MATH 257 Class Project

Hepatocellular carcinoma - HCC data classification

Shuai Li 013704886
Yuting Tao 013760071
Milica Miskovic 012446863

May 2019

Introduction

As stated in the EASL[1] Clinical Practice Guidelines, liver cancer is the fifth most common cancer and the second most frequent cause of cancer-related death globally. Hepatocellular carcinoma represents about 90% of primary liver cancers and constitutes a major global health problem. HCC has become increasing concern since the number of occurrences has been growing worldwide.

The purpose of this project is to explore the two populations of patients that lived and died and comparing mean vectors from those two populations. Since the common risk factor connected to the health problems, it was interesting to comparing mean vectors from populations of patients that do and do not consume alcohol. The first part of analysis is the exploratory analysis of the data set, the underlying distribution of the data and Principal Component Analysis. Classification tree algorithm and Random Forest method gave interesting insight in how influential are the variables in terms of the prior classification of patients regarding whether they lived or died during this research.

Regarding multivariate classifiers, we used quadratic disciminant rule which gave us the most accurate classification of the data compared to previous methods.

[1]European Association for the Study of the Liver

1 Data set

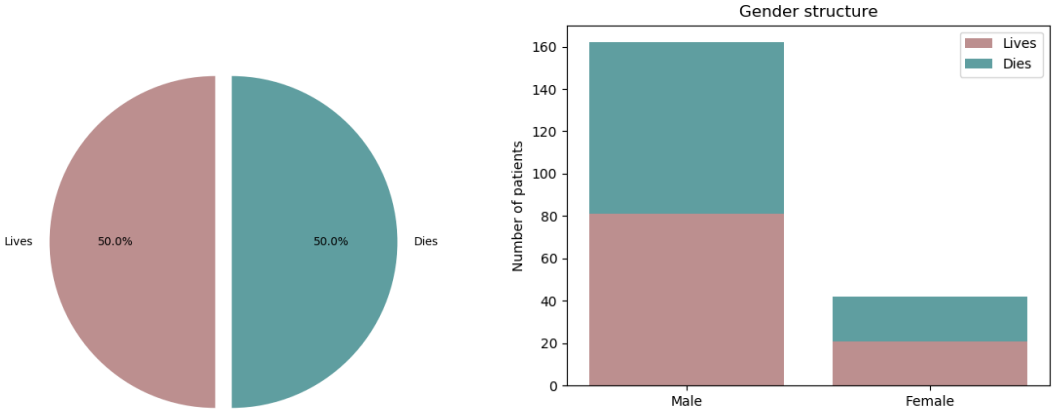
HCC dataset was obtained at a University Hospital in Portugal and contains several demographic, risk factors, laboratory and overall survival features of 204 real patients diagnosed with HCC. The dataset contains 49 features selected according to the EASL-EORTC.

This is a heterogeneous dataset, with 23 quantitative variables, and 26 qualitative variables, among which four are multilevel ordinal ('PS', 'Encephalopathy', 'Ascites' and 'Nodule')

1.1 Patients

As already stated, in this data set, we have 204 patients, perfectly balanced in sense that the same number of patients has value 0 and 1 of the "Class" variable.

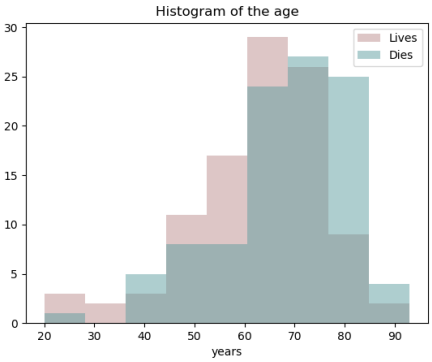
The sample is imbalanced gender wise, it consists out of larger proportion of men (79.41%, or 162 patients are male) but same proportion of male and female patients that live or die exactly 50%. So 50% of patients that died were male, therefore 39.7% off all patients are male and died during this research.



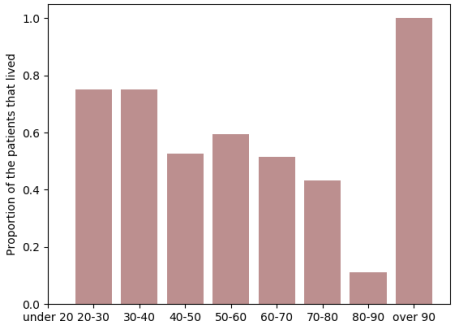
(a) Patients by Class.

(b) Patients by Gender.

Figure 1: Class and gender structure of patients.



(a) Patients by Age.



(b) Patients by Age groups.

Figure 2: Age structure of patients.

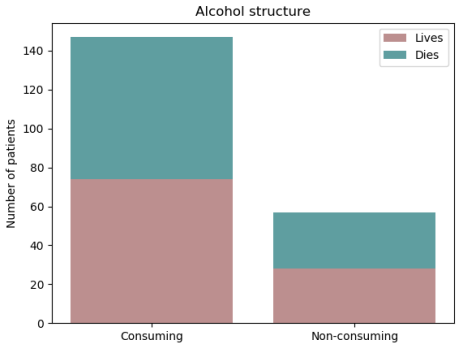
The average in years is 65.52, with youngest patient number 29 being 20 years old and the oldest patient number 100 being 93 years old (both have lived). Mode age (the age that is most frequent) is 71. Median age is 66.

If we split patients into age groups, the most populated one would be 60-70 year-old group with 72 patients, and that is also the group that has the highest number of people that lived (37 patients lived). Interestingly, the group with highest proportion of patients that lived is 90-and-older, since there is only one patient and he lived.

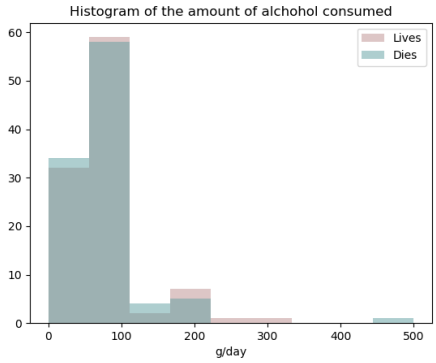
In the literature the usually mentioned risk factor for majority of liver related diseases is alcohol consumption. Our data set tracked it through two variables, categorical "Alcohol" and continuous "Grams per day" which tracks daily consumption in grams.

