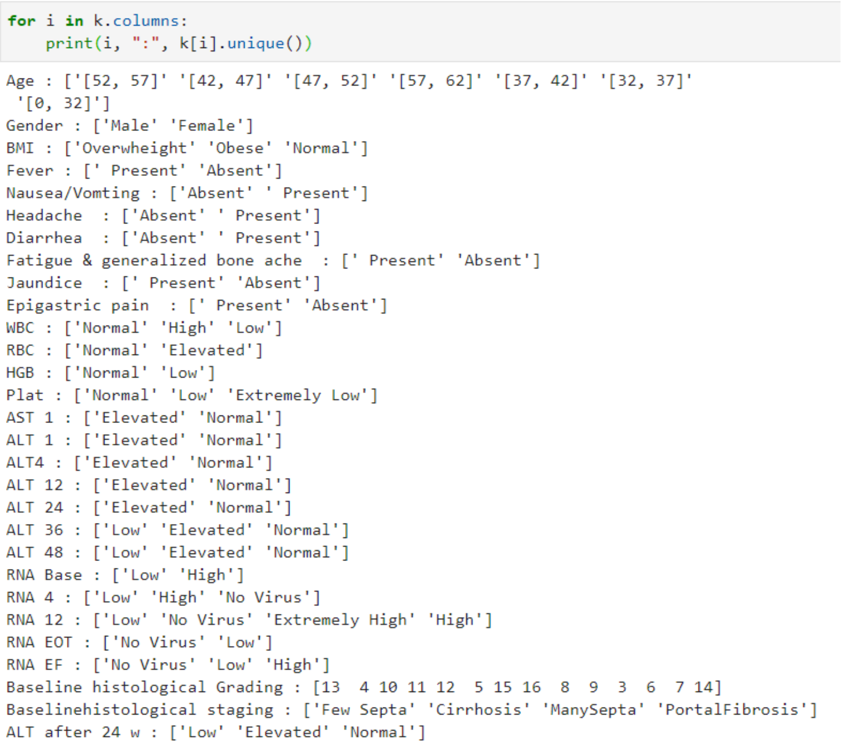
SRIHARSHA DAPARTI
UX46295
DATA 606 – CAPSTONE IN DATA SCIENCE

A VIEW ON HCV DATA

DATA EXPLORATION AND OBSERVATIONS MADE

In the earlier stages of the project, I have concentrated on understanding the crude dataset and the supporting files along with it. It was recognized that the supporting file was a list of Discretization parameters. Using these parameters, I had intended to break the continuity in the data making the data discrete and useful for better modelling.

This delivery marks the end of data exploration and manipulation phase. All the main contributors to the target variable were identified in this stage. After identification of the variables, a much needed background research on what each of the tests like AST, ALT and RNA signify and how their results work. It was known that the AST and ALT are the enzymes released by the river when it is infected. These results tell us how badly the liver is affected. RNA test is the HCV RNA levels in the blood. This test gives the viral load in the blood. Post research, it was necessary to break down the discretized ranges to understandable English terms. One interesting observation was that writing the discretized ranges to disk and reading back makes all the ranges into String parameters. The reason behind it is that when written onto disk as a .csv file, the file format does not support list/range type data points and hence considers them as character inputs. Reading back from disk and converting the already converted strings to another string was much simpler.



```
for i in k.columns:
    print(i, ":", k[i].unique())
```

```
Age : ['[52, 57]' '[42, 47]' '[47, 52]' '[57, 62]' '[37, 42]' '[32, 37]'
 '[0, 32]']
Gender : ['Male' 'Female']
BMI : ['Overwheight' 'Obese' 'Normal']
Fever : [' Present' 'Absent']
Nausea/Vomting : ['Absent' ' Present']
Headache  : ['Absent' ' Present']
Diarrhea  : ['Absent' ' Present']
Fatigue & generalized bone ache  : [' Present' 'Absent']
Jaundice  : [' Present' 'Absent']
Epigastric pain  : [' Present' 'Absent']
WBC : ['Normal' 'High' 'Low']
RBC : ['Normal' 'Elevated']
HGB : ['Normal' 'Low']
Plat : ['Normal' 'Low' 'Extremely Low']
AST 1 : ['Elevated' 'Normal']
ALT 1 : ['Elevated' 'Normal']
ALT4 : ['Elevated' 'Normal']
ALT 12 : ['Elevated' 'Normal']
ALT 24 : ['Elevated' 'Normal']
ALT 36 : ['Low' 'Elevated' 'Normal']
ALT 48 : ['Low' 'Elevated' 'Normal']
RNA Base : ['Low' 'High']
RNA 4 : ['Low' 'High' 'No Virus']
RNA 12 : ['Low' 'No Virus' 'Extremely High' 'High']
RNA EOT : ['No Virus' 'Low']
RNA EF : ['No Virus' 'Low' 'High']
Baseline histological Grading : [13  4 10 11 12  5 15 16  8  9  3  6  7 14]
Baselinehistological staging : ['Few Septa' 'Cirrhosis' 'ManySepta' 'PortalFibrosis']
ALT after 24 w : ['Low' 'Elevated' 'Normal']
```

