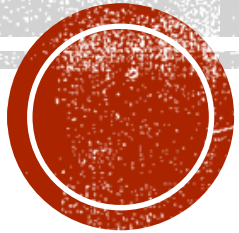
DATA 606 CAPSTONE IN DATA SCIENCE

A VIEW ON HCV DATA

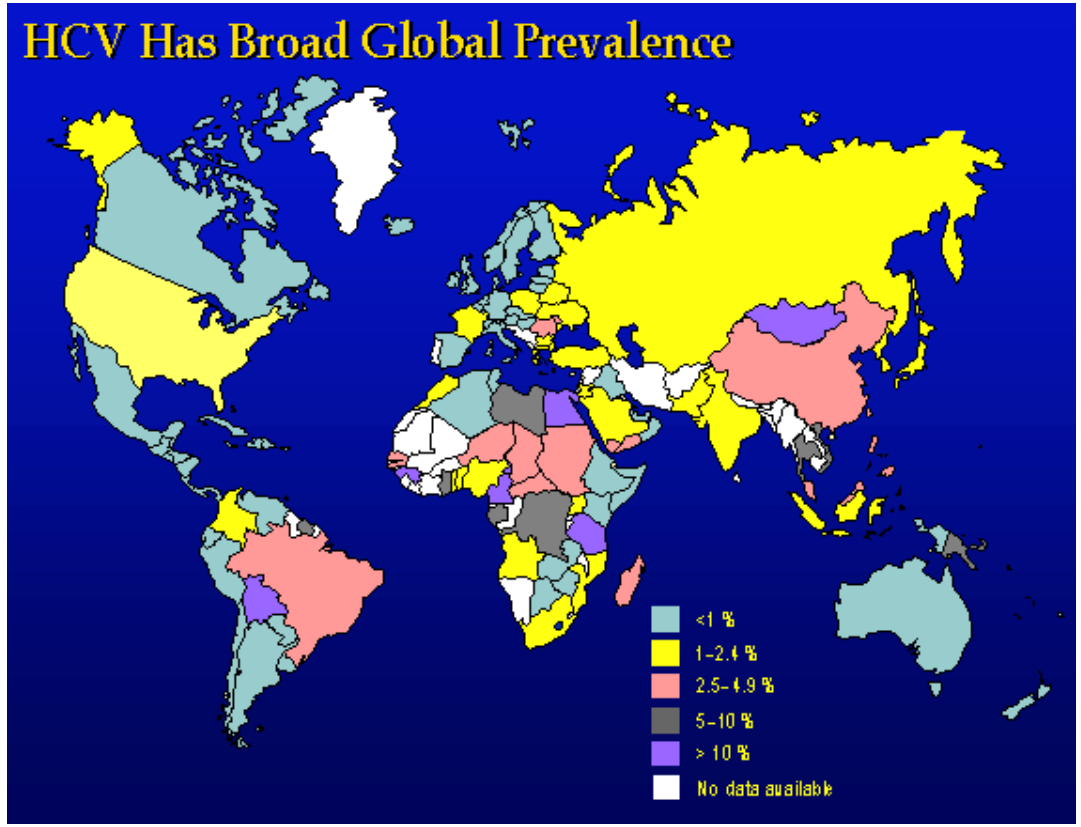
SRIHARSHA DAPARTI

UX46295

Instructor : DR. ERGUN SIMSEK



INTRODUCTION



- Hepatitis C virus (HCV) is a major health problem worldwide. In 2015, the global prevalence of HCV infection was 1.0%, with the highest prevalence in the Eastern Mediterranean Region (2.3%) followed by the European one (1.5%). The annual mortality due to HCV-related complications is estimated to be approximately 700000 deaths.

