SRIHARSHA DAPARTI

DATA 606 – CAPSTONE IN DATA SCIENCE

SPRING 2020

INSTRUCTOR – Dr. Ergun Simsek

A VIEW ON HCV DATA

AIM

The aim of the project is to delve into the Hepatitis C Virus (HCV) for Egyptian patients dataset thoroughly and gain as much information as possible. Along this process, I expect the data to throw me problems which I intend to solve in the span of this semester. The goal is to develop a predictive model or suggest one that exists which will be able to predict the progression of the disease in a given patient.

MOTIVATION

The sensitivity towards the suffering has always been a motivating factor for me. In addition to this, I had a first hand experience with a HCV case in the family. The progression of the disease in the patient was unexpected and shockingly intriguing. When I pursued the progression of the disease that is when I have come across the data set. Even with as low as 1000 patient records, the number of factors involving in the patient health and stage of infection drove me to work on this data set.

DATASET

Hepatitis C Virus (HCV) for Egyptian patients. This data was obtained from UCI Machine Learning Repository. Citation: Dua, D. and Graff, C. (2019). 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
- BMI
- Nausea/Vomiting
- Diarrhea
- Jaundice
- WBC Count – White Blood Cell Count
- HGB - Haemoglobin
- AST 1 – Aspartate Transaminase ratio
- ALT 4 – ALT Week 4
- ALT 24 – ALT Week 24
- ALT 48 – ALT Week 48
- RNA Base
- RNA 12
- RNA EF – RNA Elongation Factor
- Baseline Histological Staging
- Gender
- Fever
- Headache
- Fatigue & Generalized bone ache
- Epigastric Pain
- RBC Count - Red Blood Cell Count
- Platelets Count
- ALT 1 – Alanine Transaminase ratio Week 1
- ALT 12 – ALT Week 12
- ALT 36 – ALT Week 36
- ALT after 24 w – ALT after 24 weeks
- RNA 4
- RNA EOT – RNA at End Of Treatment
- Baseline Histological Grading

Data Link: Hepatitis C Virus (HCV) for Egyptian patients.

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 paper[1] 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[2] mentioning health data were also examined and a classification approach is decided to be the best way to tackle this question.

DATA CLEANING AND WORK DONE

Data Cleaning has been a very important and excruciatingly painful part of the project. Upon the review of the primary research publication that made use of this data set, it has been confirmed that the data set available at the data source has been mangled with. Working with such a dataset is very troublesome. 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. Going further deep into the hurdles presented during this phase, 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. With the help of a few supporting functions, a ‘refined’ discretized data set was produced. Following this cleaning process, the correlation between different attribute was visualized using a color coded Correlation Matrix plotted using ‘matplotlib’ package to get an idea of which direction this problem is supposed to be handled. This has shown that RNA levels at the end of 12 weeks and at the End of Treatment has shown to be playing an important role. The next step is to check which other attributes (symptoms) collide with this attribute to result in the said Baseline Histological Staging (whether the liver has Fewer Fibroids, Many Septa or is Cirrhotic).

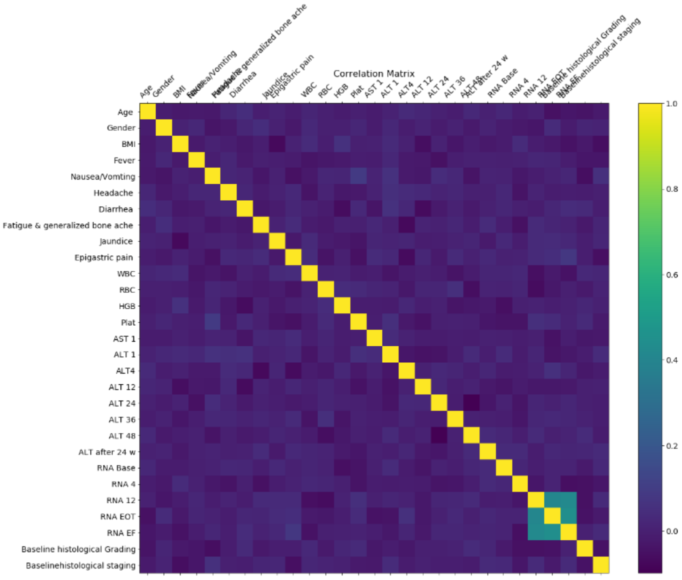
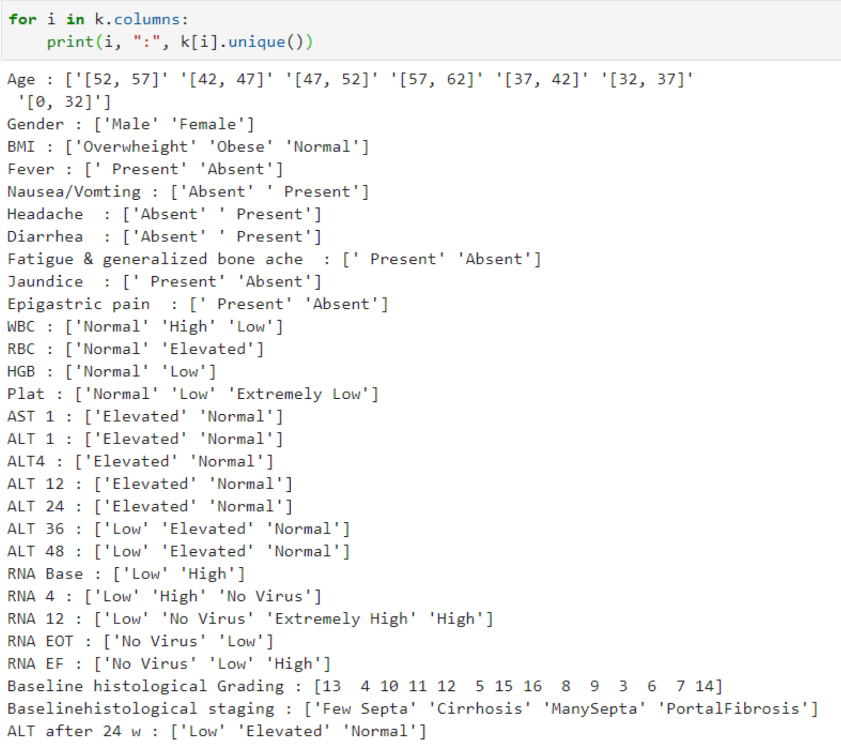


Figure 1. Correlation between all features.

