CAPSTONE: Hepatitis C Virus (HCV) for Egyptian patients Data Set

CAPSTONE

Prepared by **Lucia Oviedo**
August 12, 2020

TABLE OF CONTENTS

OVERVIEW

Problem: Liver biopsies are typically the standard diagnosis of liver progression. However, it is associated with serios complications, inconvenient to patients and expensive. The goal is to generate a reliable model implementing machine learning algorithms to substitute liver biopsy and determine the degree of liver fibrosis.

Getting to know our data

Egyptian patients who underwent treatment dosages for HCV about 18 months. Discretization should be applied based on expert recommendations as follows.

Feature Names	Feature Values	Discretization (Items)
Age	32:61	[0; 32],]32; 37],]37; 42],]42; 47],]47; 52],]52; 57],]57; 62]
Gender	Male,Female	[Male], [Female]
BMI(Body Mass Index)	22:35	[0; 18:5[[18:5; 25[, [25; 30[, [30; 35[, [35; 40[
Fever	Absent, Present	[Absent], [Present] -
Nausea/Vomiting	Absent, Present	[Absent], [Present] -
Headache	Absent, Present	[Absent], [Present] -
Diarrhea	Absent, Present	[Absent], [Present] -
Fatigue	Absent, Present	[Absent], [Present] -
Bone ache	Absent, Present	[Absent], [Present] -
Jaundice	Absent, Present	[Absent], [Present] -
Epigastria pain	Absent, Present	[Absent], [Present] -
WBC(White Blood Cells)	2991:12101	[0; 4000[, [4000; 11000[, [11000; 12101]
RBC(Red Blood Cells)	3816422:5018451	[0; 3000000[, [3000000; 5000000[,[5000000; 5018451]
HGB(Hemoglobin)	2:20	If (Gender==[Male]):[2; 14[, [14; 17:5],]17:5; 20]If(Gender==[Female]):[2; 12:3[, [12:3; 15:3],]15:3; 20]
Plat(Platelet)	93013:226464	[93013; 100000[, [100000; 255000[,[255000; 226465[
AST1(1 week)	0.088888889	[0; 20[, [20; 40],]40; 128]
ALT1(1 week)	0.088888889	[0; 20[, [20; 40],]40; 128]
ALT4(4 weeks)	0.088888889	[0; 20[, [20; 40],]40; 128]
ALT12(12 weeks)	0.088888889	[0; 20[, [20; 40],]40; 128]
ALT24(24 weeks)	0.088888889	[0; 20[, [20; 40],]40; 128]
ALT36(36 weeks)	0.088888889	[0; 20[, [20; 40],]40; 128]
ALT48(48 weeks)	0.088888889	[0; 20[, [20; 40],]40; 128]
RNA Base	0:1201086	[0; 5],]5; 1201086]
RNA 4	0:1201715	[0; 5],]5; 1201715]
RNA 12	0:3731527	[0; 5],]5; 3731527]
RNA EOT	0:808450	[0; 5],]5; 808450]
RNA EF(Elongation Factor)	0:808450	[0; 5],]5; 808450]
Baseline Histological Grading	1:16	[1]; [2]; [3]; :::[16]
Baseline Histological Staging	F0:F4	[No Fibrosis], [Portal Fibrosis],Staging (Class Label) [Few Septa], [Many Septa], [Cirrhosis]

We have 29 columns with 1385 observations.

- General Info: Age, Gender, BMI
- Symptoms: Fever, Nausea/Vomiting, Headache, Diarrhea, Fatige, Bone ache, Jaudice, Epigastria pain
- Labs: WBC, RBC, HGB, plat, AST1, ALT1, ALT4, ALT12, ALT24, ALT36, ALT48, RNA Base, RNA 4, RNA EOT, RNA EF
- Baseline: Historical Grading and Historical Staging

PREPARE AND EXPLORE THE DATA

Data Types

Nominal values: Fender, Fever, Nausea & Vomiting, Headache, Diarrhea, Fatigue and Bone Ache, Jaundice, Epigastric pain, Baseline Historical Grading, Baseline Historical Staging.