We have 147 patients that consume alcohol, out of which 74 live, that makes 72% of total number of patients. Out of 57 patients that do not consume alcohol, 28 live, which brings us to paradox situation that higher percentage of people that do consume alcohol lived (49% of people that drink in comparison to the 50.3% of non-drinking people). But that may be the case since the data is unbalanced.

In terms of amount of alcohol, our average patient would drink 75.29 grams of alcohol per day, the heaviest drinker is the patient number 10 and he/she would drink 500 grams of alcohol per day. This can be considered to be an unusual observation, but since there is no clear upper boundary for daily consumption of alcohol it is hard to assume either way. Majority of our



(a) Patients by Alcohol consumption.



(b) Patients by Alcohol daily intake.

Figure 3: Alcohol structure of patients.

patients drink between 0 and 100 grams of alcohol a day.

1.2 Data distribution

Pairwise plot is not very informative since there is a big number of variables, both categorical and continuous. Correlogram of 22 continuous variables is a little bit more informative and gives us overall shape of the probability distributions of the data set. Exploratory data analysis has to be done on smaller feature sets, or pairwise or by the multivariate data display tools like paralel plot and Andrews curves.

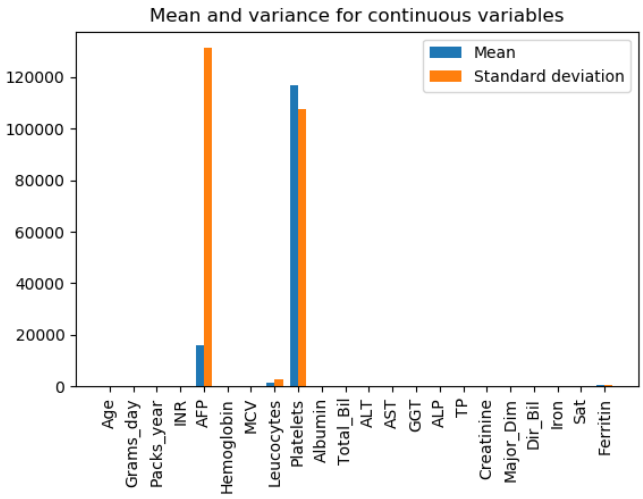
Parallel plot of continuous variables indicates an unusual value for AFP variable of patient number 72. Since there is not a hard cutoff for the value of the tumor marker AFP (Alpha-fetoprotein), we can consider leaving that unusual observation unchanged in the data set. The normal AFP levels of a healthy person are less than 10 ng/ml, and values over 500 ng/ml are considered extremely high, therefore the value of 1810346.00 ng/ml is possible, but not very probable. Also, there are some more unusually high values (over 100000), so domain consultation is needed in this case [2].

Based on the data described, we have following box plot of continuous variables, not very informative since the scale of the columns differ greatly, which imply that before we dive into more thorough analysis, standardization

[2]Alpha-Fetoprotein Tumor Marker (Blood) - Health encyclopedia

Age	65.529412
Grams_day	75.294118
Packs_year	21.411765
INR	1.443961
AFP	15882.189853
Hemoglobin	12.715686
MCV	94.818627
Leucocytes	1500.310784
Platelets	116677.460931
Albumin	3.417108
Total_Bil	3.165000
ALT	66.372549
AST	96.024510
GGT	271.441176
ALP	221.176471
TP	9.500490
Creatinine	1.137794
Major_Dim	6.888235
Dir_Bil	1.740637
Iron	87.455392
Sat	39.948333
Ferritin	435.257353

(a) Mean values of variables.



(b) Mean and variance barplot.

Figure 4

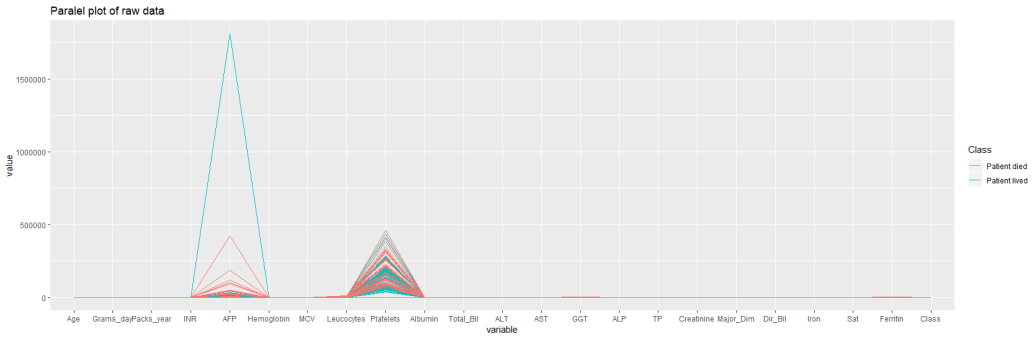


Figure 5: Parallel plot of continuous variables, raw data.

of data is in order. In spite that it indicates that there is an obviously unusual value in AFP feature column.

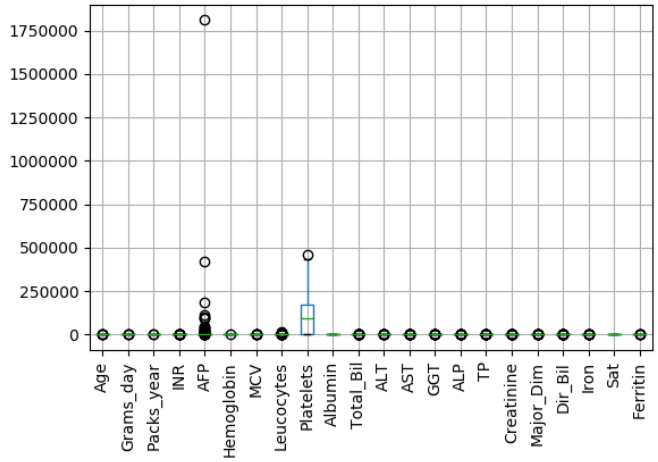


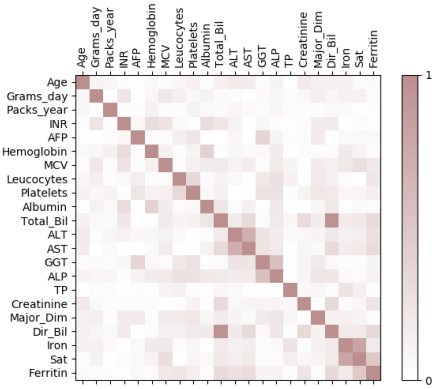
Figure 6: Boxplot of continuous variables, raw data.

