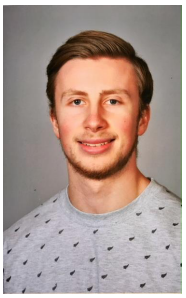
Technical University of Denmark

02450: Introduction to machine learning and data mining

Project 1: Indian Liver data set, description and visualization.

Authors:	**Studynumbers**
Mads Christian B. Andersen	s173934
Johannes Boe Reiche	s175549
August Semrau Andersen	s183918





Mads C. B. Andersen, August S. Andersen, Johannes B. Reiche

October 1, 2019

1 Introduction

1.1 Problem of interest

Half a million new liver patients are diagnosed every year in India, and the cause ranges from alcohol, obesity relateded disorders, Hepatitis B and C [1]. Hence an accurate and cheap way of treating and diagnosing patients is needed. In this project, a data set containing a collection of test samples from North East of Andhra Pradesh, India is analyzed to investigate the feasibility of creating machine learning models, that can be assist doctors in handling and diagnosig patients. The samples include both descriptive and medical measurements from Indian liver patients The data was obtained from Kaggle.com, uploaded by a Mr. Jeevan Nagaraj.

As mentioned the data could be beneficial in several different ways.
First off the data set could be used to train a classifier, that could diagnose people with having a liver disease or not. Solving this classification problem could be helpful in areas were there are no hepatologists available, but local doctors are able to obtain medical measurements. A properly trained model could potentially classify the patients as being sick and further which disease they might have. With several attributes and a known label for each patient, there is an immediate clear and straightforward setup for this machine learning task. There are several ways of handling this classification problem. One approach is to create a neural network, and train on some of the data available, and test on the remaining, however neural networks often require a lot of data, and 600 samples may not be enough to efficiently train the network. One solution could be to algorithmic expand the data set, but this requires advanced methods. Alternatively the classification problem could be solved by clustering the data and trying to see if it possible to cluster the two types of patients in a high-dimensional space.
Another use case for the data set could be to catch tendencies between attributes, that allow doctors to estimate medical values. This could be done with a multivariate regression model, with which predicting an a medical measurement (attribute) could substitute a potentially high-cost test. The increase in efficiency of diagnosis and lowering of costs associated with testing could potentially save lives.

1.2 Description of data sampling and processing

The data set has not been through a lot of processing, but there are several things worth mentioning regarding to the gathering and data handling process. When the data was gathered, patients whose age exceeded 89 years have been listed as 90 years. (This seems like a reasonable threshold since a medical measurements in the late stage of a persons life is heavily influenced by lifestyle and other factors, and therefor it would seem unlikely that a few years above the age of 90 itself should have a big influence). Another point worth mentioning is the classification of objectives as a patient or not has been done by professionals (Hepatologist). There is no correct definition of when one is a liver patient or not, therefor there is some element of subjectivity and a possible bias to this category.

2 A detailed explanation of the attributes of the data

The dataset contains 416 patients suffering from a liver-bound disease and 167 non-liver patients, totaling 583 samples. There are 10 different attributes and a label per person, where eight of the attributes are medical measures of different proteins and compounds. The last two attributes describe age and gender. All of the attributes is presented in Table 1:

Attribute:	**Type:**	**Describtion:**
age	Discrete, Ratio	Age in years
gender	Nominal	Gender (Male/Female)
tot_bilirubin	Continuous, Ratio	Amount of Direct and Indirect Bilirubin
direct_bilirubin	Continuous, Ratio	Amount of Direct Bilirubin
tot_proteins	Discrete, Ratio	Total proteins amount in patients blood
albumin	Discrete, Ratio	Albumin amount in patients blood
ag_ratio	Discrete, Ratio	Ratio between proteins Albumin and Globulin
sgpt	Continuous, Ratio	Alamine Aminotransferase amount in patient
sgot	Continuous, Ratio	Aspartate Aminotransferase amount in patient
alkphos	Continuous, Ratio	Alkaline Phosphotase amount in patient
is_patient	Nominal	Liver patient (Yes/No)

Table 1: Attribute types and descriptions.