Numeric values: Age, BMI, WBC, RBC, HGB, Plat, AST1, ALT1… ALT48, ALTafter24w, RNABase, RNA12, RNAEOT, RNAEF, BHSGrading, BHStaging

As per expert recommendations, discretization will be applied to all feature and all of them will be nominal value.

	count	mean	std	min	25%	50%	75%	max
Age	1385.0	4.631913e+01	8.781506	32.0	39.0	46.0	54.0	61.0
Gender	1385.0	1.489531e+00	0.500071	1.0	1.0	1.0	2.0	2.0
BMI	1385.0	2.860866e+01	4.076215	22.0	25.0	29.0	32.0	35.0
Fever	1385.0	1.515523e+00	0.499939	1.0	1.0	2.0	2.0	2.0
NauseaVomting	1385.0	1.502527e+00	0.500174	1.0	1.0	2.0	2.0	2.0
Headache	1385.0	1.496029e+00	0.500165	1.0	1.0	1.0	2.0	2.0
Diarrhea	1385.0	1.502527e+00	0.500174	1.0	1.0	2.0	2.0	2.0
FatigueBoneAche	1385.0	1.498917e+00	0.500179	1.0	1.0	1.0	2.0	2.0
Jaundice	1385.0	1.501083e+00	0.500179	1.0	1.0	2.0	2.0	2.0
EpigastricPain	1385.0	1.503971e+00	0.500165	1.0	1.0	2.0	2.0	2.0
WBC	1385.0	7.533386e+03	2668.220333	2991.0	5219.0	7498.0	9902.0	12101.0
RBC	1385.0	4.422130e+06	346357.711599	3816422.0	4121374.0	4438465.0	4721279.0	5018451.0
HGB	1385.0	1.258773e+01	1.713511	10.0	11.0	13.0	14.0	15.0
Plat	1385.0	1.583481e+05	38794.785550	93013.0	124479.0	157916.0	190314.0	226464.0
AST1	1385.0	8.277473e+01	25.993242	39.0	60.0	83.0	105.0	128.0
ALT1	1385.0	8.391625e+01	25.922800	39.0	62.0	83.0	106.0	128.0
ALT4	1385.0	8.340578e+01	26.529730	39.0	61.0	82.0	107.0	128.0
ALT12	1385.0	8.351047e+01	26.064478	39.0	60.0	84.0	106.0	128.0
ALT24	1385.0	8.370903e+01	26.205994	39.0	61.0	83.0	107.0	128.0
ALT36	1385.0	8.311769e+01	26.399031	5.0	61.0	84.0	106.0	128.0
ALT48	1385.0	8.362960e+01	26.223955	5.0	61.0	83.0	106.0	128.0
ALTafter24w	1385.0	3.343827e+01	7.073569	5.0	28.0	34.0	40.0	45.0
RNABase	1385.0	5.909512e+05	353935.357602	11.0	269253.0	593103.0	886791.0	1201086.0
RNA4	1385.0	6.008956e+05	362315.132786	5.0	270893.0	597869.0	909093.0	1201715.0
RNA12	1385.0	2.887536e+05	285350.674511	5.0	5.0	234359.0	524819.0	3731527.0
RNAEOT	1385.0	2.876603e+05	264559.525070	5.0	5.0	251376.0	517806.0	808450.0
RNAEF	1385.0	2.913783e+05	267700.691713	5.0	5.0	244049.0	527864.0	810333.0
BHSGrading	1385.0	9.761733e+00	4.023896	3.0	6.0	10.0	13.0	16.0
BHStaging	1385.0	2.536462e+00	1.121392	1.0	2.0	3.0	4.0	4.0

Statistical Analysis

- AGE mean value is 46 years old with a min of 32 and a max of 61
- BMI goes from 22 to 35 with a mean of 28.
- WBC min of 2991 max of 12101
- HGB is from 10 to 15 and it depends of the gender.
- Plat min of 93013 and max of 226464
- AST1, ALT1, ALT4, ALT12, ALT24 min of 39 and max of 128
- ALT36, ALT48 min of 5 and max of 128
- ALT after24w min of 5 and max of 45.
- Non near zero- variable variables were found .
- No missing values were found.
- There is a high RNA12 value of almost 3731527 that doesn't follow the RNA12 pattern. Record was removed.
- There are few values where ALT48, ALT36 and ALT afer24W that are very close to 0 and don't follow same pattern as other data, this discovered as part of pairplot visualization. # Records got removed.
- No Fibrosis data is not found in the data, it would be recommendable to collect it.