Figure 1. List of all the unique variables in each attribute .

Coming to the visualizations section, I have started to compare each of the attributes with the target attribute which is the Baseline Histological Staging. The results are as follows:

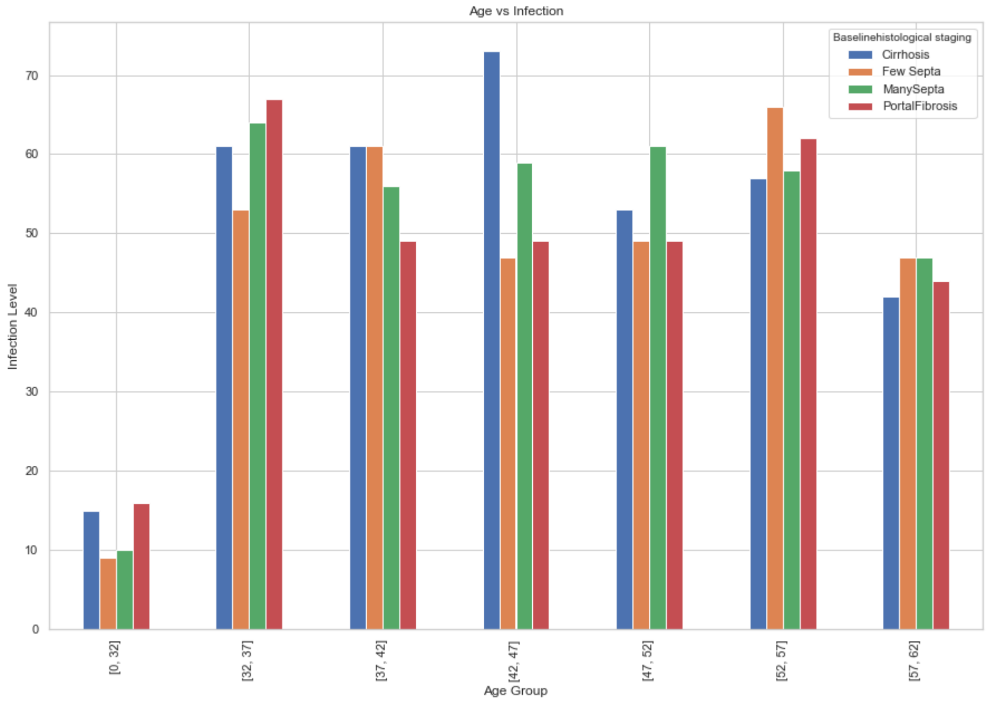


Figure 2. Age group VS Infection in Patients

Observation: Cirrhosis (Advanced level of infection, the stage just before cancer) is observed mostly in the age group of 42yrs to 47yrs.