Andrews plot tells us that there are two distinct groups, or that curves representing rows of the data set of patients that died are separable from the curves of the patients that died. The variables in the data set are sorted according to sample variance, so that Andrews plot depicts the data as good as possible.

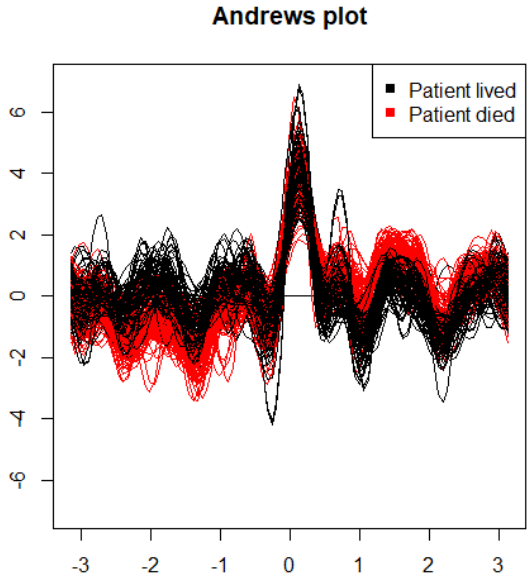
Variance covariance matrix will be given in appendix. Overall data features seem to be reasonably correlated. We have obvious clusters of correlated varables (ALT, AST, GGT, ALP and TP) which is to be expected since all of the variables in this groups are tumor markers, therefore it is reasonable to assume that values are going to be similar. The other cluster is Ferritin, Hemoglobin, Iron and Sat, which are blood related parameters. The most correlated features are direct bilirubin (Dir_Bil) and total bilirubin (Total_Bil) with correlation coefficient of 0.98.

1.3 Assessing multivariate normality

Since most of the multivariate statistical techniques we discussed in class assume that every observation comes from multivariate normal distribution



(a) Correlation matrix of continuous variables, raw data.



(b) Andrews of continuous variables, raw data.

Figure 7

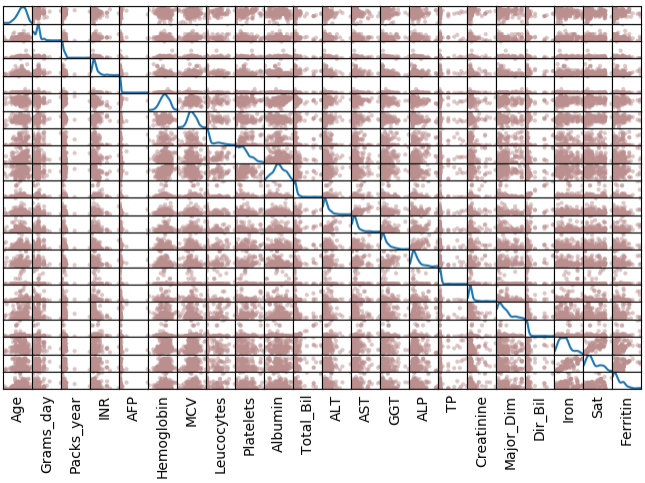
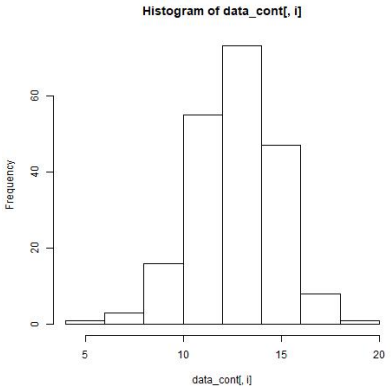
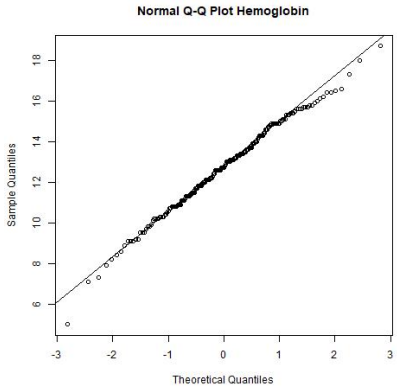


Figure 8: Pairwise plot of continuous variables, raw data.

vital part of this analysis is checking normality of every variable separately and whether they together describe a population that has underlying multivariate normal distribution. Since we have moderately big data set, normality assumption for every particular observation is less crucial for the quality of the inference.

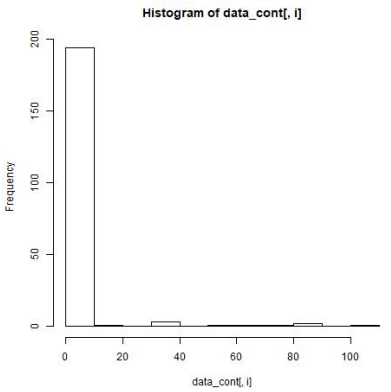


(a) Histogram.

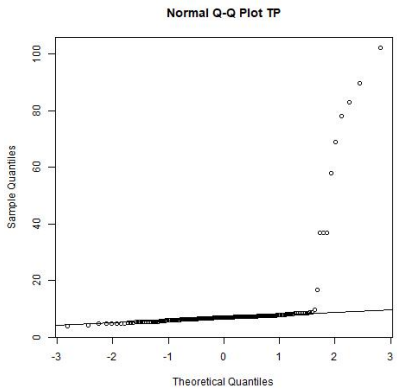


(b) QQ normal plot.

Figure 9: Variable Hemoglobin.



(a) Histogram.