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 2. 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:

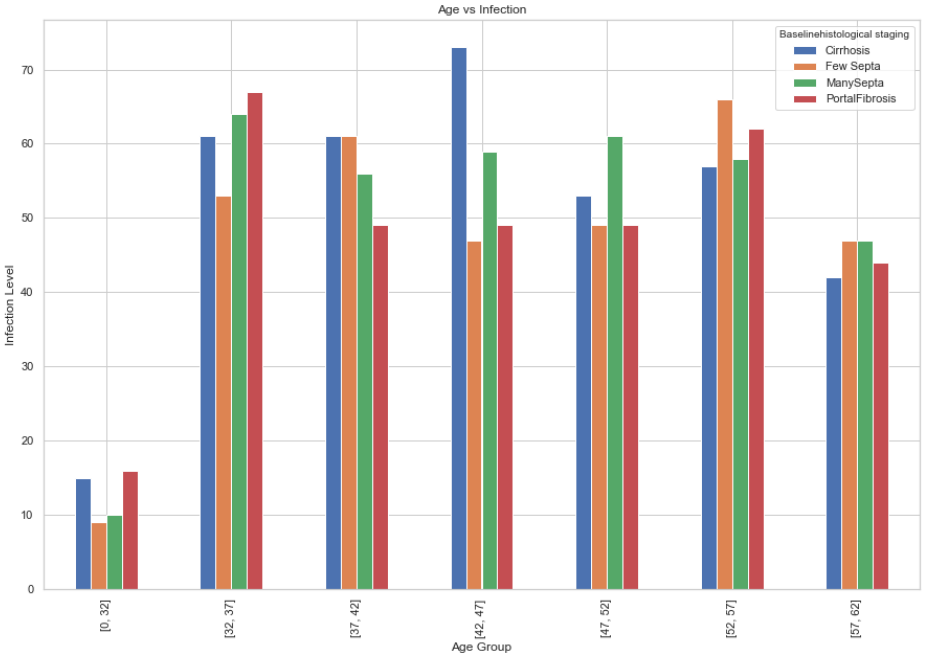


Figure 3. Age group VS Infection in Patients

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

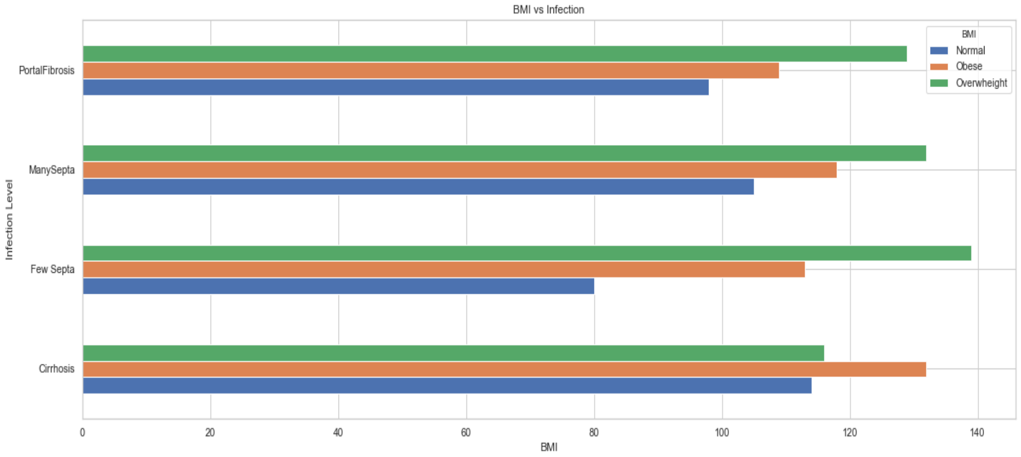


Figure 4. BMI Vs Infection Level in patients

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.

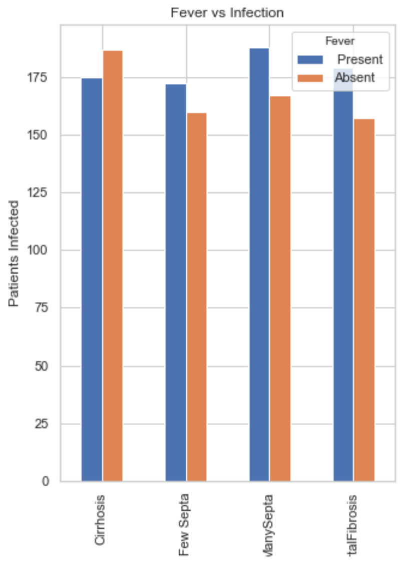


Figure 5. Fever in patients VS Infection Level in patients

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.

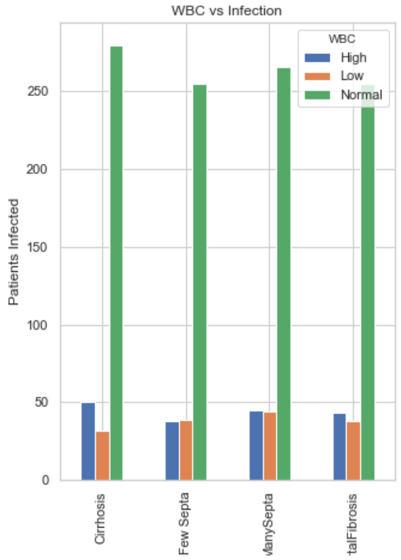


Figure 6. WBC Count VS Infection Level In Patients

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.

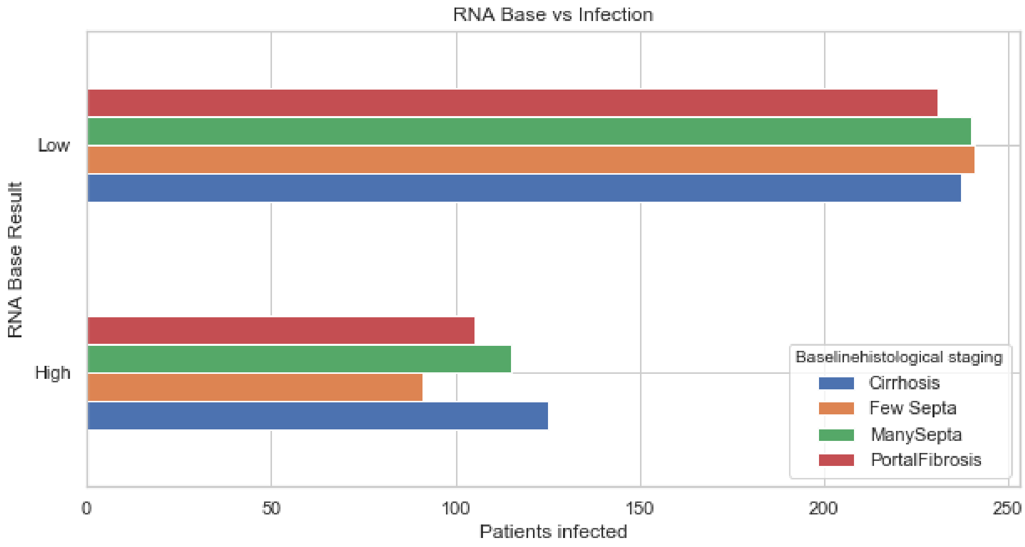


Figure 7. Initial RNA counts at the start of the treatment VS Infection Level in patients

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.