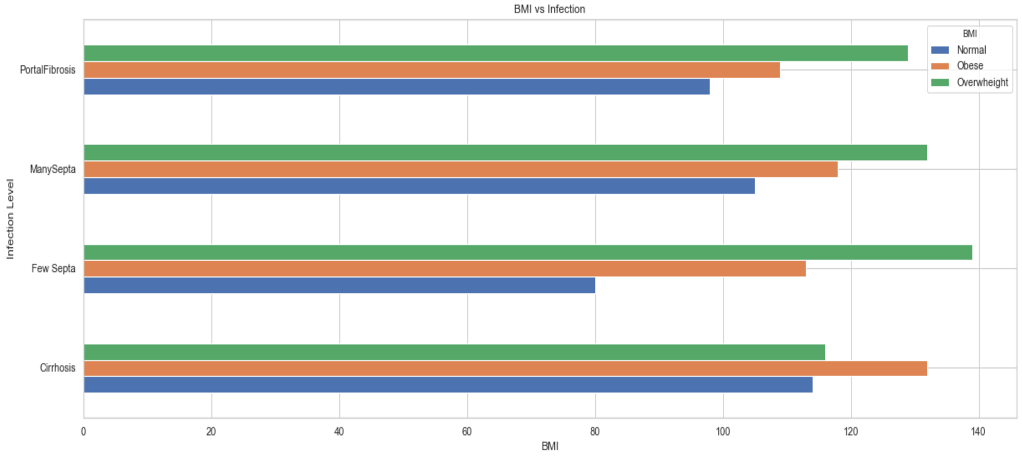


Figure 3. BMI Vs Infection Level in patients

Observation: In the above chart an overall trend can be observed where Overweight people suffer with higher infection levels. One anomaly is that Cirrhotic patients are obese more than they're overweight. Here an assumption can be done that most of the overweight patients have crossed over to the cancer stage along the course of the study and were no longer part of it.

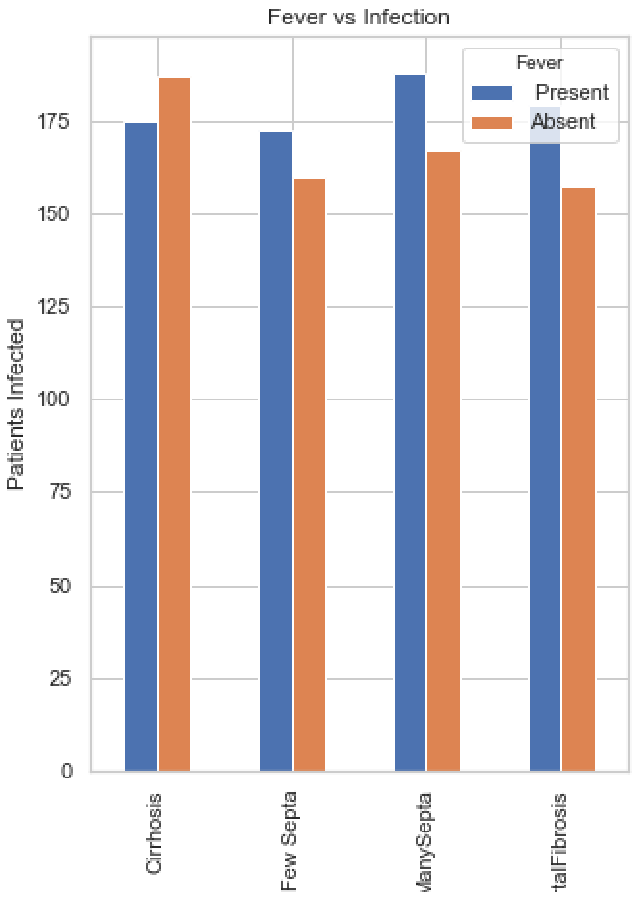


Figure 4. Fever in patients VS Infection Level in patients

Observation: This is a 50-50 case where the infection is causing the fever. A fever is generally caused when the WBC react with the virus. Another reason that causes a spike in temperature is the patients being under constant medication.