(b) QQ normal plot.

Figure 10: Variable TP.

In our data set, the normal QQ plots and histograms of some variables (Age, Hemoglobin, Albumin, MCV) indicate that they have obvious normal

distribution. In the graph of the others, we can see influence of outliers, but since that is small number of points in comparison to the rest of the data set, it can be disregarded. Overall, even though there are some violations of the normality assumption, because of the nature of the medical data we can rightfully assume that underlying distribution that is sampled by this data set is multivariate normal [3]

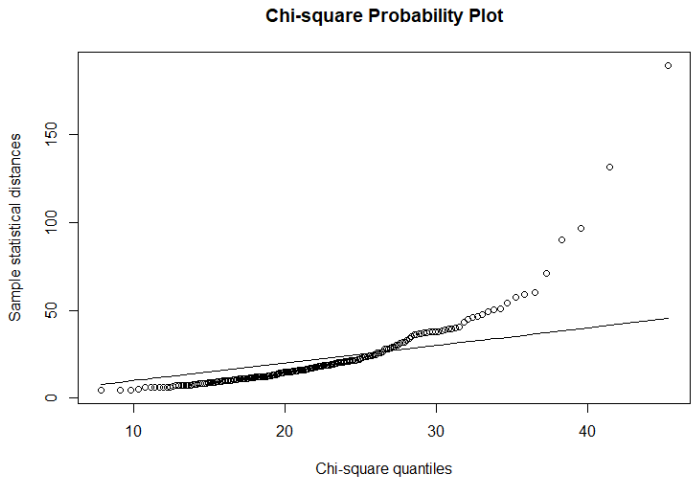


Figure 11: Chi square plot of ordered statistical distances, raw data.

Chi square QQ plot of our data resembles straight line through the origin with the slope 1, except the slight deviation towards the tail that make it appear curved upwards. Overall data seems reasonably normal.

1.4 Standardization of the data

Since medical parameters are measured in different units and depict different aspects of health conditions, the nature of the medical data is that it consists out of variables that can have large scale differences on the one hand and are moderately correlated on the other. To overcome the potential problem during Principal Component Analysis (and to mitigate influence of the outliers and unusual observations), it is beneficial to standardize the data. From each observation we subtracted the sample mean and divided by sample variance, which resulted in different (and more informative) graphs.

[3] All graphs are given in appendix.

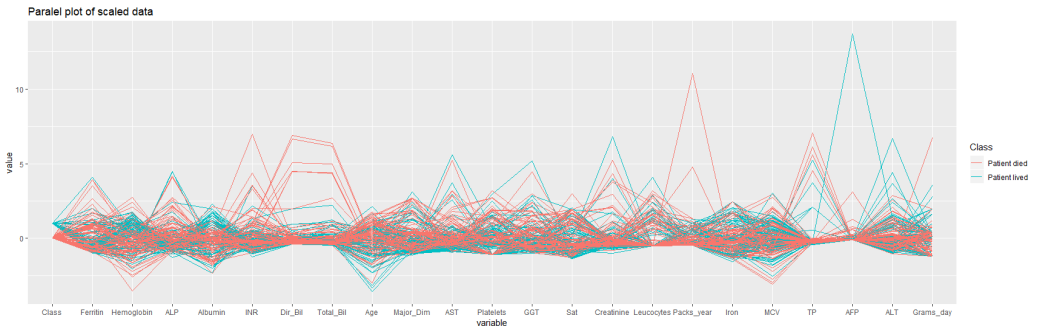


Figure 12: Parallel plot, standardized data.

$$\mathbf{Z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \mu)$$

Parallel plot of standardized data shows decreased influence of outlier in AFP variable, and draws attention to the other variables with potential unusual observed values. It also shows the obvious correlation between two bilirubin variables DirBil and TotalBil. By turning the option

```
order ="allClass"
```

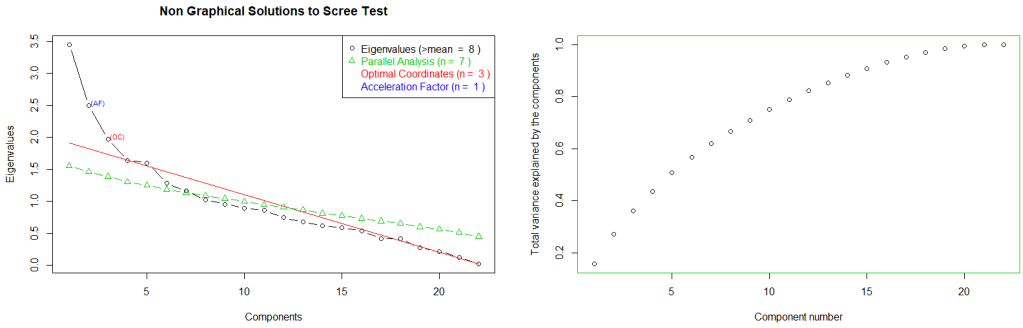
we sorted variables by F statistic from univariate ANOVA for comparing mean vectors of the two populations patients that lived and died during this research (variable Class has value 1 or 0). That gives us an incite on which variable has most influence, in this case that is variable Ferritin.

2 Analysis

2.1 PCA analysis

Principal Component Analysis of data set will show us potential for the data reduction, and, more importantly, connections between variables that will make the interpretation and understanding of the future inferences' results easier and more meaningful.

Scree plot shows us that we do have three very influential eigenvalues and an "elbow" between fourth and fifth eigenvalue, but after that gradual decrease in influence continues. Scree plot also shows us two different suggestions for the number of the components that we should consider (n=3



(a) Scree plot.

(b) Cumulative variance explained.

Figure 13: Principal Components Analysis.

and n=s). Those eigenvalues cumulatively explain only 36% and 61% of the data variance and since the majority of the variability probably comes from unusual values that those variables contain, reduction of the data will not bring great ratio of approximation and savings in computing time.

