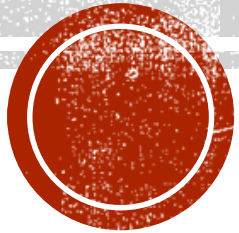
DATA 606 CAPSTONE IN DATA SCIENCE

A VIEW ON HCV DATA

SRIHARSHA DAPARTI

UX46295

Instructor : DR. ERGUN SIMSEK



RECAP

- The dataset consists of HCV(Hepatitis C Virus) infected patients in Egypt who were a part of a study. This data was obtained from UCI Machine Learning Repository.

  Citation: Dua, D. and Graff, C. (2019). Link.
- The dataset has two files. The first is the dataset itself which shows the anonymous records of Egyptian patients who underwent treatment dosages for HCV about 18 months and the second file contains the discretization parameters for each and every attribute in the first file.
- The dataset contains about 1000 patient records with 29 attributes such as age, bmi, symptoms, test results etc, for each record explaining the treatment.
- At the end of about three weeks from the project decision, a repetitive and considerable amount of cleaning and viewing has been done on the data set for it to be fit for further analysis.
- Discretization was a major chunk of the problem, Discretizing the attributes of the data set one by one and breaking the continuity of the data, due to their varied nature, has been an issue.

WORK DONE AND OBSERVATIONS

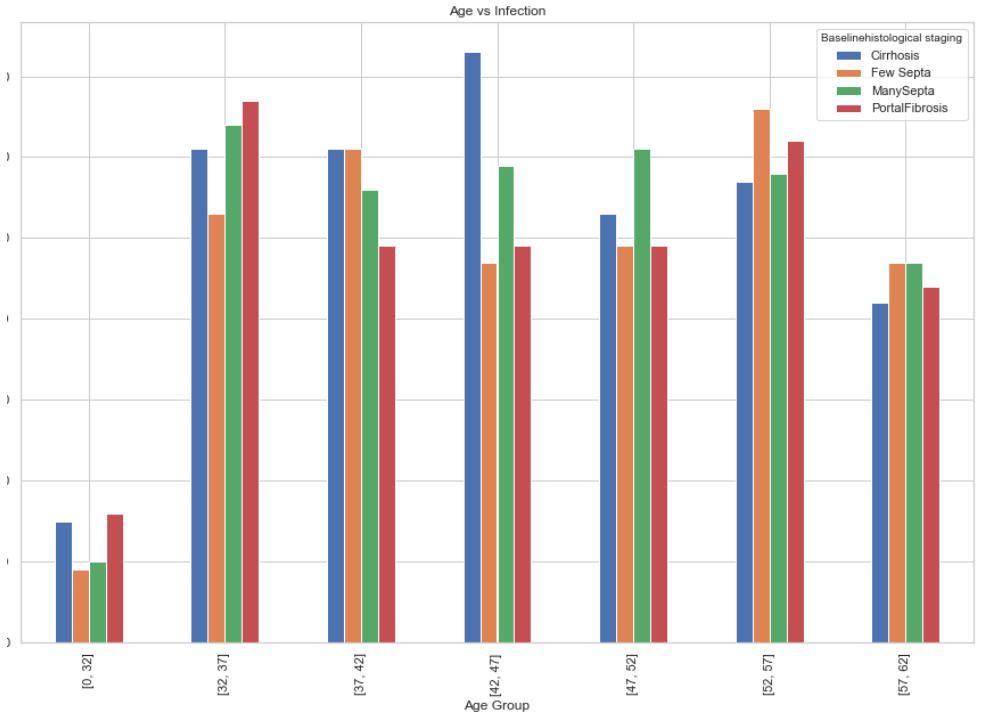
End of all major data manipulation and exploration chunk.

Background research on all the attributes and understanding the meanings of the tests and result brackets was crucial.

Converted all range parameters into simple English terms making the data discrete yet understandable

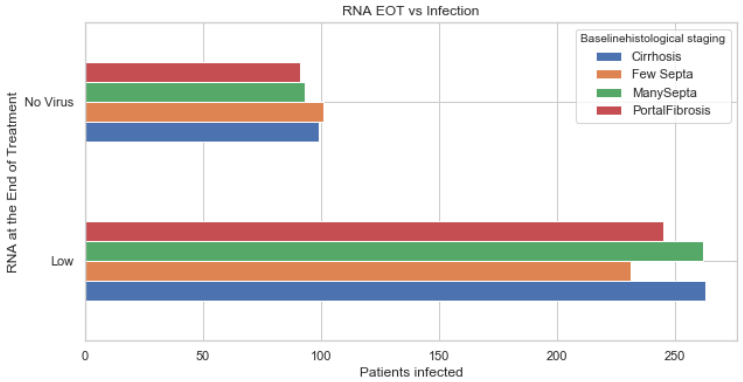
Visualizations between the attributes and the supposed target entity have been plotted. Using these observations, some important insights were made pointing towards the importance of features.

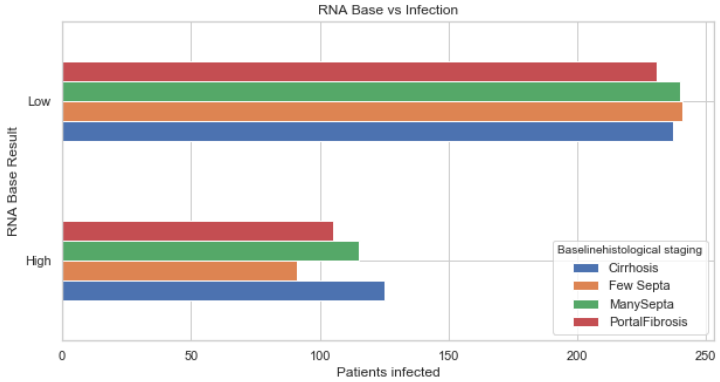
CONTD.



- One of the visualizations is this Age Group VS Infection level in patients graph.
- It shows that the age group of people with most infection level and count were in the range of 42 to 47 years.

CONTD.





- These two charts show the **HCV RNA count** in the blood (**signifies the amount of virus in the blood**) at the start and end of the treatment.
- We can observe that the counts at the **start of the treatment** were **high – low** where as at the ***end of treatment*** were ***low - no virus*** at all.

MODELLING

Linear regression performed on the crude data gave a training accuracy score of 31% and a test score of around 35%.

Logistic Regression came out with a consistent 30% in both training and testing situations.

My assumption would be the data has too much variation for predicting the target.

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.

The next algorithms expected to work better are Decision Trees and random forests

A comparison between performances the raw, continuous data and discretized data will help give better insights.

CONCLUSION AND FUTURE WORK

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.

In the next delivery, I intend to deliver the working model trained and tested on the discretized data.

The end goal would be test and play with more fitting algorithms such as Decision Trees and Random forests.

Planning to also present a consolidated score panel with the help of pipelining and suggesting a suitable algorithm for the datasets of the chosen kind.