Introduction to Machine Learning and Data Mining: Project 1

Jens Leysen (s191908)
Maud Leclerc (s191975)

September 2019

Everyone (that hasn't been living under a rock) has seen James Cameron's 1997 blockbuster "Titanic", starring the young and handsome Leonardo DiCaprio and Kate Winslet. Although most people are only interested in the dramatic aspect of "modern history's deadliest peacetime commercial marine disaster", we are mainly interested in finding underlying patterns in the survivability of the different groups of passengers and predicting whether a certain person would have been more likely to survive or die.

The death-toll is astonishing: 1502 out of 2224 passengers and crew were killed. A large portion of these untimely deaths could have been avoided would there have been more lifeboats. Although there was probably some luck involved in surviving the disaster, we are interested in finding some common traits between the passengers, that could explain why they survived or not.

1 Description of the Dataset

This data set was obtained from Kaggle. It contains 891 observations (distinct passengers), together with attributes such as Age, Sex and whether they survived or not.

The most used source for any data about the Titanic is the Encyclopedia Titanica. The original source for this data set is Eaton & Haas (1994) Titanic: Triumph and Tragedy, Patrick Stephens Ltd. This book includes a passenger list created by many researchers and edited by Michael A. Findlay.

We are mainly interested in predicting whether a certain person would have been more likely to survive or die. The machine learning problem would then be of the classification type. We are given the attributes of one passenger and want to build a model that predicts whether a passenger was likely to survive or not. This is a binary classification problem since there are only two choices (survives or dies).

Data Dictionary

Variable	Definition	Key
PassengerId	Id number	0 to 891 for our data set
Survival	Survived or not	0=No,1=Yes
Name	Name of passenger	
Pclass	Ticket Class	1=1st, 2=2nd, 3=3rd
Sex	Sex	male, female
Age	Age in years	
Sibsp	Number of siblings/spouses aboard	
Parch	Number of parents/children aboard	
Ticket	Ticket Number	
Fare	Passenger Fare	
Cabin	Cabin Number	
Embarked	Port of Embarkation	C=Cherbourg, Q=Queenstown, S=Southampton

Attributes relevant to different machine learning tasks:

1. Classification:

   Relevant Attributes: Survived, Pclass, Sex, Age, SibSp, Parch, Fare

   Irrelevant Attributes: PassengerId, Name, Ticket, Embarked, Cabin

   Explanation: We are given the attributes of one passenger and have to determine whether they were likely to survive or not. The relevant attributes are attributes that can have an impact on the chance of survival of the passenger, other ones such as Name or Ticket number will not have an impact.

2. Regression:

   Relevant Attributes: Survived, Pclass, Sex, Age, SibSp, Parch, Fare

   Irrelevant Attributes: PassengerId, Name, Ticket, Cabin, Embarked

   Explanation: The fare values are described by a float between 0 and 512.3292. A regression problem would for example predict the fare of a certain passenger based on known relevant attribute values.

3. Clustering:

   Relevant Attributes: Survived, Pclass, Sex, Age, SibSp, Parch, Fare

   Irrelevant Attributes: PassengerId, Name, Cabin, Ticket, Embarked

   Explanation: The goal of clustering is to find similarity, to find different groups in the data set. A possible group we can think of (and might find) might be employees/passengers. Ignoring the Pclass attribute, we might be able to group people according to class using other attributes or visualize other underlying groups.

4. Association Mining:

   Relevant Attributes: Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked

   Irrelevant Attributes: PassengerId, Name, Ticket, Cabin

   Explanation: Association mining is about finding rules that approximately hold true. We would be looking for rules such as: ”Given the person is male and travelled 1st class, he was likely to survive.”

5. Anomaly detection:

   Relevant Attributes: Survived, Pclass, Sex, Age, SibSp, Parch, Fare

   Irrelevant Attributes: PassengerId, Name, Cabin, Ticket, Embarked

   Explanation: Anomaly detection consists in finding which observations deviate significantly from the rest of them. This can mean, for instance, finding out if a passenger is very different from the others, according to his attributes.