	PC1	PC2	PC3
Age	-0.09552388	0.063271338	-0.008188025
Grams_day	0.03896835	-0.140176481	-0.076840533
Packs_year	-0.04792292	-0.014248063	0.162669337
INR	0.12364683	-0.155078384	-0.389904123
AFP	0.04268349	0.265652607	0.116777092
Hemoglobin	-0.00477391	-0.074468133	0.475641546
MCV	0.17670603	-0.270549473	-0.006697182
Leucocytes	0.09739201	0.289635422	0.079951340
Platelets	0.08339509	0.365578948	0.106828297
Albumin	-0.15598601	0.026540929	0.423416874
Total_Bil	0.41492100	0.001997829	-0.195080324
ALT	0.29464702	0.014730136	0.095083515
AST	0.34401955	0.017394524	0.067944002
GGT	0.16036337	0.377070872	0.116291204
ALP	0.18034605	0.407994012	-0.082167014
TP	0.01814729	-0.048100420	0.164353060
Creatinine	0.18325022	0.087876373	-0.048172227
Major_Dim	-0.07939358	0.335710998	0.121125353
Dir_Bil	0.41814781	0.041481540	-0.176426945
Iron	0.23977048	-0.276716385	0.337237276
Sat	0.28983903	-0.272108519	0.303047218
Ferritin	0.32527823	0.017363423	0.160789967

(a) Principal components.

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.8580	1.5803	1.40518	1.28069	1.26657	1.13388	1.07785
Proportion of Variance	0.1569	0.1135	0.08975	0.07455	0.07292	0.05844	0.05281
Cumulative Proportion	0.1569	0.2704	0.36017	0.43472	0.50764	0.56608	0.61889

(b) Importance of the components.

Figure 14: PCA.

First principal component has dominant factor DirBil and TotalBil which are liver enzymes and together with other liver enzymes thei are contradicting Age and Hemoglobin and MajorDim (Major Dimension of the nodule). For further interpretation od the first principal component, field specific knowledge is needed. Contrary to that for the second principal component, we can deduce that it clearly contradicts blood related measurements to liver

enzyme measurements. In PC2 liver enzymes dominate with ALP.

In cocnlusion, PCA analysis may provide some insightful way to understand the data and to interpret results. In this case, it might prove as useful tool for field experts, but data dimensionality reduction wouldn't be adviseable in this case since cost of reducing the number of variables is too high in terms of loss of explained variability of the data.

2.2 Mean vector comparison for patient populations that lived or died

The dataset was divided into two populations based on the value of the variable "Class". Patients with "Class" value 1 grouped as "lived" and with 0 grouped as "died" and mean vectors of other variables are compared.

First of all, we usled Barlett test to test the hypothesis that two populations had the same underlying variance at significance level 0.05. Results of the test gave us enough evidence to reject H_0 and to conclude that the variances of two populations are not the same. We proceeded testing for equal mean under different variance condition. Since our sample was reasonably large, we used χ squared method to test equality of the means and since we have equal sample sizes, the large sample procedure is essentially the same as the procedure based on the pooled covariance matrix[4]. Since value of our test statistics was greater than χ squared quontile, we rejected H_0 and cocnluded that means of the two populations are not the same, or that on average patients that lived and died had different values for the blood and liver markers.

Looking at the intervals, we can deduce that, on average, people that lived are younger. There is no difference in alcohol and cigarette consumption (confidence interval contains 0, which means that no difference in means of populations is plausible). Same goes for INR, AFP, MCV, ALT and others (dominantely liver enzymes). People that lived had lower levels of INR, total and direct bilirubin, major dimension of nodule and Ferritin.

[4]R. Johnson, D. Wichern, *Applied Multivariate statistical analysis*

	Lower	Upper
Age	-0.7074522	-0.03215974
Grams_day	-0.3165692	0.37056877
Packs_year	-0.4414785	0.24488783
INR	-0.7636845	-0.09248946
AFP	-0.3024723	0.38458310
Hemoglobin	0.2686290	0.92439483
MCV	-0.2537021	0.43280621
Leucocytes	-0.5548230	0.12844332
Platelets	-0.6411880	0.03811885
Albumin	0.2203959	0.88090388
Total_Bil	-0.7229310	-0.04870210
ALT	-0.3735445	0.31357883
AST	-0.6927496	-0.01649144
GGT	-0.5982108	0.08324305
ALP	-0.8962440	-0.23735895
TP	-0.4061046	0.28075714
Creatinine	-0.5655216	0.11733053
Major_Dim	-0.7053121	-0.02987636
Dir_Bil	-0.7341822	-0.06075689
Iron	-0.2496697	0.43677529
Sat	-0.5895586	0.09228491
Ferritin	-0.9521215	-0.29960954

Figure 15: 95% Confidence intervals for the difference in vector means.

2.3 Mean vector comparison for patient populations that consumed and didn't consume alcohol

Similar to the previous analysis, we did the comparison of means between populations of patients that consumed and didn't consume alcohol. Both Barlett test and rule-of-thumb lead us to conclude that underlying variancesof two populations are not the same. Therefore, we used the same test which had similar result and we will conclude that patinets that consumed and didn't consume alcohol had different mean vectors for variables in question, or in other words, at least one variable mean was significantly different between two populations.

Looking at the data, we can conclude that, on average, people that did and didn't drink alcohol had the same age, and as per expectations, people from drinking population on average consumed more alcohol per day.

People that drank alcohol had significantly lower Hemoglobin and Albumin values (blood related parameters), and higher oxygen saturation of blood and MCV (mean corposcular volume) and INR. The rest of the parameters were the same.

2.4 Classification tree

We employed Classification Tree Algorithm (CART) mainly to find out which variable can be considered the most influential when classifying patients to

	Lower	Upper
Age	-0.085979267	0.7801835963
Grams_day	1.363456893	1.7260531333
Packs_year	-0.701636440	0.4189307210
INR	0.207456820	0.7436099608
AFP	-0.087932200	0.3876047091
Hemoglobin	-0.770341329	-0.0006275901
MCV	0.150165980	0.7677814033
Leucocytes	-0.591316227	0.1885336732
Platelets	-0.752064373	0.0206338873
Albumin	-0.715058418	-0.0045065034
Total_Bil	-0.612903483	0.3964992506
ALT	-0.394078531	0.3078868837
AST	-0.287539032	0.4851983115
GGT	-0.177637994	0.4967007380
ALP	-0.721853549	0.2013504130
TP	-0.857509233	0.2160825063
Creatinine	-0.608610600	0.2661656259
Major_Dim	-0.733769954	0.0499208909
Dir_Bil	-0.719925399	0.3231375981
Iron	-0.044941649	0.6846202632
Sat	0.003726396	0.7388221405
Ferritin	-0.343410396	0.4755957873

Figure 16: 95% Confidence intervals for the difference in vector means for patients that lived or died.

populations that lived or died according to this method.