CORRELATION

Hight correlation are displayed on the left-hand side table.

Also, a high correlation was found between RNA12, RNAEOT and RNAEF.

Correlation	BHStaging
BHStaging	1
NauseaVomiting	0.054906
ALT1	0.036867
RNA12	0.034449
ALTafter24w	0.033919
RNAEF	0.030519
RNABase	0.029411
Jaundice	0.020219
WBC	0.017945
FatigueBoneAche	0.014563
Gender	0.011955
RBC	0.009623
HGB	0.002752
ALT12	0.000809

Correlation	BHStaging
Headache	-0.002
ALT24	-0.00489
Diarrhea	-0.00564
ALT36	-0.00643
ALT48	-0.01353
ALT4	-0.015
Plat	-0.01728
RNAEOT	-0.01749
Age	-0.0196
AST1	-0.02513
Fever	-0.03098
RNA4	-0.03295
BHSGrading	-0.04707
EpigastricPain	-0.05211
BMI	-0.05726

COVARIANCE ESTIMATION

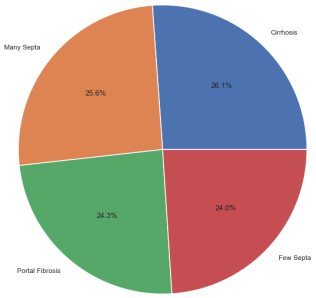
Positive trends are displayed on the left-hand side table.

Covariance	BHStaging
RNABase	11673.06174
RNA12	11023.25078
RNAEF	9161.655206
RBC	3737.529735
WBC	53.693636
BHStaging	1.257521
ALT1	1.071698
ALTafter24w	0.269052
NauseaVomting	0.030797
ALT12	0.023643
Jaundice	0.011341
FatigueBoneAche	0.008168
Gender	0.006704
HGB	0.005289

Covariance	BHStaging
Headache	-0.00112
Diarrhea	-0.003163
Fever	-0.017366
EpigastricPain	-0.029227
ALT24	-0.143646
ALT36	-0.190349
Age	-0.193003
BHSGrading	-0.212404
BMI	-0.261732
ALT48	-0.397973
ALT4	-0.446165
AST1	-0.732387
Plat	-751.91045
RNAEOT	-5187.6965
RNA4	-13385.813

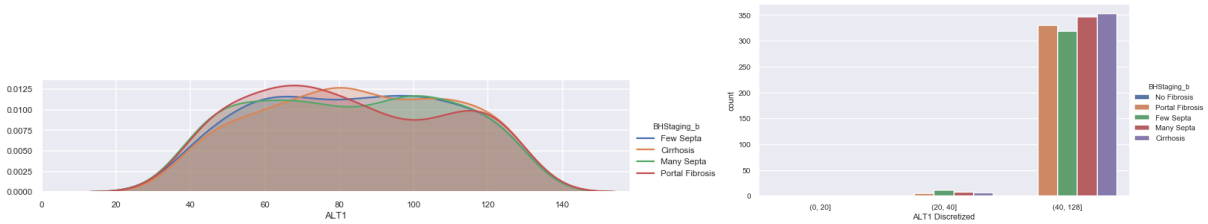
EDA

BHStaging



- There is no information about No fibrosis in the data set, it might help improve the model.