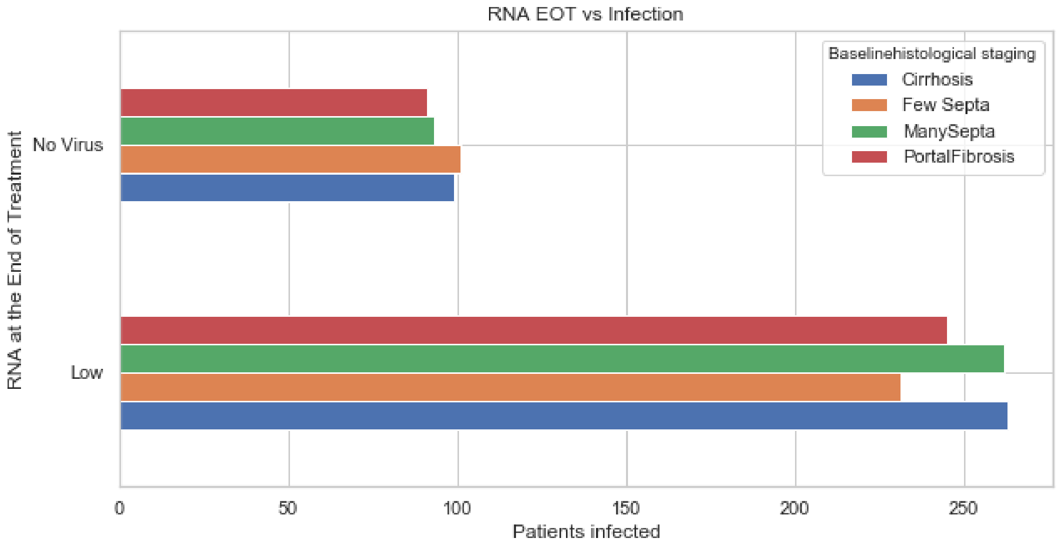


Figure 8. RNA Counts at the End of Treatment VS Infection Level in the patient

Observation:

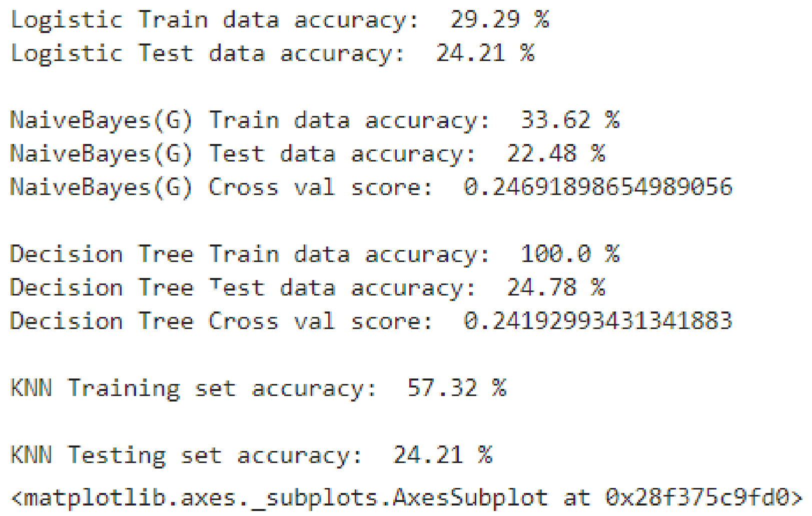
1. When compared to the RNA Base Result, the RNA count at the End of Treatment is much lower in the patients.

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

The data was discrete in nature which pushed me to go with classification techniques rather than regression algorithms. To apply machine learning on the discretized dataset, it needs to be encoded to simpler numerical values with the help of tools like OneHotEncoder etc. But the result of this encoding will come out similar to be the initial raw dataset. Therefore, to avoid redundancies and cut down runtimes, all the chosen machine learning techniques have been applied on the raw dataset and not on the discretized dataset.

First, logistic regression was performed on the dataset which gave a prediction score of 24.21% with a 75-25 train-test split. With the logistic regression scores so low, the classification was performed on Gaussian Naïve Bayes Classifier, Decision Trees Classifier and K-Nearest Neighbor Classifier. Their results were as follows:



```
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>
```