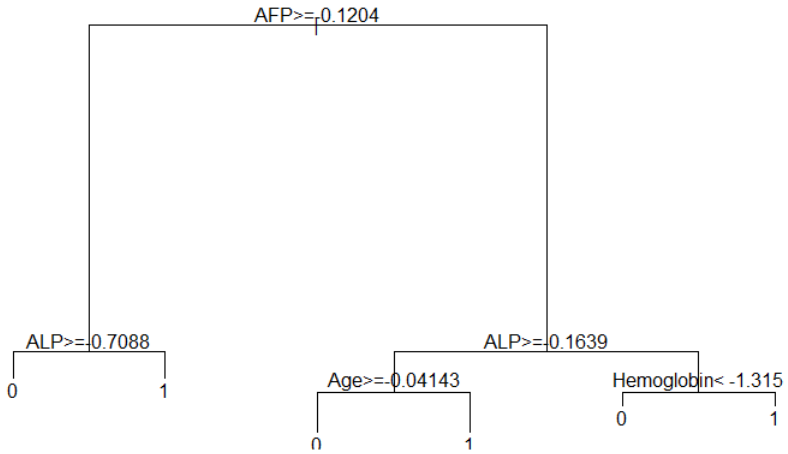


Figure 17: Classification tree.

So we fitted classification tree on whole set of variables to predict the

value of the class variable for patients that did and didn't consume alcohol.

As we can tell from the diagram, most important variable for classification by far is AFP (which if probably not because of unusual observations since AFP is dominant variable in the model even if we fit model on the dataset with mitigated extreme values). This model classified 88 observations correctly as 0 and 81 correctly as 1, which brings us to error rate of 17.16% which is reasonably good result.

2.5 Random forest algorithm

It was interesting to see that best performing random forest algorithm with four variables per node and 500 trees to vote had just a little smaller error rate of 16.67%. With slightly better performance on true zeroes and slightly worse on true ones. The most influential varaible was also AFP, but the second one was Hemoglobin and Albumin.

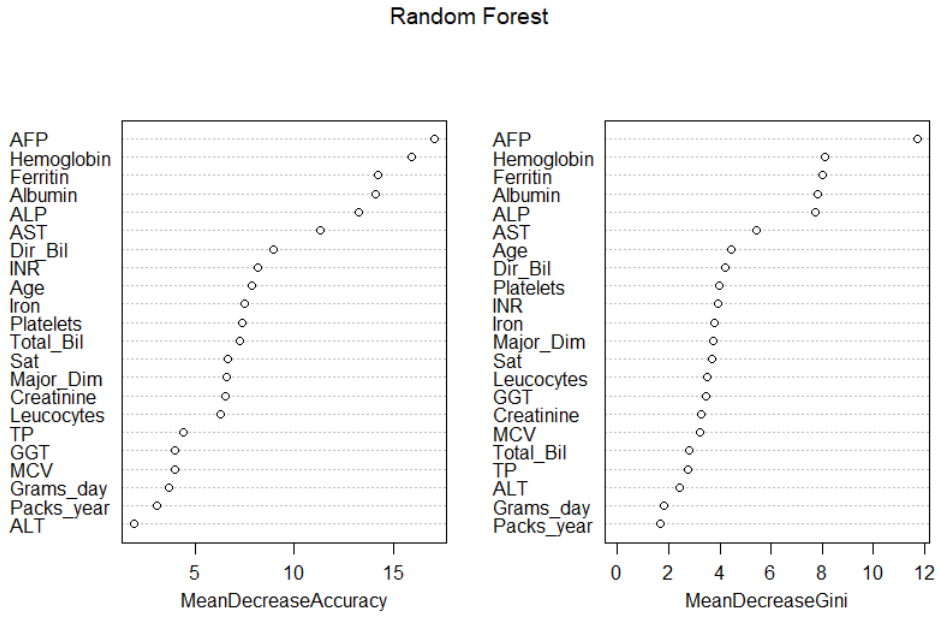


Figure 18: Random forest.

This variable importance plot tells us how removal of each variable will influence both error rate and Gini index the top variables carry more influence

and estiamte of their importance is given on x-axis. Therefore we can see that first seven variables from the topare clearly the most influential.

Gini index is the measure of contribution of each variable to the homogenity of the children nodes in comparison to the parent node. For displayed random forest Gini index is the average value of all Gini indexes for that variable in all trees that voted.

2.6 Quadratic classification analysis

Quadratic classification rule worked reasonably well, better than classification tree and random forest with apparent error rate (APR) of 13.73%.

Also, quadratic discriminant rule worked better on this dataset than linear discrimination rule which was to be expected since we have already established that variances of underlying distributions are not same.

Estimated prior probabilities were 0.5 for both classes, since we have same number of observations in both populations (data is balanced class-wise).