ALT1

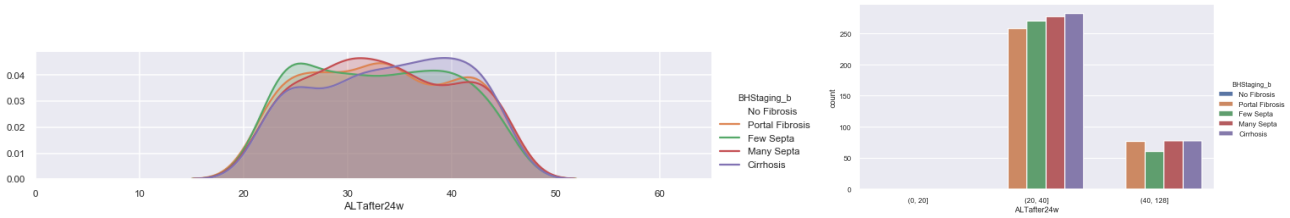


- Portal Fibrosis: ALT1 is higher around 60 to 70
- Few Septa: ALT1 range from 60 to 180
- Many Septa: ALT1 is higher from 90 to 110
- Cirrhosis: ALT1 is higher around 80
- Higher count found for Cirrhosis and Many Septa

RNA12

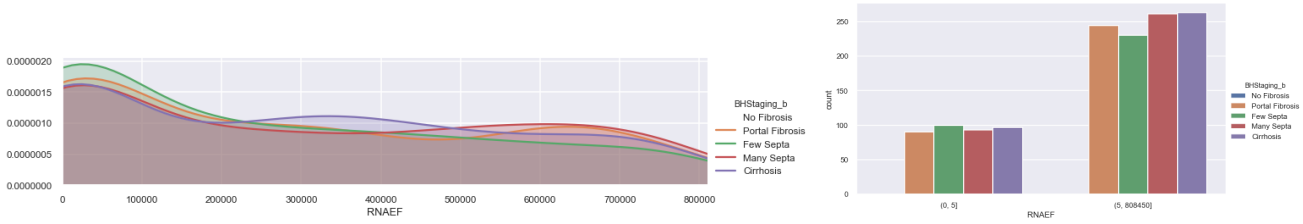
- For values less than 3000000 Cirrhosis count is less and for values higher than 3000000 count increases.
- For values less than 3000000 Portal Fibrosis count is higher, after this the count decreases.
- Many Septa and Cirrhosis has higher count for RNA from 5 to 3731527.

ALTafter24w



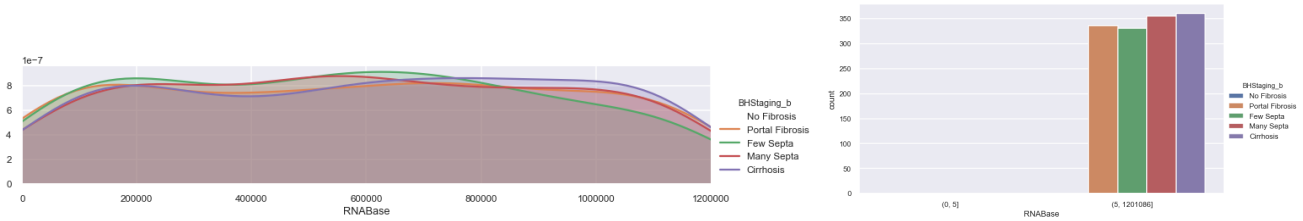
- No fibrosis has the higher count in the range of 20-40.
- Few Septa has the lower count in the range of 40-128

RNAEF



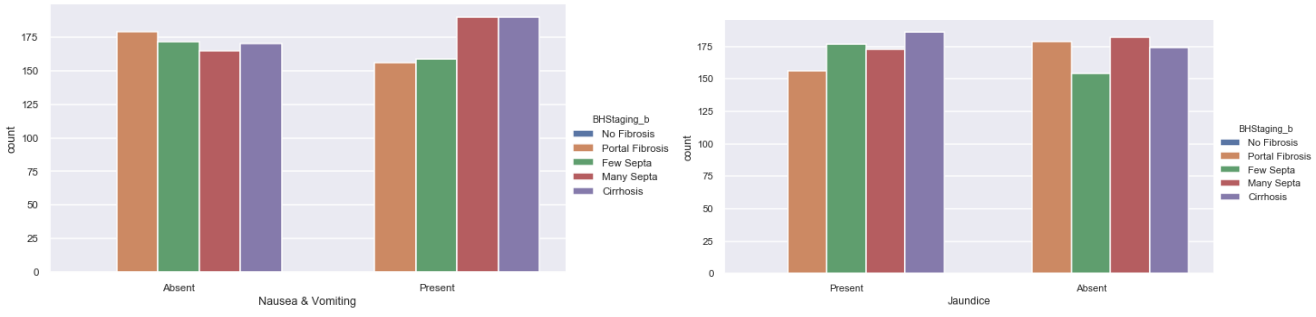
- Few Septa has a higher count on the range of 0-5