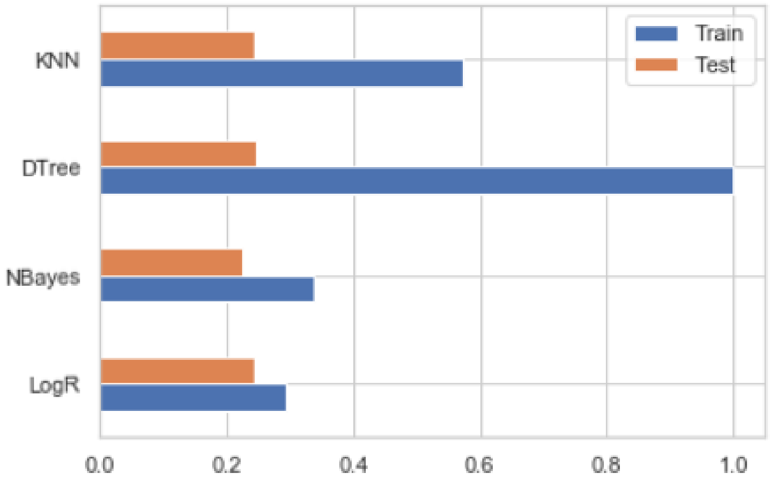


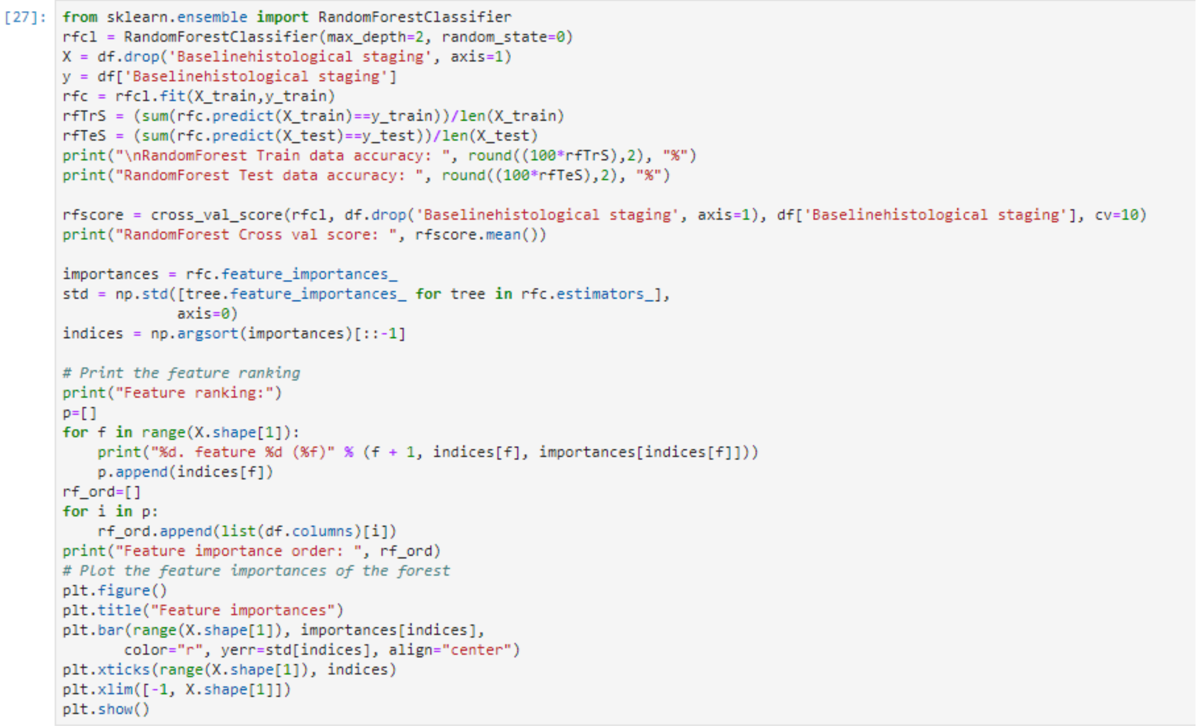
Figure 9. Train and test accuracies of chosen models with a comparison bar chart.

It was observed that Decision trees was giving a somewhat better score when compared to the other classification methods. Points to be noted:

1. KNN shows to give a better prediction on the training set but fails to provide the same with test set.

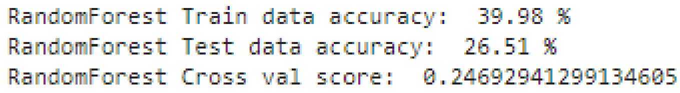
2. Decision trees classifier gives 100% prediction score on training set which questions overfitting but similar to KNN and other methods it also gives lower scores, around 24-25%, on the test set.

As decision tree came out on top among the previous techniques, Random forests was picked with the expectation to do a better job. The implementation and results are shown below:



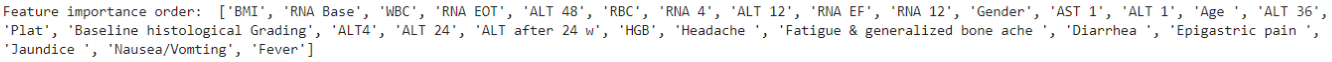
```
[27]: from sklearn.ensemble import RandomForestClassifier
      rfcl = RandomForestClassifier(max_depth=2, random_state=0)
      X = df.drop('Baselinehistological staging', axis=1)
      y = df['Baselinehistological staging']
      rfc = rfcl.fit(X_train,y_train)
      rfTrS = (sum(rfc.predict(X_train)==y_train))/len(X_train)
      rfTeS = (sum(rfc.predict(X_test)==y_test))/len(X_test)
      print("\nRandomForest Train data accuracy: ", round((100*rfTrS),2), "%")
      print("RandomForest Test data accuracy: ", round((100*rfTeS),2), "%")

      rfscore = cross_val_score(rfcl, df.drop('Baselinehistological staging', axis=1), df['Baselinehistological staging'], cv=10)
      print("RandomForest Cross val score: ", rfscore.mean())

      importances = rfc.feature_importances_
      std = np.std([tree.feature_importances_ for tree in rfc.estimators_],
                   axis=0)
      indices = np.argsort(importances)[::-1]

      # Print the feature ranking
      print("Feature ranking:")
      p=[]
      for f in range(X.shape[1]):
          print("%d. feature %d (%f)" % (f + 1, indices[f], importances[indices[f]]))
          p.append(indices[f])
      rf_ord=[]
      for i in p:
          rf_ord.append(list(df.columns)[i])
      print("Feature importance order: ", rf_ord)
      # Plot the feature importances of the forest
      plt.figure()
      plt.title("Feature importances")
      plt.bar(range(X.shape[1]), importances[indices],
             color="r", yerr=std[indices], align="center")
      plt.xticks(range(X.shape[1]), indices)
      plt.xlim([-1, X.shape[1]])
      plt.show()
```

Figure 10. Implementation of Random Forest Classifier - Prediction and Feature extraction



```
RandomForest Train data accuracy:  39.98 %
RandomForest Test data accuracy:  26.51 %
RandomForest Cross val score:  0.24692941299134605
```

Figure 11. Random Forest Classifier accuracies



```
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']
```