2.1 Data issues.

A problem with the dataset is that the units with which the attributes are measured are not available from the original source. This doesn't directly cause any problems machine learning wise, as the algorithms work with numerical values and the units in any case are removed before the learning takes place. The problem is that the results that the models provide is rather difficult to infer from, since these depend on units. It is likely that the medical test to obtain the attributes are standardized and the units problem could possibly be resolved by consulting a doctor.

Since we use the attribute "*is_patient*" as a label for the patients, the attribute is saved in a single target vector. The dataset thereby only contains 10 attributes and one target vector y, bringing the dimensions of the data set down to $579 \cdot 10 = 5790$.

Another transformation worth mentioning is performed on the gender attribute, which is origionally formatted as text either as "male" or "female". This is changed as the machine learning model has no way of interpreting text. We handle text attributes with one-out-of-K encoding, which is performed by removing the column with string values and adding two separate columns

with proxy variables (dummy variables). In the first column, the value is 1 if the sex is male and 0 if female. The second column does the opposite (male: 0, female: 1). Having the two columns is in reality not necessary, since the value of the first column infers the second, but they are kept for programming purposes.
A further note is that there were four missing *alkphos* measurements which were replaced by NaN (not a number) in the original data set. These four patients have been removed as we lack knowledge for replacing the values with proper substitute values.

2.2 Summary statistics of the attributes

Table 2 gives an overview of the attributes' statistical mean μ, standard deviation σ, range (Min - Max) and quartiles $Q1, 50\%, Q3$:

Attribute:	**Mean μ:**	**STD σ:**	**Range:**	**Q1**	**50%**	**Q3**
age	44.782	16.222	4.0 - 90.0	33.0	45.0	58.0
gender	NaN	NaN	Male/Female	NaN	NaN	NaN
tot_bilirubin	3.315	6.228	0.4 - 75.0	0.8	1.0	2.6
direct_bilirubin	1.494	2.816	0.1 - 19.7	0.2	0.3	1.3
tot_proteins	291.366	243.562	63.0 - 2110.0	175.5	208.0	298.0
albumin	81.126	183.183	10.0 - 2000.0	23.0	35.0	61.0
ag_ratio	110.415	289.850	10.0 - 4929.0	25.0	42.0	87.0
sgpt	6.482	1.085	2.7 - 9.6	5.8	6.6	7.2
sgot	3.139	0.794	0.9 - 5.5	2.6	3.1	3.8
alkphos	0.947	0.320	0.3 - 2.8	0.7	0.9	1.1
is_patient	NaN	NaN	Yes/No	NaN	NaN	NaN

Table 2: Statistical summary of attributes

As mentioned earlier a large portion of the datasets attributes are medical measurements. From the data we see that some of these medical measurements have a relatively large spread, see Table 2. Attributes such as *tot_bilirubin*, *albumin*, *tot_proteins* and *ag_ratio* have quartiles varying with magnitudes in size. From these attributes we observe quite a few outliers which can be seen in Figure 2a and 2b. An example of this is the largest measurement of *ag_ratio* measureing 4929, when compared to the $Q3$ (75%) of 87 seems way too large. In this case it is hard for interpreters of the data (us) without a medical background to know whether these outliers are indeed flawed, or if the large variation in measurements is plausible.
It is also important to take into account that more than a forth of the patients in the dataset are not classified as liver patients at the same time as 76% of the patients are male.
The attributes *sgpt* and *sgot* that seem to be close to normally distributed, and they also a smaller range of values. This can be seen in plot 7.

3 Data visualization

As mentioned before the medical measurements have a large spread, this is visualized in Figure 1. By inspecting the boxplot Figure 1, it is clear that the attributes all operate on very different scales.