RNABase

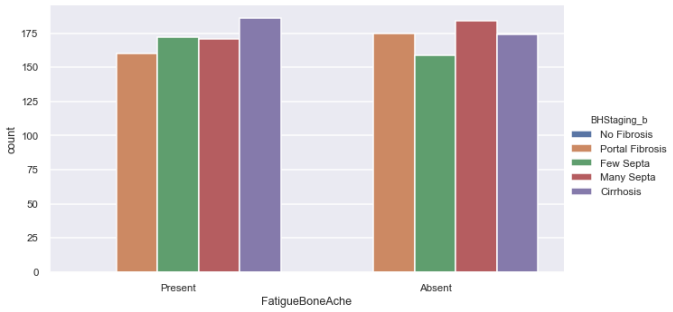


- Cirrhosis has a slightly higher count. All of the stages have around same count.

Symptoms

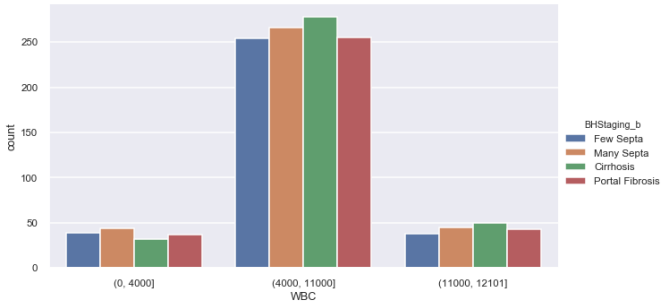


- Nausea and Vomiting has the higher correlation with BHStaging and Present count is higher for Many Septa and Cirrhosis.
- Jaundice count is higher for Cirrhosis but also a great percentage with Cirrhosis don't have this symptom.



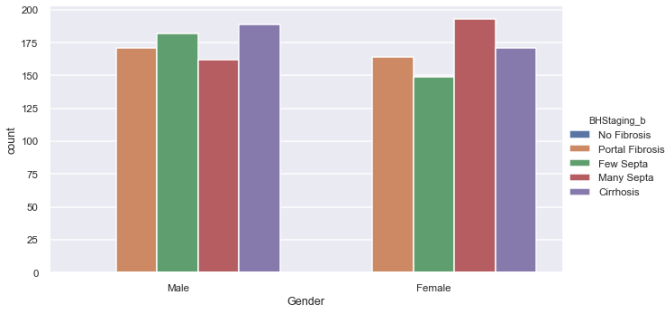
- Fatigue and bone ache count is higher for Cirrhosis but also a great percentage with Cirrhosis don't have this symptom.
- Many Septa has the higher count with absent of fatigue and bone ache.

WBC

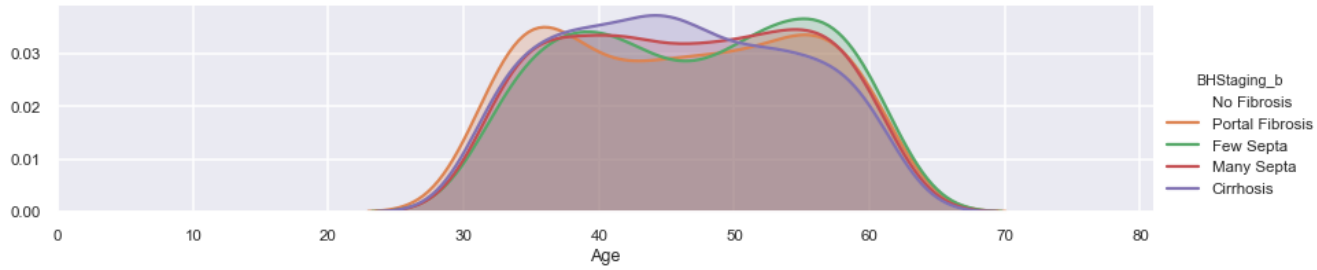


- Cirrhosis has the higher count in the range of 400-11000 and 11000 – 12101.