ABOUT THE DATASET

- **Hepatitis C Virus (HCV) for Egyptian patients.** 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 for each record explaining the treatment .
- The attributes of the patient records are:

• Age	• WBC Count – White Blood Cell Count	• ALT 48 – ALT Week 48
• Gender	• RBC Count - Red Blood Cell Count	• ALT after 24 w – ALT after 24 weeks
• BMI	• HGB - Haemoglobin	• RNA Base
• Fever	• Platelets Count	• RNA 4
• Nausea/Vomiting	• AST 1 – Aspartate Transaminase ratio	• RNA 12
• Headache	• ALT 1 – Alanine Transaminase ratio Week 1	• RNA EOT – RNA at End Of Treatment
• Diarrhea	• ALT 4 – ALT Week 4	• RNA EF – RNA Elongation Factor
• Fatigue & Generalized bone ache	• ALT 12 – ALT Week 12	• Baseline Histological Grading
• Jaundice	• ALT 24 – ALT Week 24	• Baseline Histological Staging
• Epigastric Pain	• ALT 36 – ALT Week 36	

LITERATURE REVIEW

- The research on how to tackle the presented problem and the quest for similar problems and the methodologies used to solve those problems was challenging. The primary paper1 that used this data set was thoroughly examined along with the purpose of the research, approach and methodologies. The paper talks about coming up with 'Rules' which are combinations of two or more symptoms observed. The intention of this project is to take a more broadened approach. In this paper, the expected findings are the study of each individual symptom correlated with the test results.
- Many other works mentioning health data were also examined and a classification approach is decided to be the best way to tackle this question.

DATA CLEANING

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, due to their varied nature has been an issue.

The correlation between different attribute was visualized using a color coded Correlation Matrix plotted using ‘matplotlib’ package

EXPLORATOTY ANALYSIS

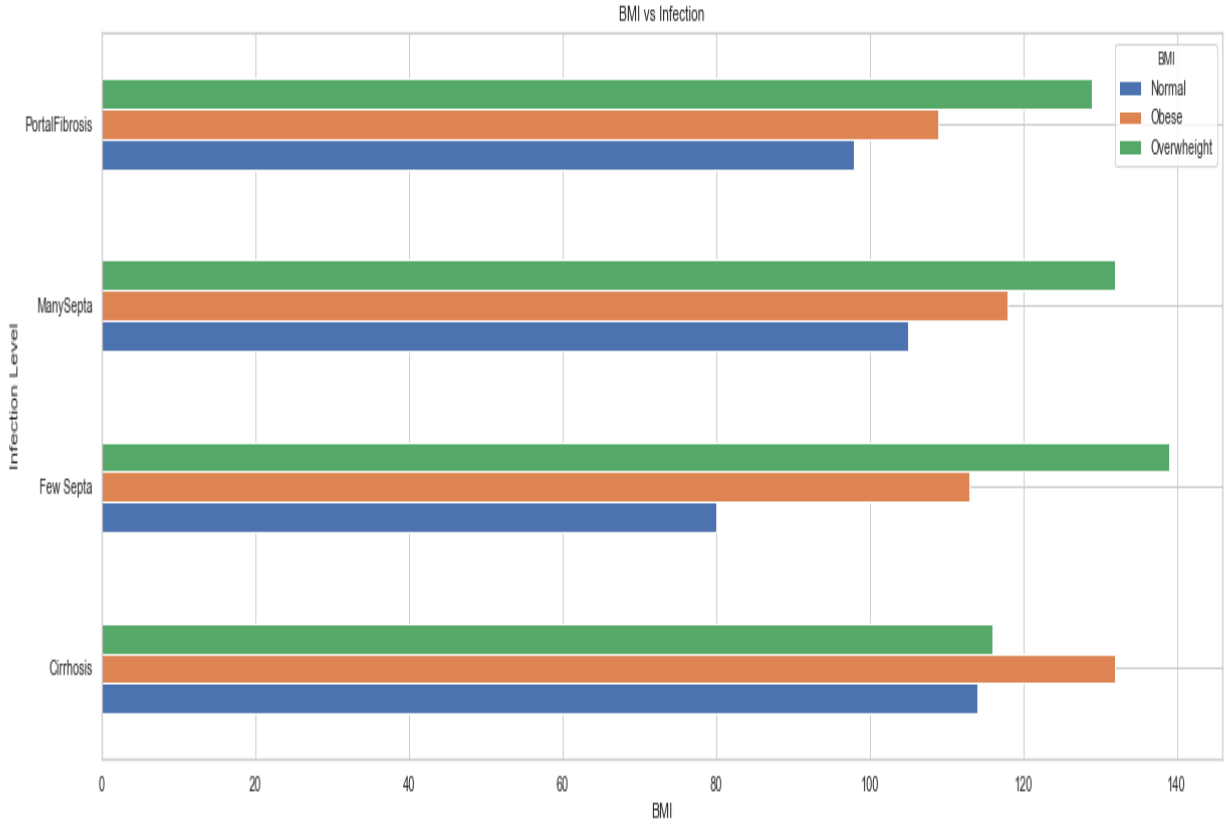
End of all major data manipulation and exploration chunk.