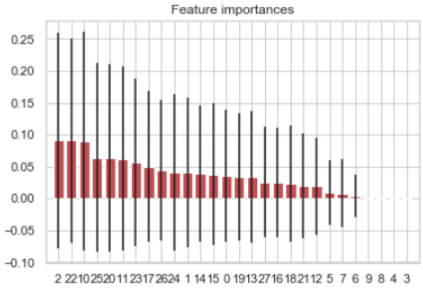
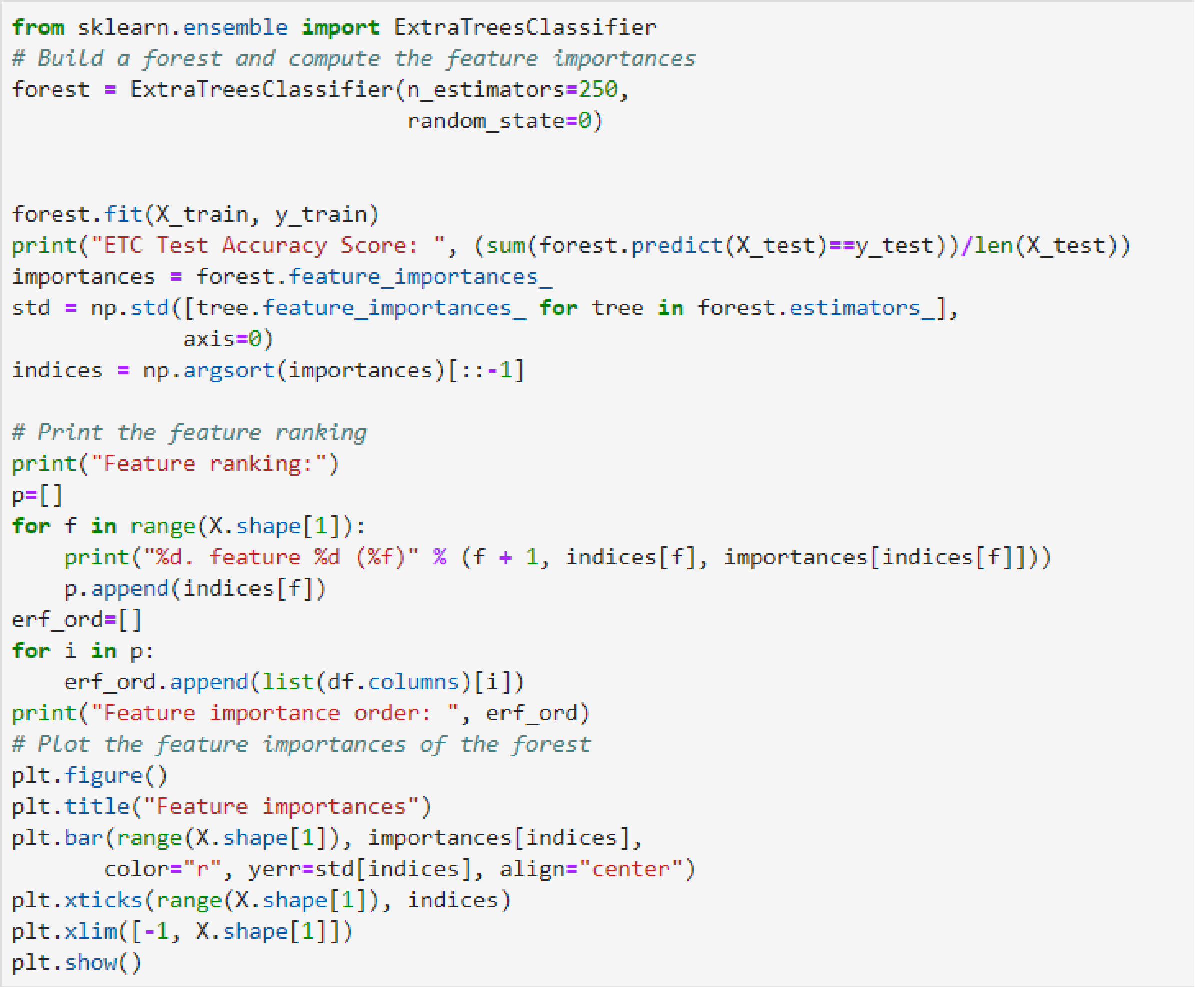


Figure 12.Random Forest Classifier - Feature importance and ranking

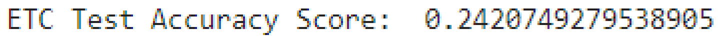
As observed from the above results, there was not much of an improvement between Decision trees classifier and random forest classifier but nevertheless there was a little. With a consistent cross validation score and feature extraction using *feature_importances_* function, important features with respect to this model were obtained.

In an attempt to improve the predictability, as a last attempt, Extremely Randomized Trees was used. The implementation and results are as follows:



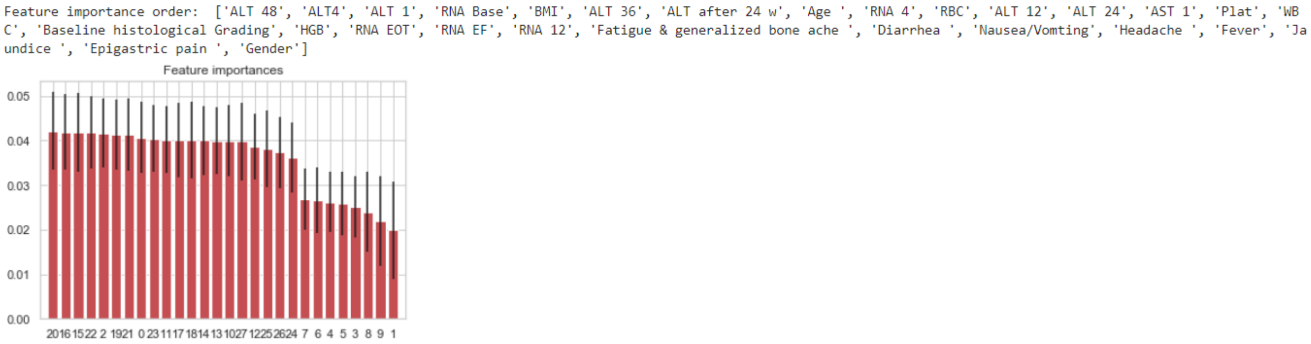
```
from sklearn.ensemble import ExtraTreesClassifier
# Build a forest and compute the feature importances
forest = ExtraTreesClassifier(n_estimators=250,
                              random_state=0)


forest.fit(X_train, y_train)
print("ETC Test Accuracy Score: ", (sum(forest.predict(X_test)==y_test))/len(X_test))
importances = forest.feature_importances_
std = np.std([tree.feature_importances_ for tree in forest.estimators_],
             axis=0)
indices = np.argsort(importances)[::-1]

# Print the feature ranking
print("Feature ranking:")
p=[]
for f in range(X.shape[1]):
    print("%d. feature %d (%f)" % (f + 1, indices[f], importances[indices[f]]))
    p.append(indices[f])
erf_ord=[]
for i in p:
    erf_ord.append(list(df.columns)[i])
print("Feature importance order: ", erf_ord)
# Plot the feature importances of the forest
plt.figure()
plt.title("Feature importances")
plt.bar(range(X.shape[1]), importances[indices],
       color="r", yerr=std[indices], align="center")
plt.xticks(range(X.shape[1]), indices)
plt.xlim([-1, X.shape[1]])
plt.show()
```