Data Issues and Data Operations By close inspection of the data set, we found some fare values equal to zero. After some research we found out that these values indicate that these passengers were servants of another passenger, or represented a special group for example the H&W Guarantee Group (Harland and Wolff employees chosen to oversee the smooth running of the Titanic's maiden voyage). We assume these persons don't represent an incorrect value but rather that they didn't pay for the trip themselves. Two examples are Mr.William Henry Harrison and Mr.Richard Fry, who were the secretary and valet of Mr.Joseph Bruce Ismay. [1]

The Encyclopedia Titanica clearly states that the H&W employees don't have fare values because they are employees of the titanic. [2]

The fare values may look odd, that is because it is per Britain's pre-decimalised currency. There were 20 shillings in £1, 12 pennies in 1 shilling and thus 240 pennies in £1. This system was used in much of the British Commonwealth until the 1960s and 1970s. [5] [6] The conversion formula to a decimalised value is: $\mathrm{Y} = \mathrm{P}+5\cdot(0.01)s + \frac{5}{12} \cdot (0.01)d$.

Where: Y is the decimalised amount in pounds, P is the number of pounds, s is the number of shillings and d is the number of pennies. [4]

The raw data is quite detailed, but is missing some information. Moreover, some of the passengers' attributes are not useful for this analysis. Firstly, the passengers' name, Id number and ticket number are irrelevant for the analysis. Secondly, the cabin number, although useful to know which deck the passenger was on, is missing for a lot of passengers and thus not very useful. We selected only the data that seemed of interest to explain the survival: the passenger's sex, age, class, number of siblings or spouses aboard, number of relatives aboard, the fare they paid, and where they embarked. All the passengers with missing data on these attributes were then removed from the data set, leaving 712 passengers.

Later, in order to do a principal component analysis, Pclass, Sex and Embarked will be coded using an one-out-of-K coding.