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 81 21
         1  7 95
```

Figure 19: Quadratic discrimination confusion matrix.

3 Conclusion

Our analysis provided us with insights in underlying distribution of the data, feature values of different subsets of our patient population and distribution of variability among principal components. It brought out differences in mean vectors and variances of patients that lived and the ones that died, that did and didn't consume alcohol. We were also able to find out which variables are the most important in classification based on the methods we used, tree and random forest algorithm. We discovered the classifying method that gave us most accurate classiification of the data and can potentially be used for similar datasets.

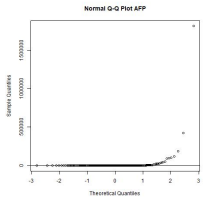
In conclusion, this analysis opened many questions which can be answered with future analysis of this health dataset, like cluster analysis and further factor analysis. Also, we disregarded 25 categorical and ordinal variables which can probably provide even more information that can be used in discrimination and classification. There are probably linear regressional relationships between variables or subsets of variables that we didn'r investigate in this project.

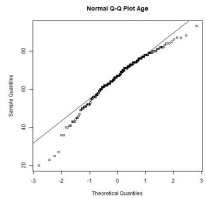
Appendix

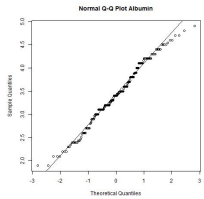
	Age	Grams_day	Packs_year	INR	AFP	Hemoglobin	MCV	Leucocytes	Platelets	Albumin	Total_Bil
Age	163.28	79.63	46.33	-0.17	-3.115271e+04	-2.05	3.25	2768.45	1.882259e+05	0.12	-9.61
Grams_day	79.63	3991.88	-54.42	7.40	1.146777e+04	-10.45	110.73	24609.03	-4.920233e+05	-4.71	-20.83
Packs_year	46.33	-54.42	1952.28	-0.51	1.784602e+05	12.95	12.57	3589.44	4.905436e+05	2.02	-13.90
INR	-0.17	7.40	-0.51	0.23	-4.046740e+03	-0.32	0.95	-76.64	-7.048900e+02	-0.10	0.61
AFP	-31152.71	11467.77	178460.16	-4046.74	1.719179e+10	9438.26	42086.57	56637146.65	3.275761e+09	-618.02	-2327.80
Hemoglobin	-2.05	-10.45	12.95	-0.32	9.438260e+03	4.75	2.20	151.92	2.472926e+04	0.57	-0.29
MCV	3.25	110.73	12.57	0.95	4.208657e+04	2.20	66.70	-1014.15	-1.247426e+05	-0.90	8.81
Leucocytes	2768.45	24609.03	3589.44	-76.64	5.663715e+07	151.92	-1014.15	7866698.85	1.008033e+08	-82.08	1576.65
Platelets	188225.87	-492023.25	490543.62	-704.89	3.275761e+09	24729.26	-124742.61	100803260.10	1.158710e+10	4174.49	78994.47
Albumin	0.12	-4.71	2.02	-0.10	-6.180200e+02	0.57	-0.90	-82.08	4.174490e+03	0.42	-0.85
Total_Bil	-9.61	-20.83	-13.90	0.61	-2.327800e+03	-0.29	8.81	1576.65	7.899447e+04	-0.85	34.27
ALT	-131.67	-131.51	-198.29	2.82	5.307813e+05	9.14	48.80	-8719.17	-2.260956e+04	-1.18	59.97
AST	-172.84	62.38	-227.60	0.24	1.864984e+04	10.37	108.41	2406.42	-1.054613e+05	-6.57	156.66
GGT	-124.46	1007.88	-488.96	-10.03	1.225795e+07	-39.11	-93.87	145845.06	7.091831e+06	2.15	56.82
ALP	-224.94	-821.95	-654.32	-2.20	2.146245e+06	-64.26	-264.42	128819.98	4.700607e+06	-21.92	186.96
TP	2.11	-31.25	19.34	0.11	-4.368971e+04	-0.07	-10.67	5039.15	3.668043e+04	0.06	-5.80
Creatinine	2.21	-4.27	-1.12	-0.02	-4.173910e+03	-0.09	-0.09	40.06	1.052029e+04	-0.01	1.75
Major_Dim	8.44	-45.08	5.90	-0.39	1.115870e+05	-0.43	-6.85	2472.13	1.084760e+05	0.28	-4.59
Dir_Bil	-6.94	-21.16	-10.07	0.34	5.427000e+02	-0.40	4.30	1567.23	7.873650e+04	-0.54	22.79
Iron	-92.57	473.82	126.83	0.19	-4.070221e+05	22.14	90.89	-5940.20	-4.690267e+05	0.54	59.15
Sat	-14.68	235.22	14.56	0.66	-2.565680e+05	6.91	68.89	-1030.47	-2.681916e+05	-1.18	34.11
Ferritin	297.44	1013.84	-411.62	4.28	4.309390e+06	41.09	740.19	269875.98	7.130535e+06	-11.95	813.58

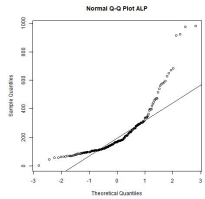
	ALT	AST	GGT	ALP	TP	Creatinine	Major_Dim	Dir_Bil	Iron	Sat	Ferritin
Age	-131.67	-172.84	-124.46	-224.94	2.11	2.21	8.44	-6.94	-92.57	-14.68	297.44
Grams_day	-131.51	62.38	1007.88	-821.95	-31.25	-4.27	-45.08	-21.16	473.82	235.22	1013.84
Packs_year	-198.29	-227.60	-488.96	-654.32	19.34	-1.12	5.90	-10.07	126.83	14.56	-411.62