Figure 13. Implementation of Extra Trees Classifier - Prediction and Feature extraction



```
ETC Test Accuracy Score:  0.2420749279538905
```

Figure 14. Extra Trees Classifier test set accuracy



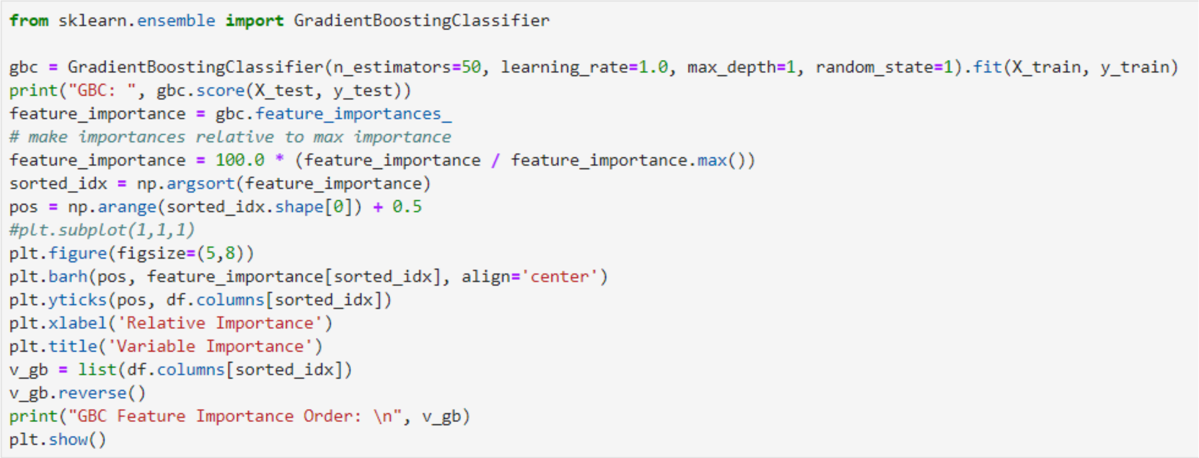
```
Feature importance order:  ['ALT 48', 'ALT4', 'ALT 1', 'RNA Base', 'BMI', 'ALT 36', 'ALT after 24 w', 'Age ', 'RNA 4', 'RBC', 'ALT 12', 'ALT 24', 'AST 1', 'Plat', 'WB
C', 'Baseline histological Grading', 'HGB', 'RNA EOT', 'RNA EF', 'RNA 12', 'Fatigue & generalized bone ache ', 'Diarrhea ', 'Nausea/Vomting',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

Figure 15. Extra Trees Classifier - Feature importance and ranking

As seen above, Extremely Randomized trees did not work as expected, however, features could still be extracted. At this point, there could only be one solution to explain these accuracy scores

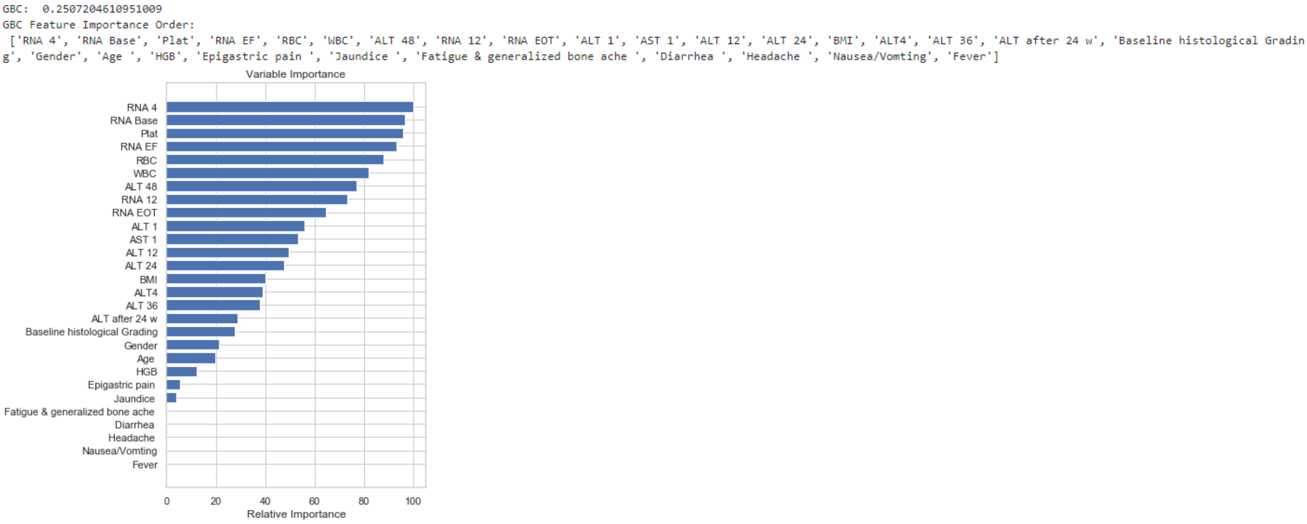
– fewer data points. If the results turn out to be consistently low, then this assumption would gain enough evidence.

Moving forward with this assumption, upon research Gradient Tree Boosting seemed to work in some occasions. The next classifier to be tested was Gradient Boosting. The implementations and results can be seen below:



```
from sklearn.ensemble import GradientBoostingClassifier

gbc = GradientBoostingClassifier(n_estimators=50, learning_rate=1.0, max_depth=1, random_state=1).fit(X_train, y_train)
print("GBC: ", gbc.score(X_test, y_test))
feature_importance = gbc.feature_importances_
# make importances relative to max importance
feature_importance = 100.0 * (feature_importance / feature_importance.max())
sorted_idx = np.argsort(feature_importance)
pos = np.arange(sorted_idx.shape[0]) + 0.5
#plt.subplot(1,1,1)
plt.figure(figsize=(5,8))
plt.barh(pos, feature_importance[sorted_idx], align='center')
plt.yticks(pos, df.columns[sorted_idx])
plt.xlabel('Relative Importance')
plt.title('Variable Importance')
v_gb = list(df.columns[sorted_idx])
v_gb.reverse()
print("GBC Feature Importance Order: \n", v_gb)
plt.show()
```

Figure 16. Implementation of Gradient Boost Classifier - Prediction and Feature extraction

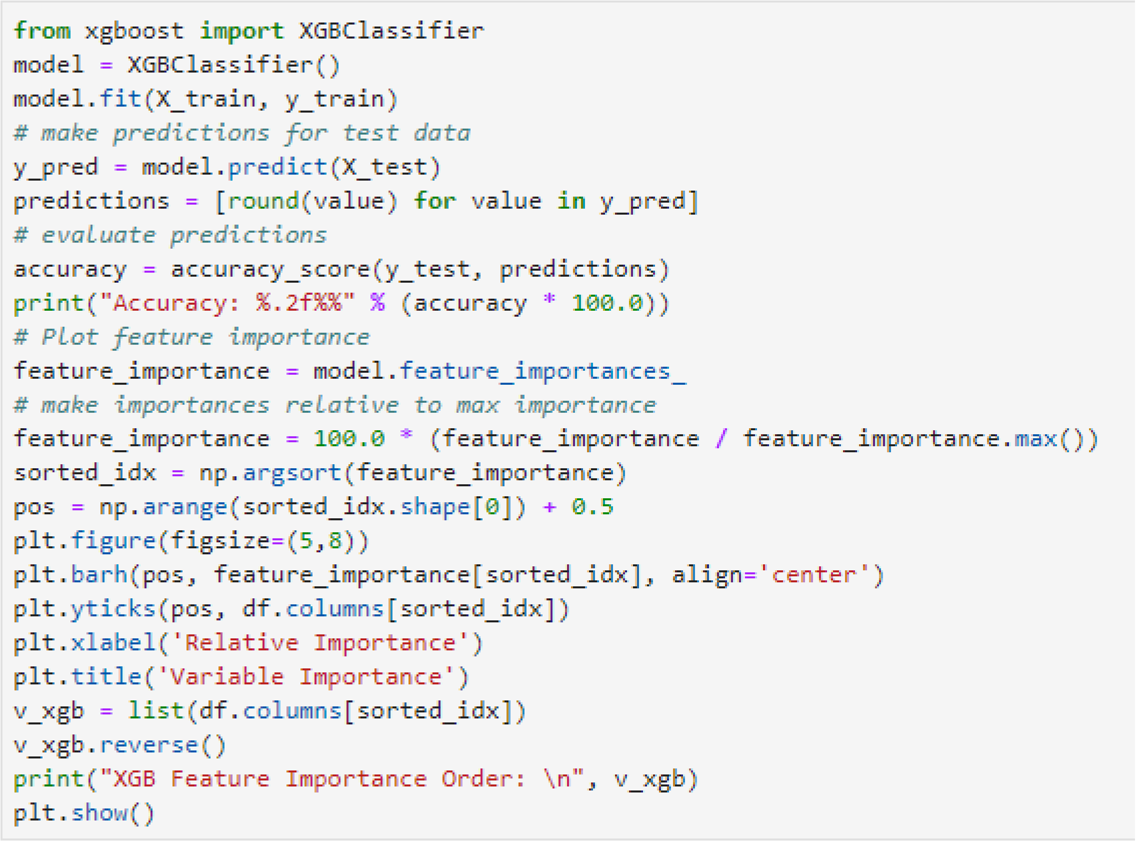


```
GBC:  0.2507204610951009
GBC Feature Importance Order: 
 ['RNA 4', 'RNA Base', 'Plat', 'RNA EF', 'RBC', 'WBC', 'ALT 48', 'RNA 12', 'RNA EOT', 'ALT 1', 'AST 1', 'ALT 12', 'ALT 24', 'BMI', 'ALT4', 'ALT 36', 'ALT after 24 w', 'Baseline histological Gradin
g', 'Gender', 'Age ', 'HGB', 'Epigastric pain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

Figure 17. Gradient Boost Classifier - Accuracy and Feature extraction

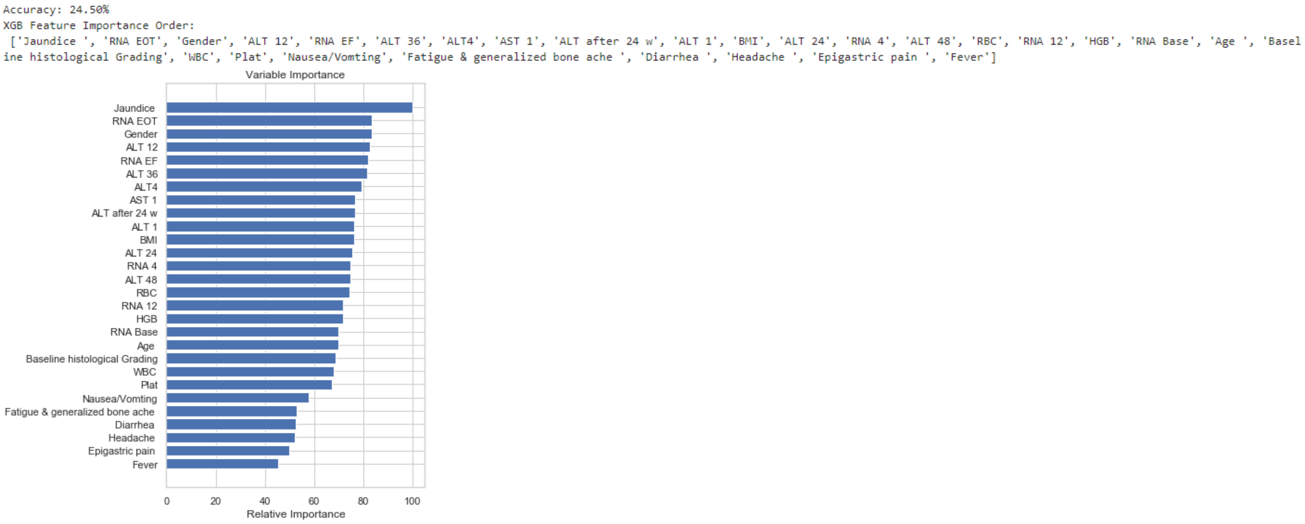
As observed in Figure 16, even Gradient boosting did help to boost the scores. This served as evidence to my above assumption and it can be said that the insufficient amount of data points(or rows) poses as a barrier to the project.