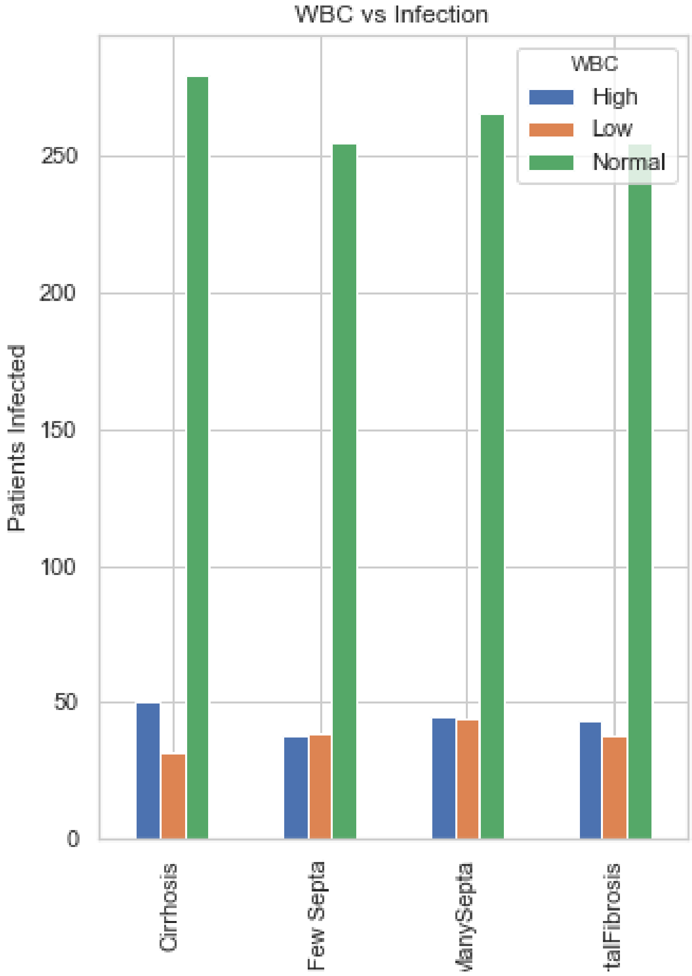


Figure 5. WBC Count VS Infection Level In Patients

Observation:

1. Not much of a correlation between WBC and infected patients / infection level. Most of the patients display normal WBC levels.

2. Resonates with the fact that WBC can't act on HCV like they do on other generic diseases and viruses.

3. Confirms the fact that the medication caused the spike in temperature in the patients.

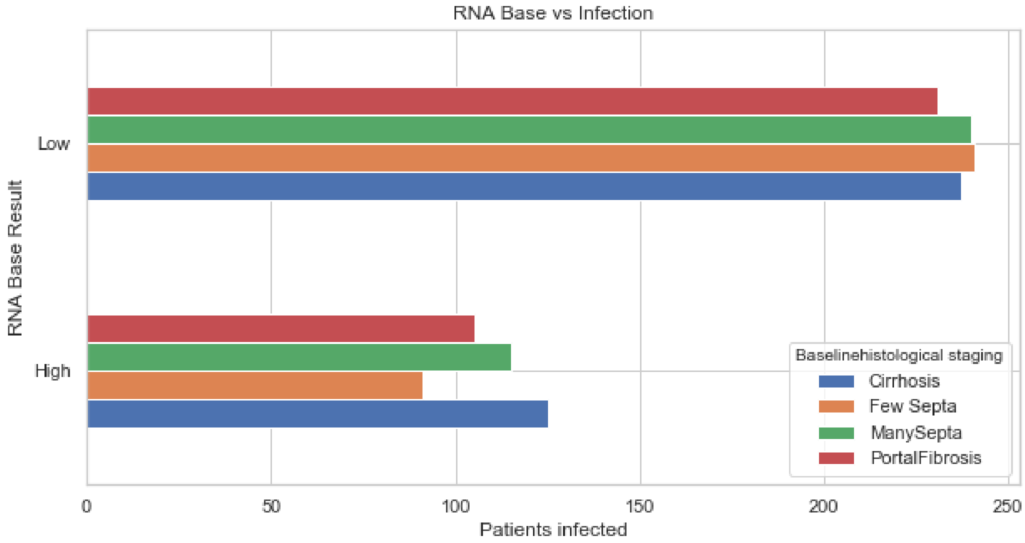


Figure 6. Initial RNA counts at the start of the treatment VS Infection Level in patients

Observation: A trend can be observed right out that most number of patients with the infection have a lower RNA Base result. But if we look closely, among people with low RNA Base result, the infection is less i.e, Few Septa and for those with higher RNA Base result, Cirrhosis is more common.