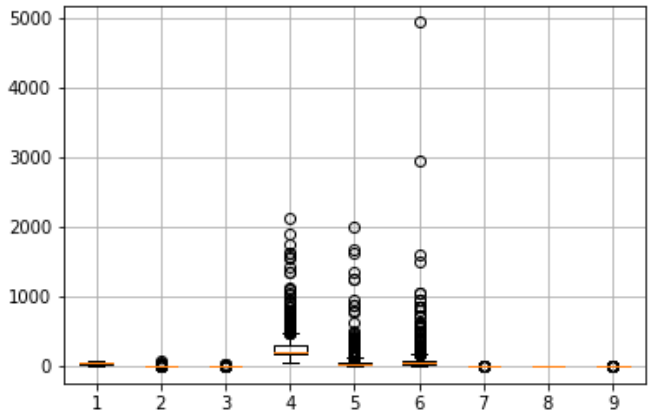
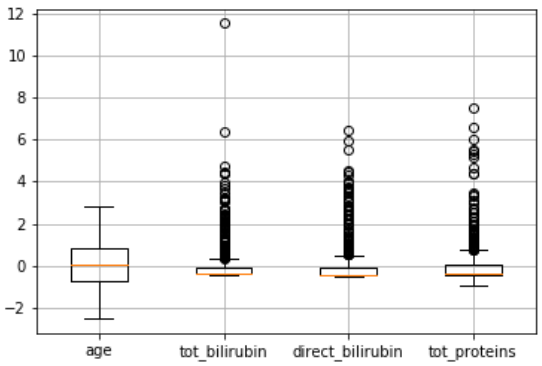
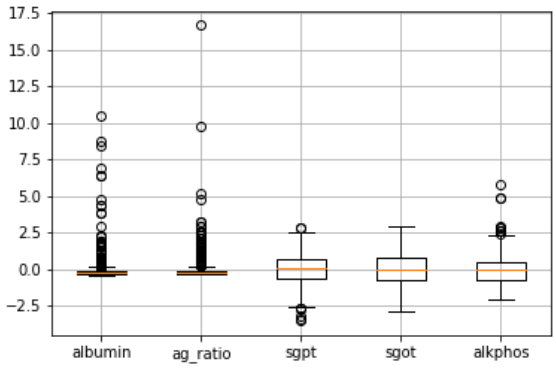


Figure 1: Boxplot of each attributes raw data

In spite of this, it is possible to see that there appears to be some outliers for some of the attributes. In order to inspect the raw data more clearly we standardize by centering and normalizing, more about this later in the section about PCA in equation 4. Plot 2a and 2b show the new standardized data set:



(a) First 4 attributes



(b) Last 5 attributes

Figure 2: Boxplot of each attributes standardized data

We choose to keep the outliers in the dataset, as the extremely high values of etc. *ag_ratio* or *tot_bilirubin* all belong to patients that are labelled as having a liver disease. Therefor the high values might be caused by their illness, or it might just be an error performed by the doctor, we dont know.

There are roughly 3 levels of how well the attributes from our dataset can be described as being normally distributed. These 'typical' levels are visualised in Figure 7.