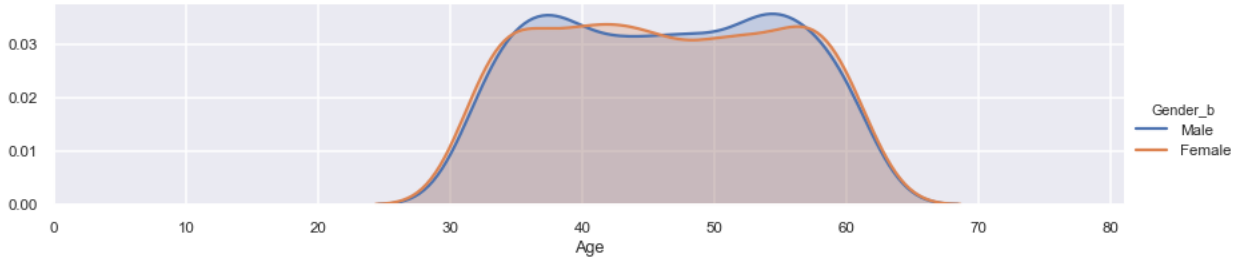
Age and Gender



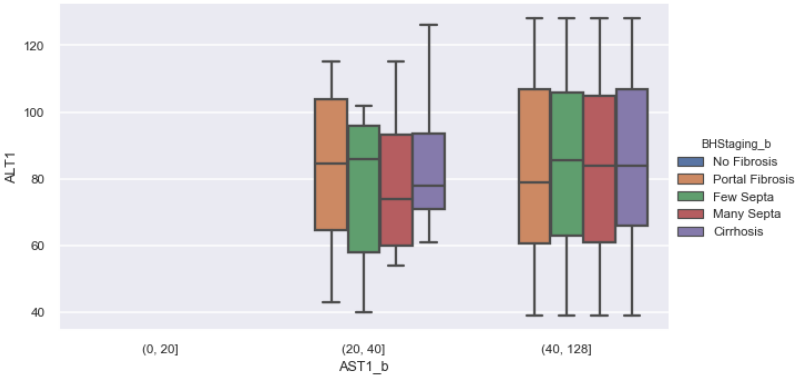
- There are almost the same count of Females and Males.
- Portal Fibrosis, Few Septa and Cirrhosis count is higher on Male.
- Many Septa count is higher on Females.



- Portal Cirrhosis count is higher on mid-thirties and mid-fifties
- Few Septa is higher on forties and mid-fifties
- Many Septa is higher on forties and mid-fifties
- Cirrhosis is higher for ages between 40 and 50

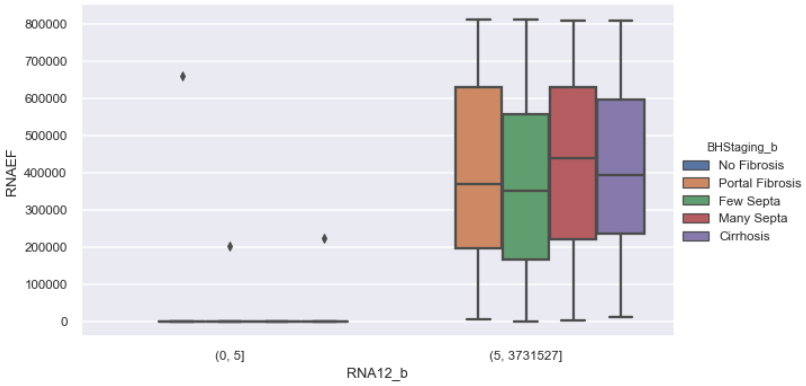


AST1 VS ALT1/ BHStaging



- Many septa has a lower average of ALT1 in the range of 20-40.
- Portal Fibrosis has a lower average of ALT1 in the range of 40-128
- There is no AST1 existing data for range 0-20.