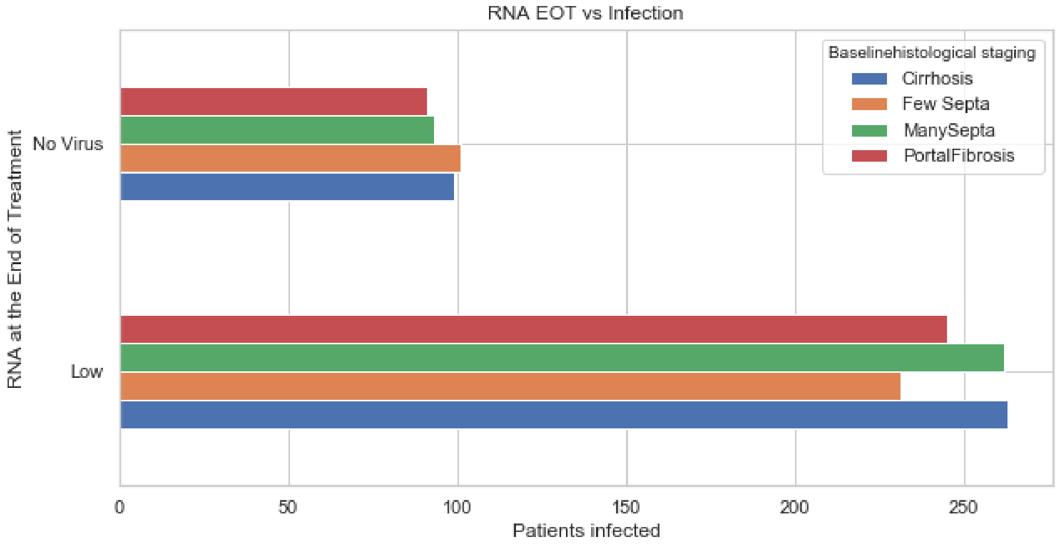


Figure 7. RNA Counts at the End of Treatment VS Infection Level in the patient

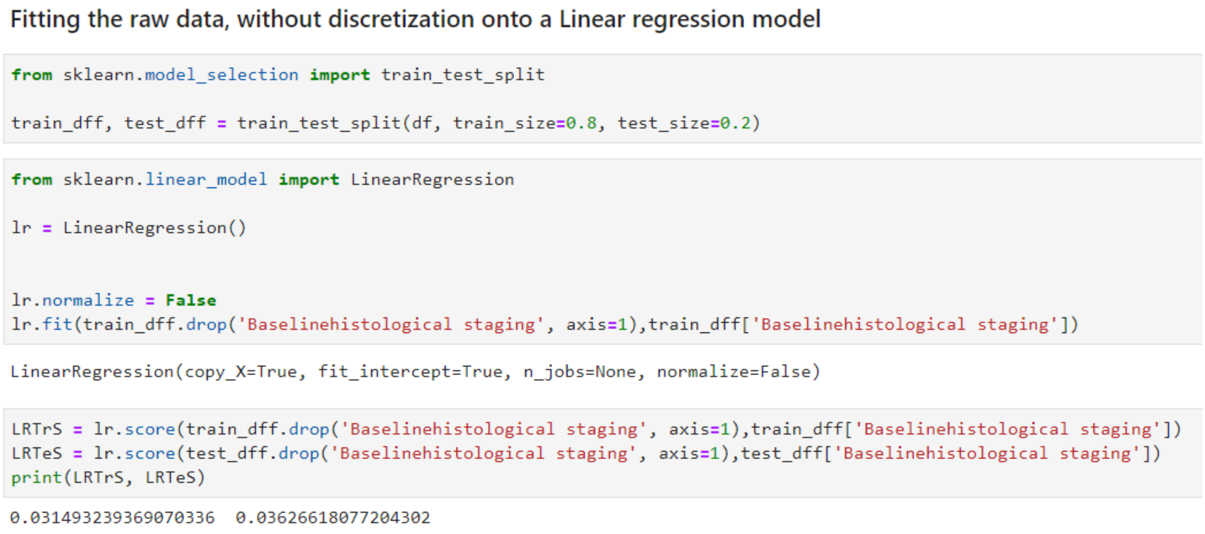
Observation:

1. When compared to the RNA Base Result, the RNA count at the End of Treatment is much lower in the patients.

2. Also, start of the treatment has patients ranging from low to high RNA counts and at the end of treatment, the patients have come down to low and No virus levels with 0% with high RNA counts.