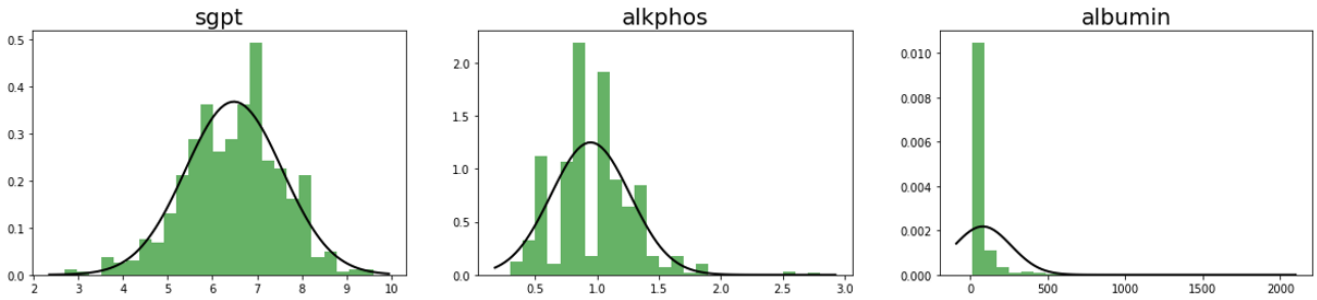


Figure 3: *Three levels of how well the attributes can be approximated as the normal distribution. Histograms fitted with a normal distribution.*

The histogram of *sgpt* seems to be be almost normally distributed, and *sgot* and *age* seem to also fit this description. *alkphos* can very roughly be said to be normally distributed. The remainder of the attributes however, (see Figure 7 in appendix) seem to be heavily distributed over a small range of values, and cannot be approximated as normally distributed. When the patients and non patients are plotted separately their distributions vary only slightly and the same attributes can be said to approximate the normal distribution.

3.1 Correlation

To evaluate whether the attributes are correlated, a correlation-matrix is used wherein each value describes the correlation between two attributes and the values of the diagonal are the attributes correlation to themselves, ie. 1.
In the correlation-matrix Figure 4, it is seen that there are some attributes that have numerical correlations above 0.5, which makes them a tad more interesting than the others. However, some of the correlations are expected, as the attributes are directly linked. Examples could be $cor(direct\_bilirubing, tot\_bilirubin) = 0.87$ and $cor(ag\_ratio, albumin) = 0.79$, where *direct_bilirubin* is a part of the *tot_rubin*. Same goes for the *ag_ratio* and the amount of *albumin* where *albumin* is one of the two measured values in *ag_ratio*.
Furthermore the $cor(sgot, sgpt) = 0.78$, for which the amount of Aspartate Aminotransferase correlates with the amount of Alamine Aminotransferase.
The last significant correlation is $cor(sgot, alkphos) = 0.69$, that is the correlation between the amount of Aspartate Aminotransferase and Alkaline Phosphotase.