2 Explanation of the attributes of the data

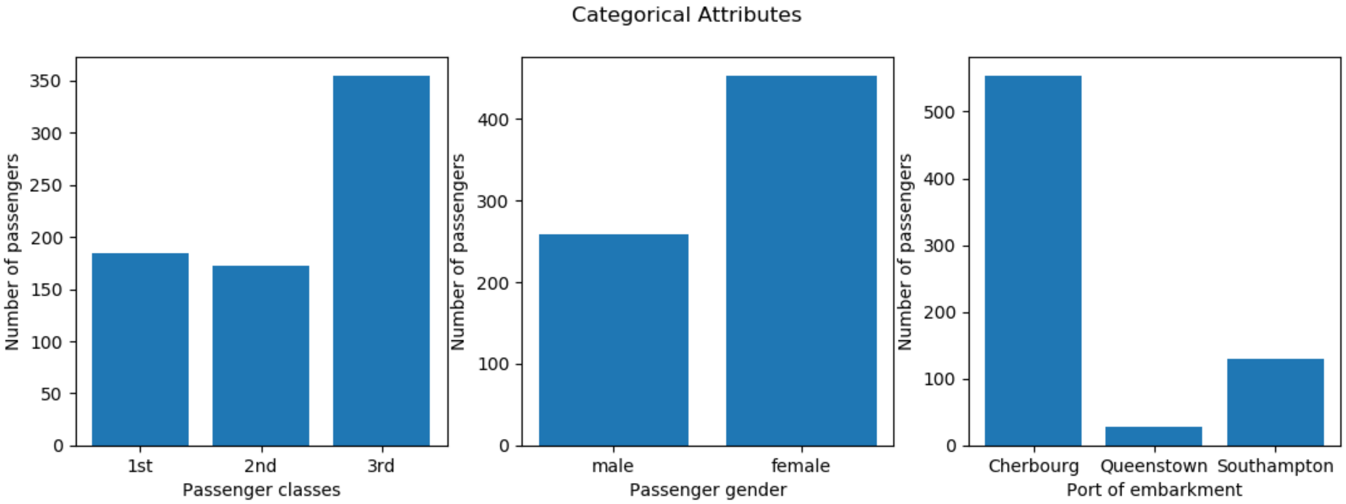
Attribute Description

Attribute	Type	Type
PassengerId	Discrete	Nominal
Survival	Binary	Nominal
Name	Discrete	Nominal
Pclass	Discrete	Ordinal
Sex	Binary	Nominal
Age	Discrete	Ratio
Sibsp	Discrete	Ratio
Parch	Discrete	Ratio
Ticket	Discrete	Nominal
Fare	Discrete	Ratio
Cabin	Discrete	Nominal
Embarked	Discrete	Nominal

Summary statistics of the attributes Note: Values have been rounded to two decimal places for readability. The exact values can be found by running our code.

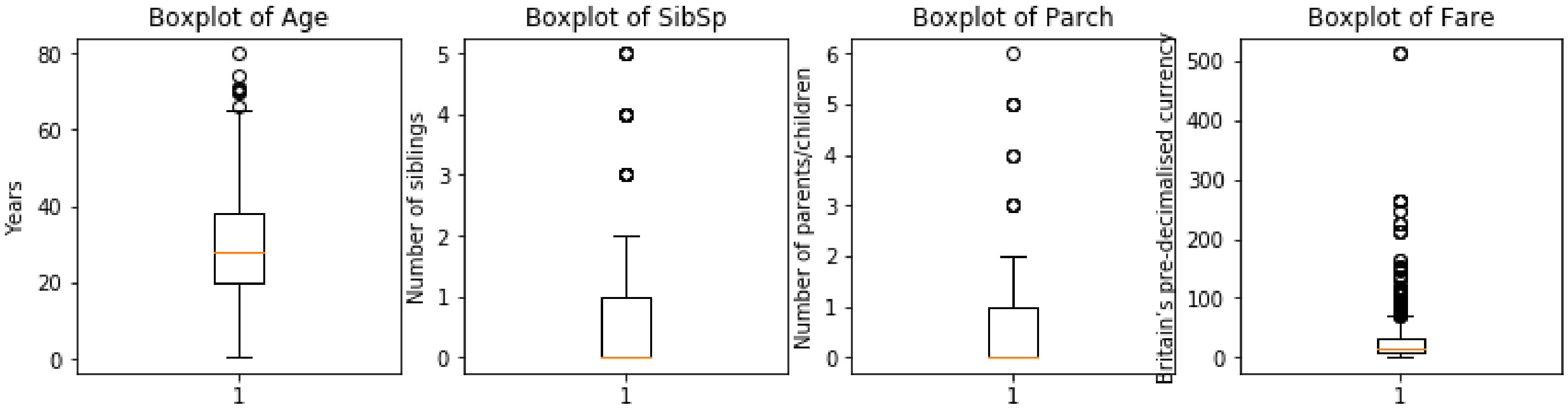
Attribute	Mean	Standard deviation	Median	Range
Age	29.64	14.49	28.0	79.58
SibSp	0.51	0.93	0.0	5.0
Parch	0.43	0.85	0.0	6.0
Fare	34.57	52.94	15.65	512.33