CLASSIFICATION AND MODELLING

First, linear and logistic regression were done on the crude dataset with out the discretization done on it. Both the regressions gave results with around a 30% score of predictability. My assumption would be the data has too much variation for predicting the target. Once we break the continuity and discretize the data, classification and regression analysis will be much fruitful. The challenge now is to deal with the discrete values each of the attributes offer. The hidden advantage of discretization is that the target variable can be varied as the data becomes all categorical and multiclass in nature.



Fitting the raw data, without discretization onto a Linear regression model

```
from sklearn.model_selection import train_test_split

train_dff, test_dff = train_test_split(df, train_size=0.8, test_size=0.2)
```

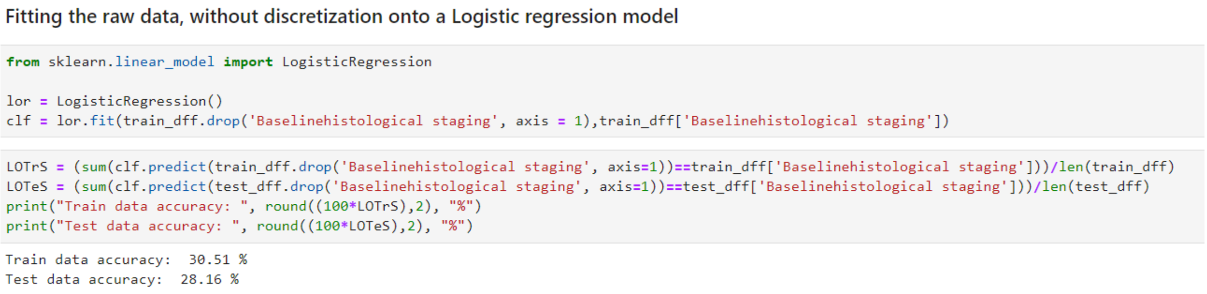
```
from sklearn.linear_model import LinearRegression

lr = LinearRegression()


lr.normalize = False
lr.fit(train_dff.drop('Baselinehistological staging', axis=1),train_dff['Baselinehistological staging'])
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

```
LRTrS = lr.score(train_dff.drop('Baselinehistological staging', axis=1),train_dff['Baselinehistological staging'])
LRTeS = lr.score(test_dff.drop('Baselinehistological staging', axis=1),test_dff['Baselinehistological staging'])
print(LRTrS, LRTeS)
```

```
0.031493239369070336  0.03626618077204302
```

Figure 8. Linear regression on Raw Data



Fitting the raw data, without discretization onto a Logistic regression model

```
from sklearn.linear_model import LogisticRegression

lor = LogisticRegression()
clf = lor.fit(train_dff.drop('Baselinehistological staging', axis = 1),train_dff['Baselinehistological staging'])
```

```
LOTrS = (sum(clf.predict(train_dff.drop('Baselinehistological staging', axis=1))==train_dff['Baselinehistological staging']))/len(train_dff)
LOTeS = (sum(clf.predict(test_dff.drop('Baselinehistological staging', axis=1))==test_dff['Baselinehistological staging']))/len(test_dff)
print("Train data accuracy: ", round((100*LOTrS),2), "%")
print("Test data accuracy: ", round((100*LOTeS),2), "%")
```

```
Train data accuracy:  30.51 %
Test data accuracy:  28.16 %
```

Figure 9. Logistic regression on Raw Data

CONCLUSION AND FUTURE WORK

As mentioned in the observation part of the report, all the attributes contributing to the target are recognized. Discretization seems very promising and better results on the new and more discrete data is expected. A raw model is already scripted and with little changes and trifle needed it should be able to work with a better score. In the next delivery, I intend to deliver the working model trained and tested on the discretized data. Also, the end goal would be test and play with more fitting algorithms such as Decision Trees and Random forests. A consolidated score panel with respect to each of the algorithms can be expected by using a pipeline model to determine the best one of them all.