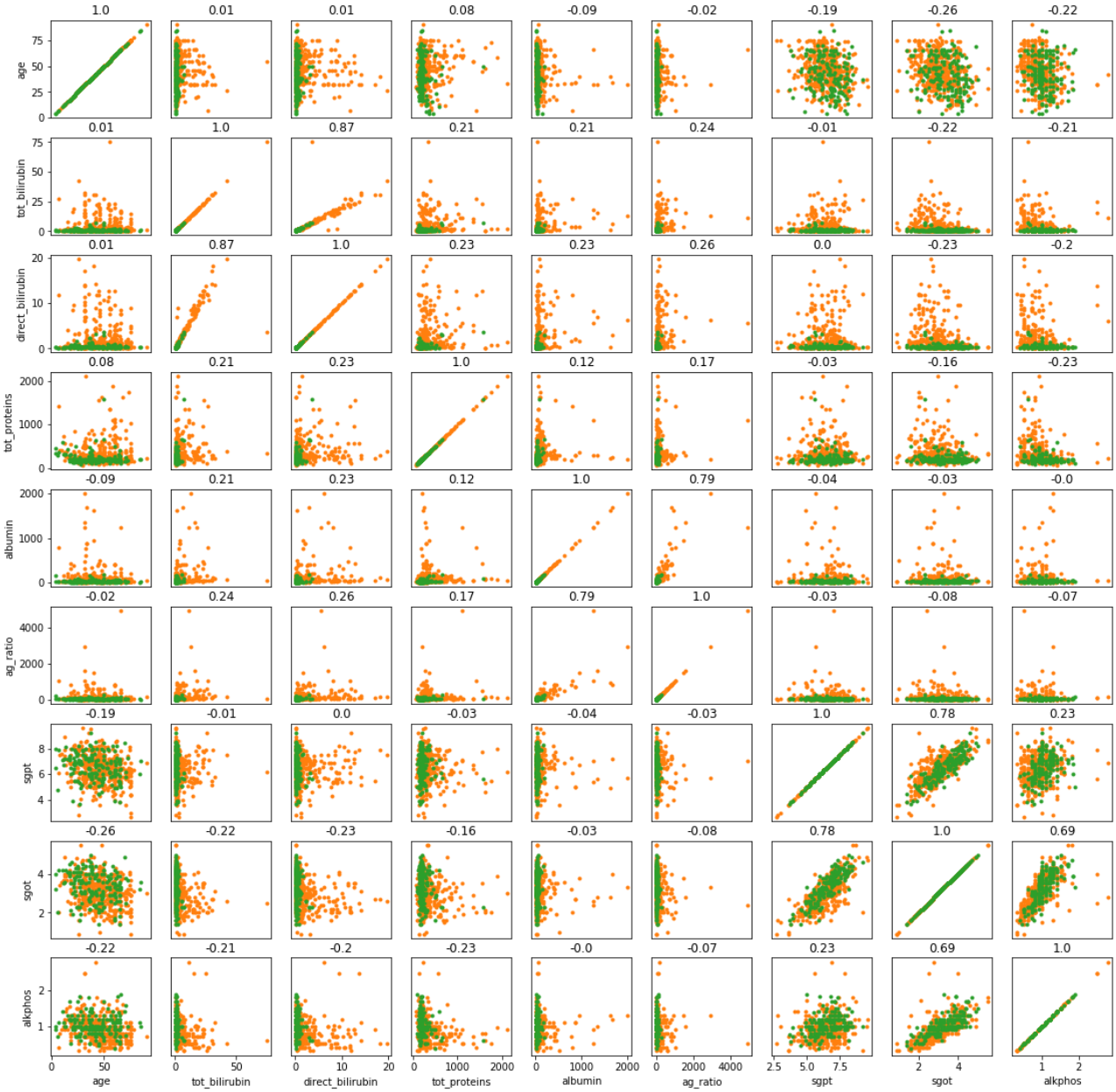


Figure 4: The 9 attributes plottet against each other, subtitle values being correlation.

3.2 PCA

With 11 original attributes, where the *is_patient* attribute is used as target, and adding an extra column with the one-out-of-K-coding for the gender attribute, the dimensions for the data set X, where X is an N x M matrix, is given by:

$$N = 579 \qquad M = 11 \tag{1}$$

Therefore, the feature space is 11-dimensional, making visualisations of every aspect of the dataset difficult. This is very unfortunate, since visualization is one of the most powerful tools for attaining an understanding of the data in general. Plotting individual attributes against each other can be helpful in some cases, but it does not give a complete impression of the

data. One method of getting a potentially meaningful representation of the entire, or at least a large part of the dataset is by reducing the dimensionality by performing principal component analysis (PCA). PCA is performed by creating a new basis in the feature space wherein the eigenvectors, also called principal components, are designed to retain as much of the variance as possible, and thereby retaining as much information from the data set as possible.
The eigenvectors are computed by 'Singular Value Decomposition' given in (2):

$$U\Sigma V^T = X \tag{2}$$

Where V contains the M eigenvectors. By creating a new basis of the two first principal components (eigenvectors), it is possible to project the original data onto them, and thereby visualising the whole data set in two dimensions. The projection of a single observation into this subspace is shown in (3):

$$b_i^T = x_i^T V_2 \tag{3}$$

For classification algorithms, this PCA plot can be helpful in visualising how the classes can be separated in a 2D space. Before applying the PCA, the data has been standardized by subtracting the mean of each column and dividing by the standard deviation. This is done mathematically in equation (4):

$$\tilde{x}_{ij} = \frac{x_{ij} - \frac{1}{N}\sum_{i=1}^{N} x_{ij}}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N} \tilde{x}_{ik}^2}} \tag{4}$$