Next technique was purely chosen for the purpose of feature extraction. To maintain consistency with the Gradient Boosting Classifier, Extreme Gradient Boosting was used as follows:



```
from xgboost import XGBClassifier
model = XGBClassifier()
model.fit(X_train, y_train)
# make predictions for test data
y_pred = model.predict(X_test)
predictions = [round(value) for value in y_pred]
# evaluate predictions
accuracy = accuracy_score(y_test, predictions)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
# Plot feature importance
feature_importance = model.feature_importances_
# make importances relative to max importance
feature_importance = 100.0 * (feature_importance / feature_importance.max())
sorted_idx = np.argsort(feature_importance)
pos = np.arange(sorted_idx.shape[0]) + 0.5
plt.figure(figsize=(5,8))
plt.barh(pos, feature_importance[sorted_idx], align='center')
plt.yticks(pos, df.columns[sorted_idx])
plt.xlabel('Relative Importance')
plt.title('Variable Importance')
v_xgb = list(df.columns[sorted_idx])
v_xgb.reverse()
print("XGB Feature Importance Order: \n", v_xgb)
plt.show()
```

Figure 18. Implementation of Extreme gradient Boost Classifier – Prediction and Feature extraction

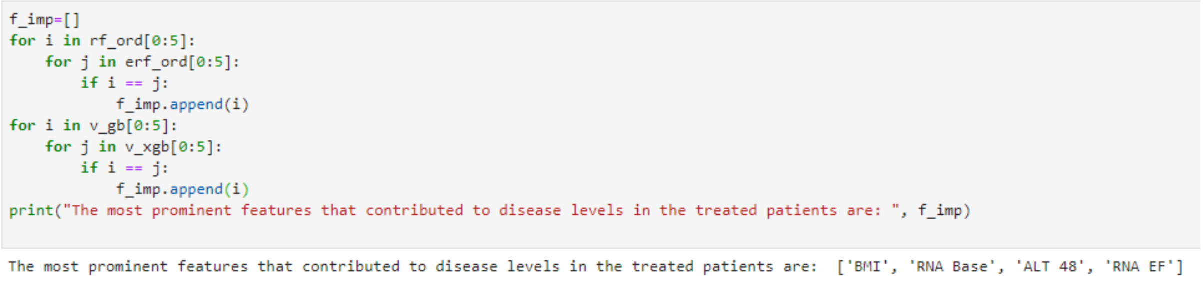


```
Accuracy: 24.50%
XGB Feature Importance Order:
 ['Jaundice ', 'RNA EOT', 'Gender', 'ALT 12', 'RNA EF', 'ALT 36', 'ALT4', 'AST 1', 'ALT after 24 w', 'ALT 1', 'BMI', 'ALT 24', 'RNA 4', 'ALT 48', 'RBC', 'RNA 12', 'HGB', 'RNA Base', 'Age ', 'Basel
ine histological Grading', 'WBC', 'Plat', 'Nausea/Vomting',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

Figure 19. Extreme Gradient Boost Classifier - Accuracy and Feature extraction

These results are also consistent with all the algorithms chosen to work with the discrete data that the Hepatitis C Virus in Egyptian Patients is.

One advantage of this work is feature extraction. So the next step was to pull the top features extracted from all four ensembles – Random Forests, Extremely Randomized trees, Gradient Boosting and Extreme gradient boosting. While the first two techniques gave features with effective splitting of trees, the later worked on loss reduction. Pooling 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.



```
f_imp=[]
for i in rf_ord[0:5]:
    for j in erf_ord[0:5]:
        if i == j:
            f_imp.append(i)
for i in v_gb[0:5]:
    for j in v_xgb[0:5]:
        if i == j:
            f_imp.append(i)
print("The most prominent features that contributed to disease levels in the treated patients are: ", f_imp)
```

```
The most prominent features that contributed to disease levels in the treated patients are:  ['BMI', 'RNA Base', 'ALT 48', 'RNA EF']
```

Figure 20. Most prominent features

CONCLUSION

In conclusion, it is proven that for better results, more data points are needed. This encourages a constant look out for data belonging to HCV patients and the history of the disease. As much as it is a classification problem, it also poses as a data collection problem. With the features extracted, some interesting facts can be established.

1. 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. This can be validated with the observations made in Figure 4.
2. 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[4]. This is validated by the EDA in Figure 7.
3. ALT 48 refers to the ALT test which tells the damage done to the liver tissue after 48 weeks of treatment. High levels of ALT may indicate liver damage from hepatitis, infection, cirrhosis, liver cancer, or other liver diseases[5]. 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.
4. RNA EF is the Elongation Factor of the RNA stands which facilitates the replication of the virus[6]. This 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.

With the above features, predictability of the progression of the disease can be monitored and explained if not treated successfully. The unseen advantage is that even with the increase in data points, these features remain important and will continue to help quantify the disease in an effective way.

REFERENCES

1. Nasr, M., El-Bahnasy, K., Hamdy, M., & Kamal, S. M. (2017). A novel model based on non invasive methods for prediction of liver fibrosis. 2017 13th International Computer Engineering Conference (ICENCO).
2. Kelman, C. W., Bass, A. J., & Holman, C. D. J. (2002). Research use of linked health data - a best practice protocol. Australian and New Zealand Journal of Public Health, 26(3), 251–255.
3. Patton, G. C., Coffey, C., Sawyer, S. M., Viner, R. M., Haller, D. M., Bose, K., … Mathers, C. D. (2009). Global patterns of mortality in young people: a systematic analysis of population health data. The Lancet, 374(9693), 881–892.
4. Shimakami, T., Yamane, D., Welsch, C., Hensley, L., Jangra, R. K., & Lemon, S. M. (2012). Base pairing between hepatitis C virus RNA and microRNA 122 3' of its seed sequence is essential for genome stabilization and production of infectious virus. Journal of virology, 86(13), 7372–7383. https://doi.org/10.1128/JVI.00513-12
5. ALT Blood Test, https://medlineplus.gov/lab-tests/alt-blood-test/
6. Li, D., Wei, T., Abbott, C. M., & Harrich, D. (2013). The unexpected roles of eukaryotic translation elongation factors in RNA virus replication and pathogenesis. Microbiology and molecular biology reviews : MMBR, 77(2), 253–266. https://doi.org/10.1128/MMBR.00059-12