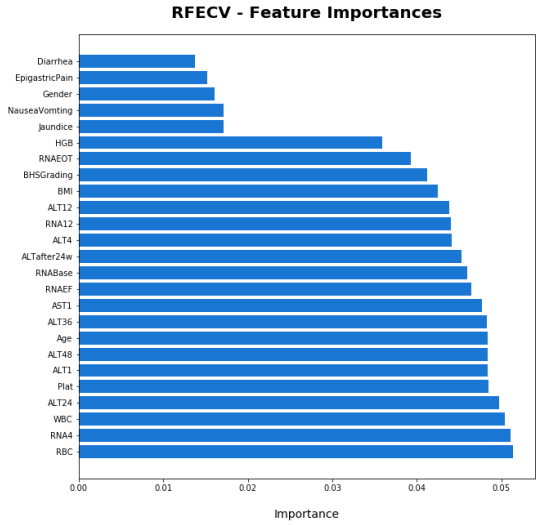
RNA12 VS RNAEF/ BHStaging



- Few Septa has a lower average of RNAEF
- Many septa has a the higher average of RNAEF
- There is no RNA12 existing data for range 0-5.

FEATURE ENGINEERING AND DIMENSIONALITY REDUCTION

To implement feature engineering recursive feature selection with cross-validation was implemented using RandomForestClassifier.



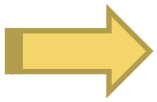
An optimal number of 25 features were found with RFE. In the graphic we can see them listed in ascen ding order by importance.

ONE-HOT ENCODING

One – Hot Encoding was implemented for the nominal features, in this case all of our attributes in the data set were discretized and hot encoded.

'Age', 'BMI', WBC', 'RBC', 'HGB', 'Plat', 'AST1', 'ALT1', 'ALT4', 'ALT12', 'ALT24', 'ALT36', 'ALT48', 'ALTafter24w', 'RNABase', 'RNA4', 'RNA12', 'RNAEOT', 'RNAEF', 'BHStaging', 'BHSGrading', 'Jaundice', 'Fever', 'Diarrhea', 'EpigastricPain', 'Headache', 'FatigueBoneAche', 'NauseaVomting', 'Gender'.

	0	1	2	3	4
Age	5	3	5	4	6
Gender	Male	Male	Male	Female	Male
BMI	3	2	3	3	3
Fever	Present	Absent	Present	Absent	Absent
NauseaVomting	Absent	Present	Present	Present	Absent
Headache	Absent	Present	Present	Absent	Present
Diarrhea	Absent	Absent	Present	Present	Absent
FatigueBoneAche	Present	Present	Absent	Absent	Present
Jaundice	Present	Present	Absent	Present	Present
EpigastricPain	Present	Absent	Absent	Absent	Present
WBC	1	2	1	1	0
RBC	1	1	1	1	1
HGB	0	0	0	0	0
Plat	1	1	1	1	1
AST1	2	2	2	2	2
ALT1	2	2	2	2	2
ALT4	2	2	2	2	2
ALT12	2	2	2	2	2
ALT24	2	2	2	2	2
ALT36	0	2	0	2	2
ALT48	0	2	0	2	2
ALTafter24w	0	2	0	1	1
RNABase	1	1	1	1	1
RNA4	1	1	1	1	1
RNA12	1	1	0	1	1
RNAEOT	0	1	1	1	1
RNAEF	0	1	1	1	1
BHSGrading	13	4	4	10	11



	0	1	2	3	4
Age_0	0	0	0	0	0
Age_1	0	0	0	0	0
Age_2	0	0	0	0	0
Age_3	0	1	0	0	0
Age_4	0	0	0	1	0
...	...	...	...	...	...
FatigueBoneAche_Present	1	1	0	0	1
NauseaVomting_Absent	1	0	0	0	1
NauseaVomting_Present	0	1	1	1	0
Gender_Female	0	0	0	1	0
Gender_Male	1	1	1	0	1

CLASSIFICATION

This problem was approached with classification Techniques. Below are the classifiers implemented to build models:

- GradientBoostingClassifier
- RandomForestClassifier
- AdaBoostClassifier
- LinearDiscriminantAnalysis
- DecisionTreeClassifier
- KNeighborsClassifier

However, we were performance measure for these models was around 25% of accuracy. So, different approach has to be taken.