This way, the data will be centered with mean of zero and normalized, hence the scale of the attributes is of very different. A rule of thumb is that for a PCA to be useful, the variance retained after projecting the dataset onto the new principal components should be at least 80%. The mentioned variance explained by principle components is calculated by (5):

$$\text{Variance Explained} = \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2} \tag{5}$$

As Plot 5 below shows, the variance explained by the first two principal component is well below the 80% mark:

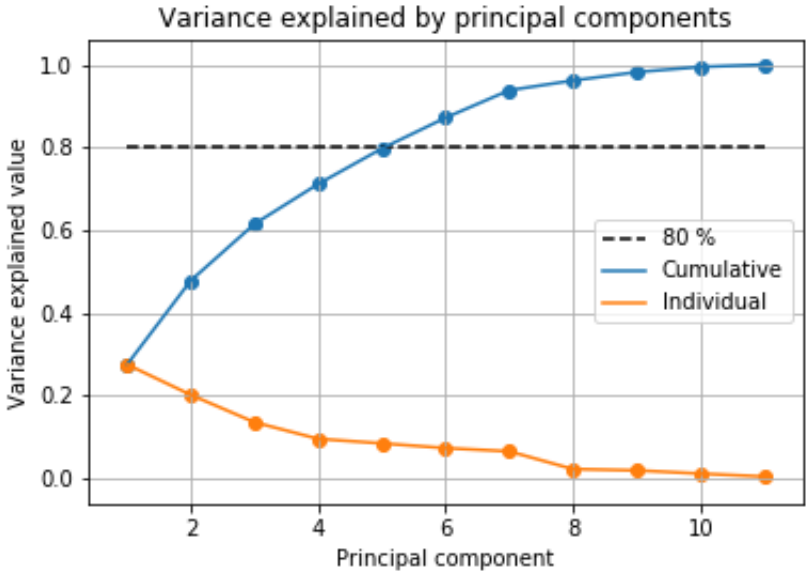


Figure 5

The plot also reveals that five principal components are required to reach the threshhold of 80%. This is troublesome and the PCA plot will therefor not be very useful in our case. Another way to analyse the results of the PCA is to look at the projection of the data in this 2-dimensional subspace, illustrated in Plot 6:

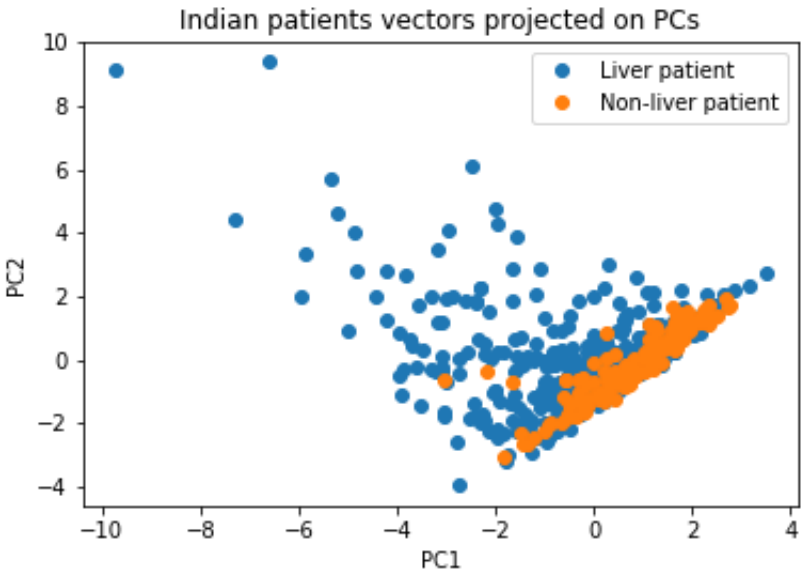


Figure 6

This plot shows that the liver patients and non-liver patients are not separated into clusters. It seems like there is a very clear threshhold in the bottom right corner. We think that a possible explanation to this could be a medical-compound threshhold that makes it 'impossible' to have a combination of biomarker-values that would get a PC1 value above -2 while the PC2 value isn't below a certain value. Furthermore, there also seems to be a part of the 'liver patient' group that get extreme attribute values that no 'non-liver patients' has - values below -2 in PC1 and above 2 in PC2. The PCA is therefore not completely useless, because there are tendencies to be found in the plot and with interpretation of the principal components, it is possible to infer what attribute may cause people to have a higher chance of being classified as a liver-patient. The two principal components are not meaningful by themselves, but by looking at the linear combination of the original attributes, it is possible to interpret which attributes set the patients apart from the non-patients. The linear combination of the first two principal components are shown in Table 3:

Principal directions:	**PC1**	**PC2**
age	-0.14	0.03
female	-0.29	-0.00
male	-0.02	0.01
tot_bilirubin	0.41	0.05
direct_bilirubin	-0.83	0.04
tot_proteins	0.16	-0.03
albumin	-0.10	0.04
ag_ratio	0.03	0.99
sgpt	-0.05	0.01
sgot	0.00	0.04
alkphos	0.01	-0.02

Table 3: Principal directions in terms of features

As mentioned, a negative value in the PC1-direction seems to increase the chance of being a liver patient. The main contributor for a negative value in the PC1-direction is *direct_bilirubin*. On the other hand, the attribute *tot_bilirubin* will have a positive contribution the PC1-direction. There is a direct relation between the two attributes, so it seems like a medical-biological analysis is needed to give an explanation. For the PC2-direction, a higher positve value seems to increase the chance of being a liver patient. In this direction there is only attribute that has an impact of matter (compared to the rest), the '*ag_ratio*'. This attribute therefore seems to have a big influence on whether a patient have higher risk of being a liver patient or not. One thing to mention about the *ag_ratio* is that it is also the attribute with the highest standard deviation and biggest range of values, see table 2 and plot 2b. This could maybe partly explain why *ag_ratio* is an important attribute, based on the statistics and PCA.
On top of the data visualization, a 2D-clustering machine learning model would only be able to differ between the patients with high values of *direct_bilirubin* and *ag_ratio*, based on the principal components from the PCA. Based on plot 5 it seems that this should at least be done in a 5-dimensional space, and it will be interesting to see how much dimensions could be brought down and still do a good classification job. The visualizations doesn't really reveal any significant aspects regarding to a classification problem with a neural network, so it remains unclear what the effects of this machine learning could have. From plot 4 it was possible to se that some attributes have high correlation, why a multivariate regression model seems to have some potential to be able to predict certain attributes.

4 Conclussion

By processing the Indian liver data set in several ways, primarily calculating descriptive statistics and creating visualisation of different aspects of the data set, it has been possible to create an understanding of the dataset. The preprocessing and error handling allowed us to get a qualitative understanding of the dataset. The dataset was in a general good shape, however there were a few missing values and several outliers. A more general problem was the lack of units, but as mentioned this has no machine learning implications and it could be resolved by consulting a doctor. Furthermore it has been possible to get a grasp of the content of the dataset,

first by calculating descriptive statistics such as the mean, quartiles and maximum values and secondly by visualizing the attribute distributions and their pairwise dependence. Another aspect was that the attributes operated on very different scales, why they where standardized and normalized before further data analysis. The last important thing that was gained insights of was how feasible the dataset seem to be for the machine learning algorithms we want to apply to them. For classification the PCA did not allow a 2D space with clearly separated classes (patients/non-patients), which means that to classify people, more dimensions are necessary. Luckily, the 11 dimensional vector space is available and it should be possible for the machine learning models to learn data patterns that can distinguish the patients from the non-patients.