INR	2.82	0.24	-10.03	-2.20	0.11	-0.02	-0.39	0.34	0.19	0.66	4.28
AFP	530781.30	18649.84	12257954.41	2146245.05	-43689.71	-4173.91	111587.03	542.70	-407022.09	-256567.97	4309390.27
Hemoglobin	9.14	10.37	-39.11	-64.26	-0.07	-0.09	-0.43	-0.40	22.14	6.91	41.09
MCV	48.80	108.41	-93.87	-264.42	-10.67	-0.09	-6.85	4.30	90.89	68.89	740.19
Leucocytes	-8719.17	2406.42	145845.06	128819.98	5039.15	40.06	2472.13	1567.23	-5940.20	-1030.47	269875.98
Platelets	-22609.56	-105461.28	7091831.22	4700607.07	36680.43	10520.29	108475.97	78736.50	-469026.73	-268191.58	7130534.81
Albumin	-1.18	-6.57	2.15	-21.92	0.06	-0.01	0.28	-0.54	0.54	-1.18	-11.95
Total_Bil	59.97	156.66	56.82	186.96	-5.80	1.75	-4.59	22.79	59.15	34.11	813.58
ALT	2805.98	3112.30	3203.04	1497.65	41.40	3.76	-23.81	41.96	359.94	241.78	6377.37
AST	3112.30	6664.98	4833.93	2682.73	-12.59	7.57	-24.43	108.20	726.46	446.45	11086.36
GGT	3203.04	4833.93	63343.57	23947.58	47.69	22.00	221.73	62.13	645.18	195.97	10378.19
ALP	1497.65	2682.73	23947.58	28511.19	-24.66	14.67	183.57	150.43	-583.39	-205.26	7857.43
TP	41.40	-12.59	47.69	-24.66	171.43	-0.91	-5.66	-2.26	175.18	56.58	-148.76
Creatinine	3.76	7.57	22.00	14.67	-0.91	0.90	-0.06	1.29	1.49	2.12	99.07
Major_Dim	-23.81	-24.43	221.73	183.57	-5.66	-0.06	23.35	-2.48	-36.19	-13.34	77.89
Dir_Bil	41.96	108.20	62.13	150.43	-2.26	1.29	-2.48	16.05	40.90	22.50	593.09
Iron	359.94	726.46	645.18	-583.39	175.18	1.49	-36.19	40.90	3069.76	1270.33	4589.83
Sat	241.78	446.45	195.97	-205.26	56.58	2.12	-13.34	22.50	1270.33	832.04	6170.99
Ferritin	6377.37	11086.36	10378.19	7857.43	-148.76	99.07	77.89	593.09	4589.83	6170.99	193995.23

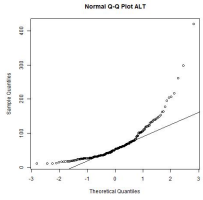
Figure 20: Variance covariance matrix.

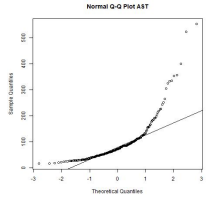

Normal Q-Q Plot AFP
Sample Quantiles
Theoretical Quantiles

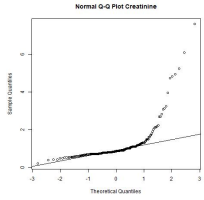

Normal Q-Q Plot Age
Sample Quantiles
Theoretical Quantiles

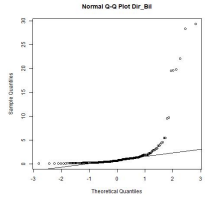

Normal Q-Q Plot Albumin
Sample Quantiles
Theoretical Quantiles

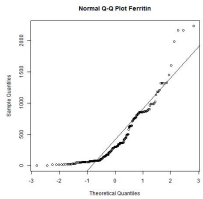

Normal Q-Q Plot ALP
Sample Quantiles
Theoretical Quantiles

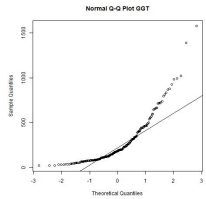

Normal Q-Q Plot ALT
Sample Quantiles
Theoretical Quantiles

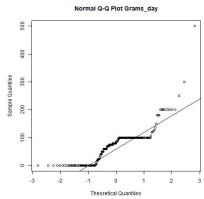

Normal Q-Q Plot AST
Sample Quantiles
Theoretical Quantiles

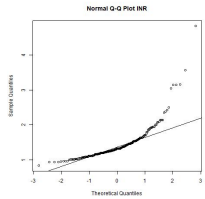

Normal Q-Q Plot Creatinine
Sample Quantiles
Theoretical Quantiles

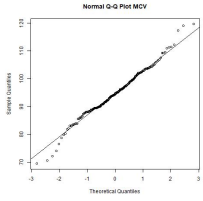

Normal Q-Q Plot Dir_Bil
Sample Quantiles
Theoretical Quantiles

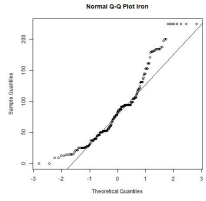

Normal Q-Q Plot Ferritin
Sample Quantiles
Theoretical Quantiles

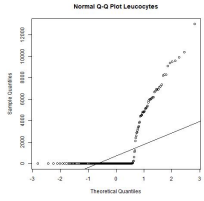

Normal Q-Q Plot GGT
Sample Quantiles
Theoretical Quantiles

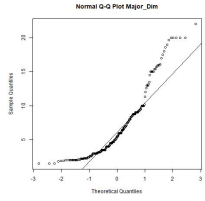

Normal Q-Q Plot Grams_day
Sample Quantiles
Theoretical Quantiles

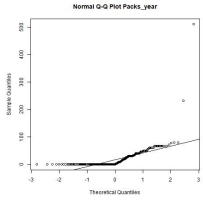

Normal Q-Q Plot INR
Sample Quantiles
Theoretical Quantiles

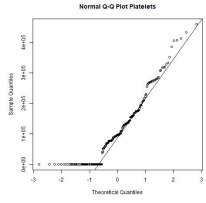

Normal Q-Q Plot MCV
Sample Quantiles
Theoretical Quantiles

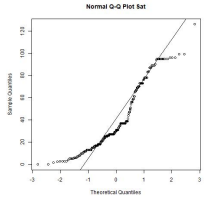

Normal Q-Q Plot Iron
Sample Quantiles
Theoretical Quantiles

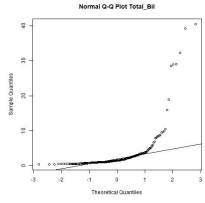

Normal Q-Q Plot Leucocytes
Sample Quantiles
Theoretical Quantiles


Normal Q-Q Plot Major_Dim
Sample Quantiles
Theoretical Quantiles


Normal Q-Q Plot Packs_year
Sample Quantiles
Theoretical Quantiles


Normal Q-Q Plot Platelets
Sample Quantiles
Theoretical Quantiles


Normal Q-Q Plot Sat
Sample Quantiles
Theoretical Quantiles


Normal Q-Q Plot Total_Bil
Sample Quantiles
Theoretical Quantiles

References

[1] R. Johnson, D Wichern. Applied Multivariate Statistical analysis 6th edition

[2] M. Bishop. Patern recognition and machine learning

[3] EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma

[4] Package 'rpart' documentation

[5] Package 'randomForest' documentation

[6] Package 'caret' documentation

[7] Package 'MASS' documentation