Summary of categorical attributes

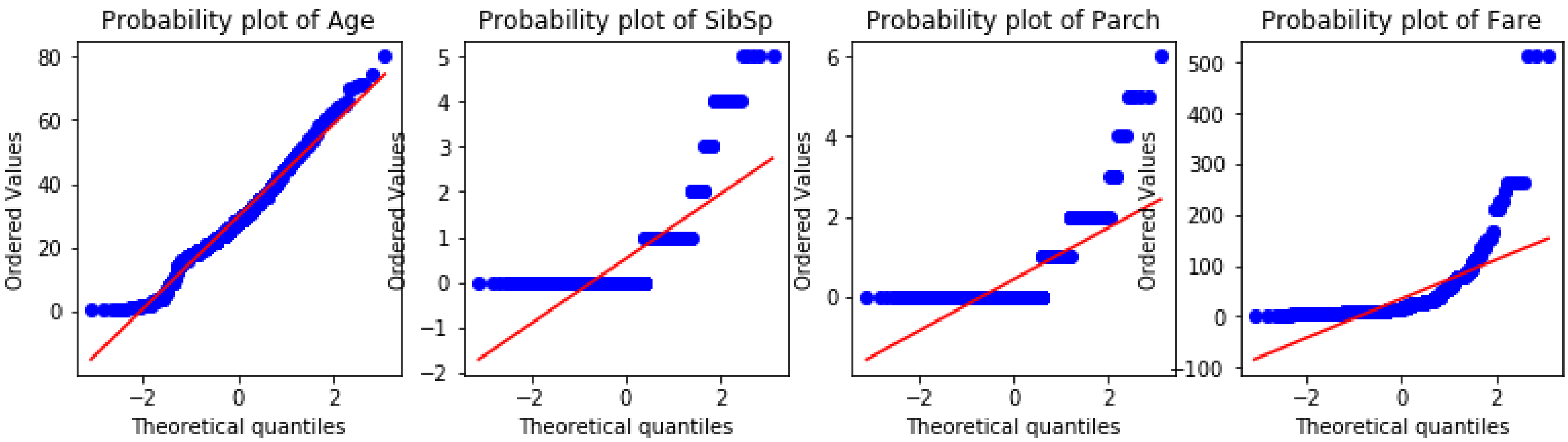


3 Data visualizations and PCA

Boxplots The fare boxplot is the most interesting one, the maximum fare value encountered is 512.3292 (as noted in the table above). It's stated in the description of Mr Thomas Drake Martinez Cardeza that he occupied one of the two most luxurious suites on board (B51/3/5, ticket 17755, £512, 6s). There are no strong reasons to think that the outliers represented in the boxplots below are erroneous values and we will thus not remove any of these outliers. [3]



Normal Distribution Analysis We used a normal probability plot as a graphical method for comparing the relevant attributes to a normal distribution. The points follow a strongly linear pattern for Age, while the other 3 attributes have a very large deviation from a linear pattern. The fare data is clearly right-skewed because of its exponential curve. We thus conclude that the Age attribute can be approximated by a normal distribution while the others can't.