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

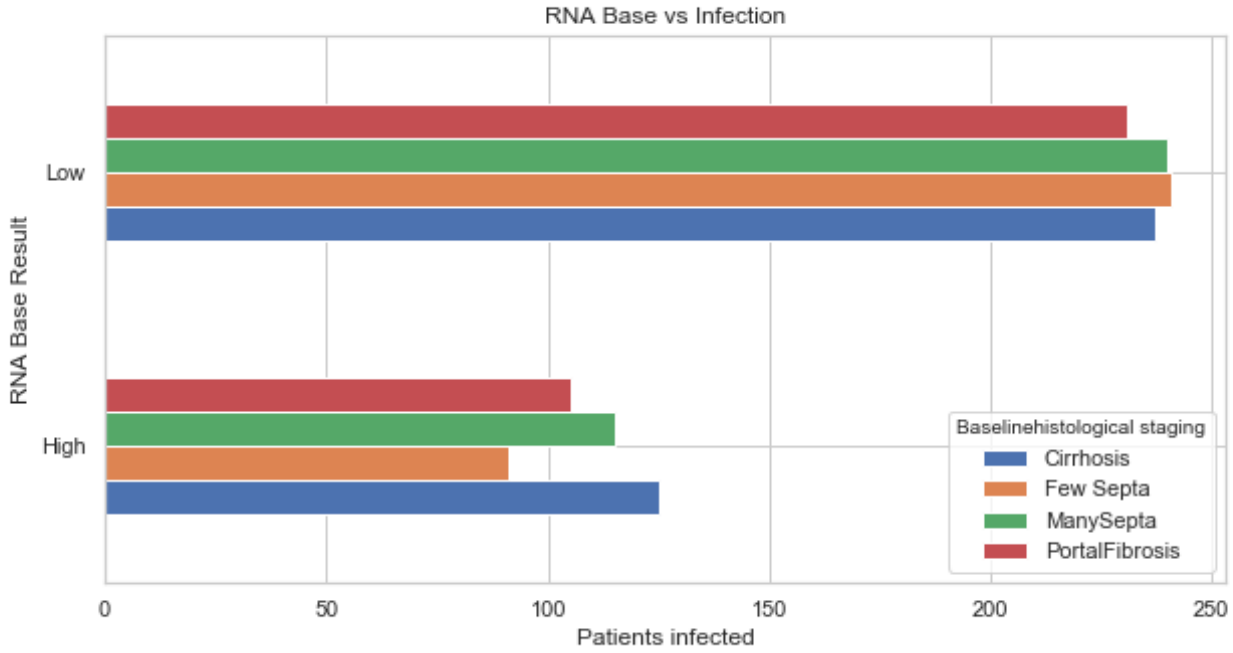
Visualizations between the attributes and the supposed target entity have been plotted. Using these observations, some important insights were made pointing towards the importance of features.

EXPLORATORY ANALYSIS



- 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.
- Possible Assumption: 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.

EXPLORATORY ANALYSIS



- A trend can be observed right out that most number of patients with the infection have a lower RNA Base result.
- 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.

