:

* Influenza A rapid test\_negative: 0.3990042303549945
* Influenza B rapid test\_negative: 0.347424494071442
* Patient age quantile\_1.0: 0.2429134177521585
* Eosinophils: 0.010657857821405043
* Mean corpuscular hemoglobin: 0.0

1. Produce the ROC curve for all different regression models.



* Based on roc curve it’s concluded that default curve and curve with cv proves as best models and yes, covid-19 positive patients characteristics are Neutrophils, Eosinophils present, Influenza A rapid test\_negative and Influenza A rapid test\_positive to be true by highly affecting patient age quantile\_1.0.

Task 4: Predictive modeling using Neural Networks (5.5 marks)

1. Build a Neural Network model using the default setting.
2. Explain the parameters used in building this model

* Net architecture: test/train set division: test set 0.3x dataset number of seeds: 10

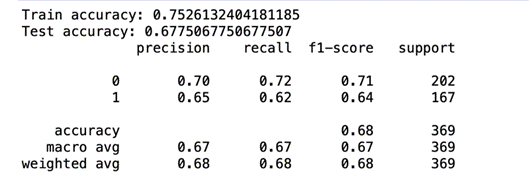


* Iterations: Generally, as we have a large dataset, we expect to have the best result by setting the number of iterations around 100. (the best result achieved using iterations between 70 and 90)
* Activation functions: Relu (rectified linear unit) is used as the activation function. It is the most used activation function in NN, especially in Convolutional Neural Networks (CNNs). It can be used as the default activation function for designed networks when the target activation function is not defined. (usually the first option)

1. Classification accuracy on training and test datasets

The Train accuracy is: 0.7526132404181185

The Test accuracy is: 0.6775067750677507



1. Did the training process converge and resulted in the best model?

* It is expected to see much better results by improving the network (tuning).

1. Refine this network by tuning it with GridSearchCV. Report the trained model.
2. The parameters used in building this model

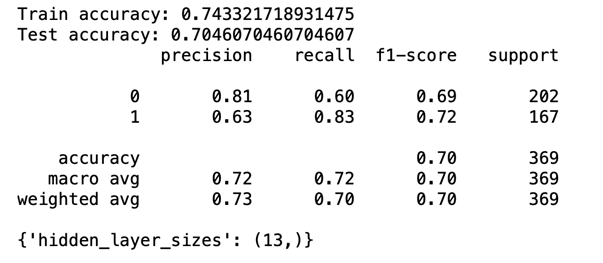
* Net architecture: test/train set division: test set 0.3x dataset

number of seeds: 10

* Iterations: Generally, as we have a large dataset, we set the number of iterations around 100 again. (the best result achieved using iterations between 70 and 90) Also, in order to have more accurate results, we set the batch size to 10. Changing this factor made the model perform %0.3 better on test set.
* Activation functions: Relu (rectified linear unit) is used as the activation function.

1. What is classification accuracy on training and test datasets?

* The results were as followings based on the specifications mentioned above:



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

* It seems that we have achieved the best working model and there is no sign of over-fitting.

Final remarks: Decision making (2 marks)

From the below given data we can see that the model made with regression model that i.e.'cv' has covered maximum curve area and also the index value on test data (0.8090351812366737) hence that will be the best suitable model for this data.