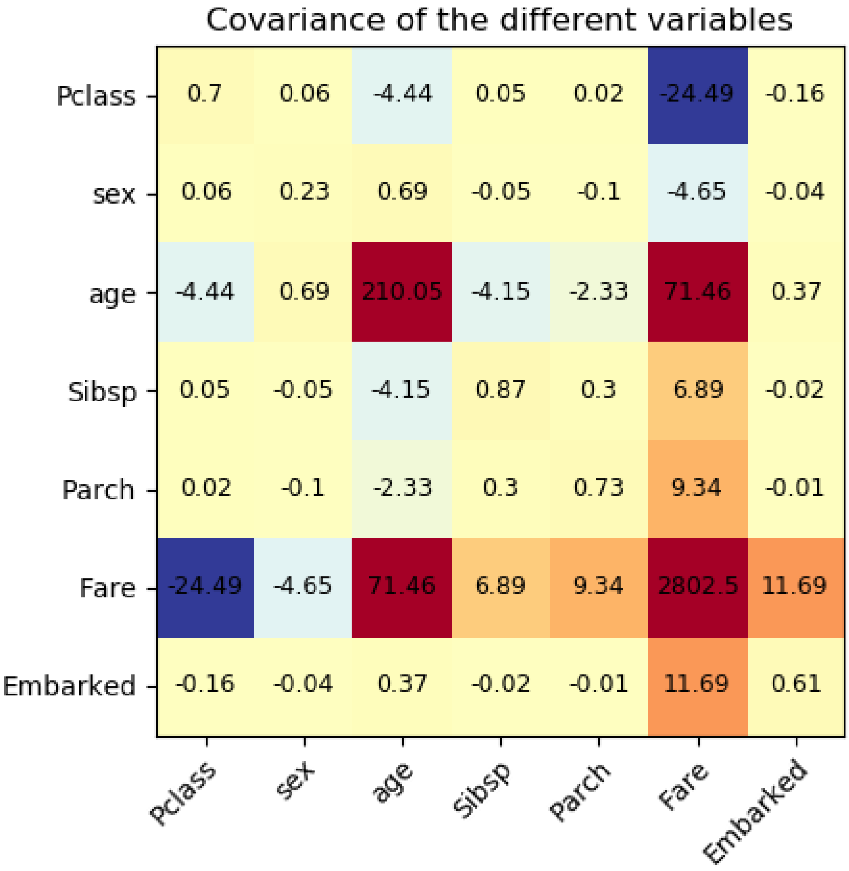

Probability plot of Age

Probability plot of SibSp

Probability plot of Parch

Probability plot of Fare

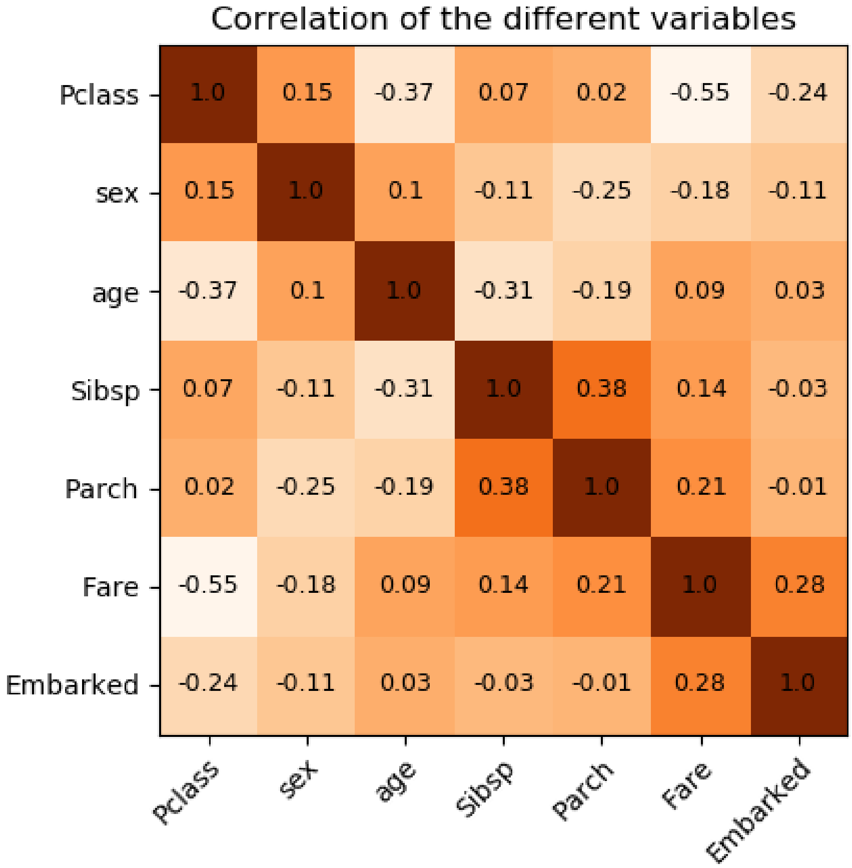
Covariance and correlation of the attributes To see if some of the variables were correlated, we first plotted a covariance matrix, to display the covariance between each of the variables.