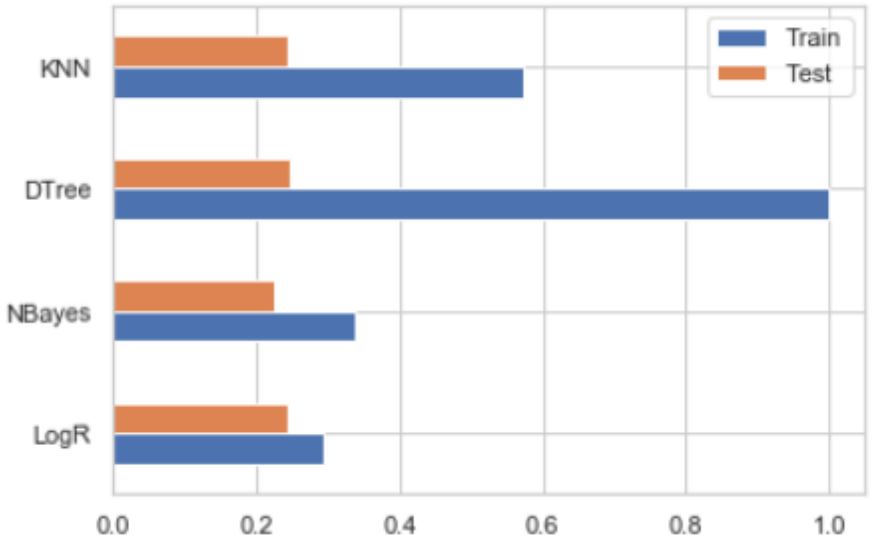
```
Logistic Train data accuracy:  29.29 %
Logistic Test data accuracy:  24.21 %

NaiveBayes(G) Train data accuracy:  33.62 %
NaiveBayes(G) Test data accuracy:  22.48 %
NaiveBayes(G) Cross val score:  0.24691898654989056

Decision Tree Train data accuracy:  100.0 %
Decision Tree Test data accuracy:  24.78 %
Decision Tree Cross val score:  0.24192993431341883

KNN Training set accuracy:  57.32 %

KNN Testing set accuracy:  24.21 %
<matplotlib.axes._subplots.AxesSubplot at 0x28f375c9fd0>
```



CLASSIFICATION & MODELLING

- Discrete data calls for classification rather than regression
- Train - Test split is 75:25
- Comparing the test accuracies of all four models,
  - KNN shows to give a better prediction on the training set but fails to provide the same with test set.
  - Decision trees classifier gives 100% prediction score on training set which questions overfitting but similar to KNN and other methods it also gives lower scores, around 24-25%, on the test set.

CLASSIFICATION AND MODELLING

Next steps were Random Forests and Extremely randomized trees(Extra Tree Classifier)

Reason to choose was the score of Decision tree classifier which was more than any other classification technique

They both gave scores of 26.51% and 24.20%

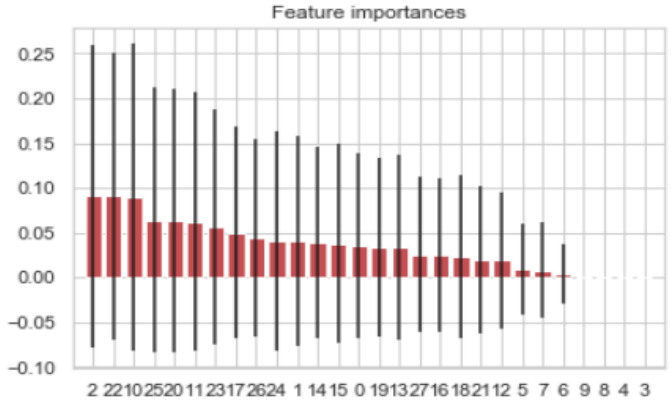
Consistently low accuracies since the start of classification.

Usual rule of thumb is prediction accuracies of Decision Trees < Random Forests < Extra Tress. Clearly, the scores are out of fashion.