DECISION RULES

Decision rules were implemented to create rules based on each one of the classifications for BHStaging:

- Portal Fibrosis
- Few Septa
- Many Septa
- Cirrhosis

To implement decision rules we used lw.RIPPER from Wittgenstein library.

MODEL TUNING

Form the selected classifiers, Random Forest Classifier was tuned using different values for n_estimators and max_depth.

Gradient Boosting Classifier was tuned using different values for n_estimators, learning_rate, subsample and max_depth.

Model Tuning was also applied to lw.RIPPER in order to obtain a better performance.

MODEL EVALUATION DECISION RULES

- Four decision rules were generated to determine Staging level from these 4 rules we got a performance measure from 72% to 76%
- For Portal Fibrosis and Few Septa we obtain a precision of 0 to determine when it is positive, which means these two decision rules perfume better to determine false.
- For Many Septa we obtain a precision of .40 for true results.
- For Cirrhosis we got the higher precision of .60 for true results.

BHStaging	Rule	Accuracy	Classification Report	Cross Validation
Portal Fibrosis	[ALTafter24w_2=1^Headache_Absent=1^BMI_2=0^Gender_Female=0^Age_2=0^FatigueBoneAche_Absent=0]	0.763689	precision recall f1-score support 0 0.77 0.98 0.87 270 **1 0.00 0.00 0.00 77**	[0.77142857 0.74285714 0.76811594 0.75362319 0.72463768]
Few Septa	[Gender_Female=0^Diarrhea_Absent=1^Fever_Absent=0^Jaundice_Absent=0^Age_3=1]	0.755043	precision recall f1-score support 0 0.76 1.00 0.86 263 **1 0.00 0.00 0.00 84**	[0.74285714 0.77142857 0.73913043 0.71014493 0.75362319]
Many Septa	[Headache_Absent=1^BMI_3=1^EpigastricPain_Absent=1^Fever_Absent=0^WBC_1=1^Age_2=0^Age_5=0]	0.720461	precision recall f1-score support 0 0.73 0.96 0.83 253 **1 0.40 0.06 0.11 94**	[0.71428571 0.72857143 0.73913043 0.71014493 0.68115942]
Cirrhosis	[EpigastricPain_Absent=1^BMI_2=1^BHSGrading_12=1^Age_1=0]	0.737752	precision recall f1-score support 0 0.74 0.99 0.85 255 **1 0.60 0.03 0.06 92**	[0.72857143 0.67142857 0.75362319 0.72463768 0.72463768]

MODEL EVALUATION FOR CLASIFICATION

- From all three models for classification method, the one who perform better is RFE using 25 columns before discretization. Many septa had a higher precision of 0.36 compared with the rest of the staging level.

Technique	Model	Accuracy	Classification Report	Cross Validation
RFE (No discretized, 25 columns)	RandomForestClassifier	0.285303	precision recall f1-score support 1 0.21 0.25 0.23 77 2 0.25 0.21 0.23 84 **3 0.36 0.28 0.31 94** 4 0.32 0.39 0.35 92	[0.25714286 0.22857143 0.30434783 0.26086957 0.26086957]
ohe(discretized)	LinearDiscriminantAnalysis	0.270893	precision recall f1-score support 1 0.27 0.22 0.24 77 2 0.22 0.24 0.23 84 3 0.30 0.31 0.30 94 4 0.29 0.30 0.30 92	[0.18571429 0.2 0.33333333 0.33333333 0.23188406]
Raw (no discretized)	AdaBoostClassifier	0.270893	precision recall f1-score support 1 0.27 0.26 0.26 77 2 0.24 0.24 0.24 84 3 0.29 0.27 0.28 94 4 0.28 0.32 0.30 92	[0.31428571 0.1 0.14492754 0.26086957 0.31884058]

RECOMMENDATIONS:

- Include “No fibrosis” Staging data
- Discretization for RNA Base, RNA 4, RNA EOT, RNA EF is [0; 5],]5; 1201086], contemplate dividing on more than two bins.
- Need to investigate and compare other techniques and algorithms of Decision Rules to improve performance.