References

[1] D. A. Sahney, *Is liver disease the next major lifestyle disease of india after diabetes and bp?* https://timesofindia.indiatimes.com/life-style/health-fitness/health-news/is-liver-disease-the-next-major-lifestyle-disease-of-india-after-diabetes-and-bp/articleshow/58122706.cms.

A Bilag

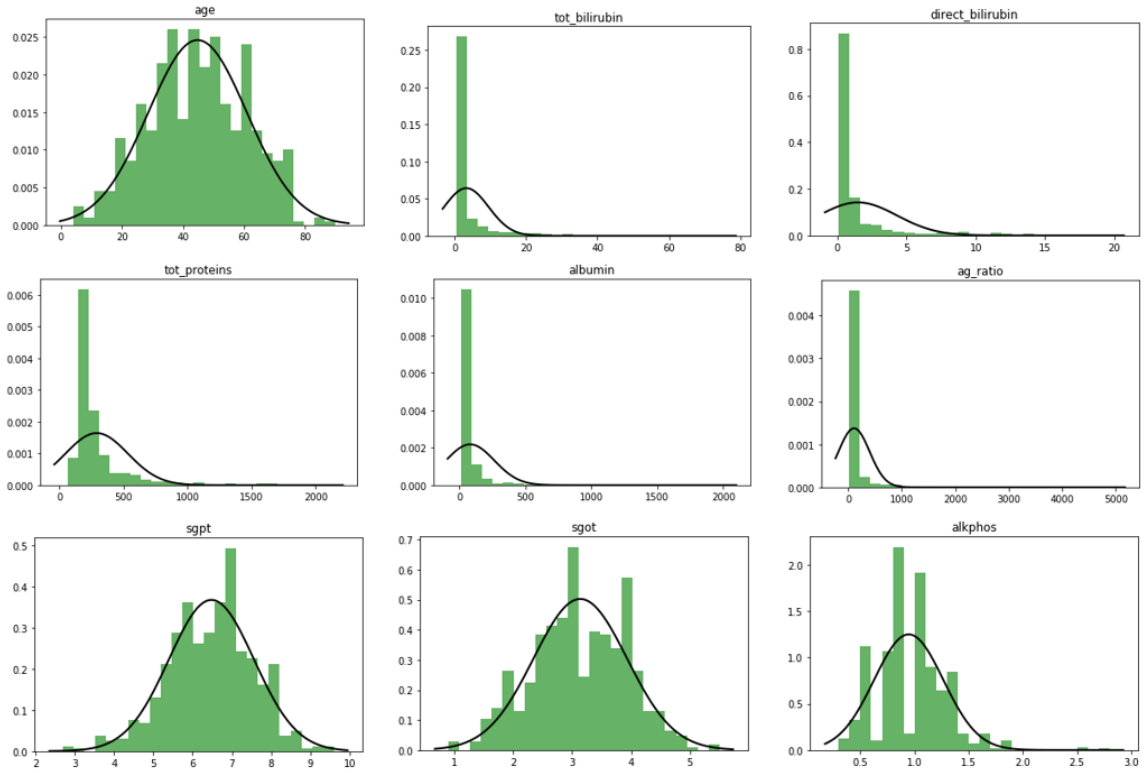


Figure 7: Histograms of the data with a normal distribution plotted on top of it.

Table 4: Responsibility table

Section\Name	Mads C. B. Andersen S173934	August S. Andersen S183918	Johannes B. Reiche S175549
1.1 Problem at interest	X	O	O
1.2 Description of data sampling and proccesing	X		
2.0 Detailed explanation of dataset.		X	
2.1 Data issues	X	O	O
2.2 Summary Statistics	O	X	O
3.0 Data visualization	O		X
3.1 Correlation		X	
3.2 PCA	O		X
4. Conclusion	X	O	O
Editing	O	O	O

The person in charge of a section is marked with a 'X' and have been at least 50% responsible for the section. Co-editor(s) have received an 'O'.