Feature extraction is done to bring out the most important features

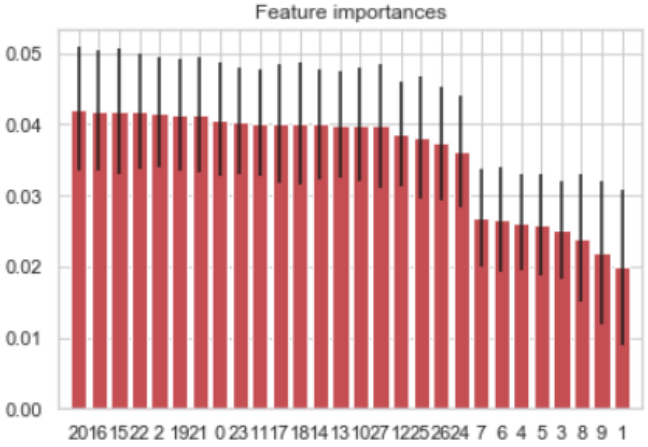
CLASSIFICATION & MODELLING

```
Feature importance order:  ['BMI', 'RNA Base', 'WBC', 'RNA EOT', 'ALT 48', 'RBC', 'RNA 4', 'ALT 12', 'RNA EF', 'RNA 12', 'Gender', 'AST 1', 'ALT 1', 'Age ', 'ALT 36',
'Plat', 'Baseline histological Grading', 'ALT4', 'ALT 24', 'ALT after 24 w', 'HGB', 'Headache ', 'Fatigue & generalized bone ache ', 'Diarrhea ', 'Epigastric pain ',
'Jaundice ', 'Nausea/Vomting', 'Fever']
```



Random Forrest Classifier –
Feature extraction and ordering

```
Feature importance order:  ['ALT 48', 'ALT4', 'ALT 1', 'RNA Base', 'BMI', 'ALT 36', 'ALT after 24 w', 'Age ', 'RNA 4', 'RBC', 'ALT 12', 'ALT 24', 'AST 1', 'Plat', 'WB
C', 'Baseline histological Grading', 'HGB', 'RNA EOT', 'RNA EF', 'RNA 12', 'Fatigue & generalized bone ache ', 'Diarrhea ', 'Nausea/Vomting', 'Headache ', 'Fever', 'Ja
undice ', 'Epigastric pain ', 'Gender']
```



Extra Tree Classifier –
Feature extraction and ordering

CLASSIFICATION & MODELLING

Later steps included Gradient boosting trees and Extreme Gradient Boosting

Reason to choose was to reduce loss of data while modelling(if any)

Both techniques show consistent scores with the previous algorithms

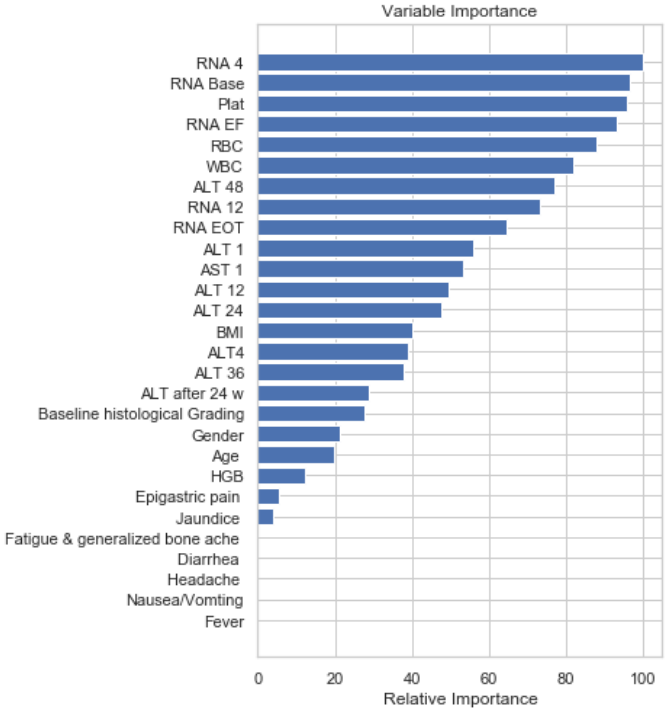
Scores are 25.07% and 24.50%

Problem found to be less number of data points(or patient records)

Feature extraction was done in both cases again

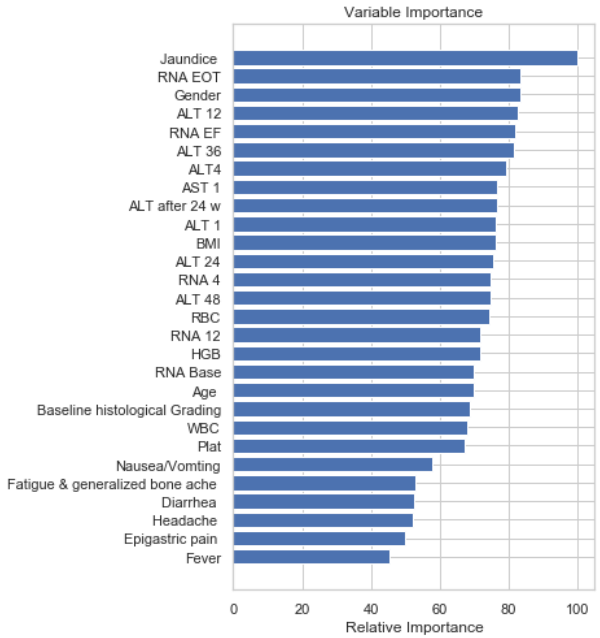
CLASSIFICATION & MODELLING

```
GBC:  0.2507204610951009
GBC Feature Importance Order:
 ['RNA 4', 'RNA Base', 'Plat', 'RNA EF', 'RBC', 'WBC', 'ALT 48', 'RNA 12', 'RNA EOT', 'ALT 1', 'AST 1', 'ALT 12', 'ALT 24', 'BMI', 'ALT4', 'ALT 36', 'ALT after 24 w', 'Baseline histological Gradin
g', 'Gender', 'Age ', 'HGB', 'Epigastric pain ', 'Jaundice ', 'Fatigue & generalized bone ache ', 'Diarrhea ', 'Headache ', 'Nausea/Vomting', 'Fever']
```



Gradient Boosting Classifier – Feature extraction and ordering

```
Accuracy: 24.50%
XGB Feature Importance Order:
 ['Jaundice ', 'RNA EOT', 'Gender', 'ALT 12', 'RNA EF', 'ALT 36', 'ALT4', 'AST 1', 'ALT after 24 w', 'ALT 1', 'BMI', 'ALT 24', 'RNA 4', 'ALT 48', 'RBC', 'RNA 12', 'HGB', 'RNA Base', 'Age ', 'Basel
ine histological Grading', 'WBC', 'Plat', 'Nausea/Vomting', 'Fatigue & generalized bone ache ', 'Diarrhea ', 'Headache ', 'Epigastric pain ', 'Fever']
```



Extreme Gradient Boosting Classifier – Feature extraction and ordering

RESULTS

The first two techniques gave features with effective splitting of trees, the later worked on loss reduction.

Pooling was done by selecting the top 5 features from the results of each of the techniques

The most prominent features that contributed most to Hepatitis C disease levels are BMI, RNA Base, ALT 48, RNA EF.

CONCLUSION

- With the help of observations made in the EDA and feature extraction using ensemble classifiers, the following facts can be established.
  - BMI (Body Mass Index) of the patient is more important than attributes like age and gender where trends are usually seen most of the time.
  - RNA Base refers to the one or more of the 4 RNA bases that make up the human RNA structure. The more the RNA Base count, the more the patient is likely to suffer infection as it facilitates the stabilization of the virus in the body.
  - ALT 48 - High levels of ALT may indicate liver damage from hepatitis, infection, cirrhosis, liver cancer, or other liver diseases. This indicates that at the end of 48 weeks, if the patients ALT levels are low, then the chance of the disease to progress are next to none.
  - RNA EF(Elongation Factor) shows that increased RNA EF factor is directly proportional to the rate of replication of the virus which leads to the progress of the infection in the patients.

Thank You