Covariance of the different variables

	Pclass	sex	age	Sibsp	Parch	Fare	Embarked
Pclass	0.7	0.06	-4.44	0.05	0.02	-24.49	-0.16
sex	0.06	0.23	0.69	-0.05	-0.1	-4.65	-0.04
age	-4.44	0.69	210.05	-4.15	-2.33	71.46	0.37
Sibsp	0.05	-0.05	-4.15	0.87	0.3	6.89	-0.02
Parch	0.02	-0.1	-2.33	0.3	0.73	9.34	-0.01
Fare	-24.49	-4.65	71.46	6.89	9.34	2802.5	11.69
Embarked	-0.16	-0.04	0.37	-0.02	-0.01	11.69	0.61

This covariance matrix indicates that some of the attributes are linked together. For in-

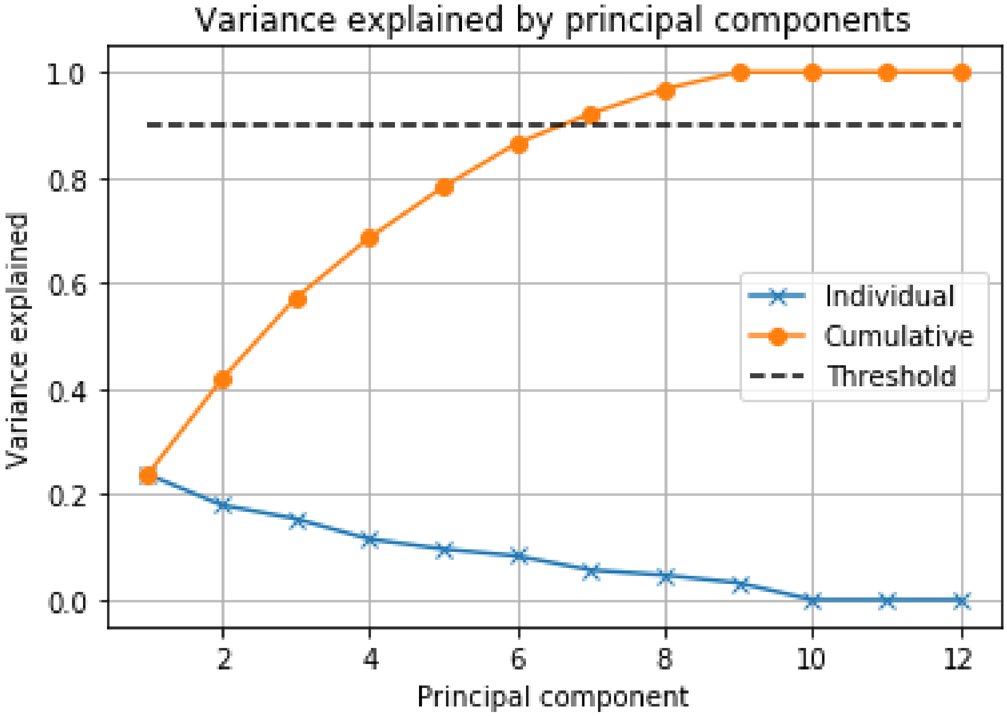
stance, the passenger's class seem to decrease when the fare paid increases (keeping in mind that 1st class is better than 3rd class, hence the variation in this direction). On the contrary, the fare seems to increase with the age of the passenger.
To get a better understanding of how powerful these relationships are, we then calculated the correlation matrix.



Overall, the different attributes are not extremely correlated. The biggest correlation is between the fare paid by a passenger and the class he was in : the coefficient is -0.55. This seems logical : a passenger travelling in 1st class is likely to pay more than a passenger travelling in 3rd class.
The passenger's class is also correlated to age, which can be explained by the fact that the older a passenger is, the more likely he is able to pay for a first class ticket.
The last noticeable correlation is between the number of parents or children and siblings or spouse aboard. This is probably because people tended to travel either alone, or with their family (spouse and children).

Since the attributes are not strongly correlated with each other, every attribute is probably very important to explain the variance. We can the expect the PCA to not reduce in an extreme way the number of dimensions of the data set.

Variation of PCA components As can be seen from the plot below, the first principal component accounts for nearly 25% of variation in the data. The first 7 principal components account for 90% of variation.



Principal Directions of PCA components Vectors:

PC1: [0.43 -0.04 -0.34 0.32 -0.32 0.12 0.04 0.12 0.41 -0.37 0. 0.39]

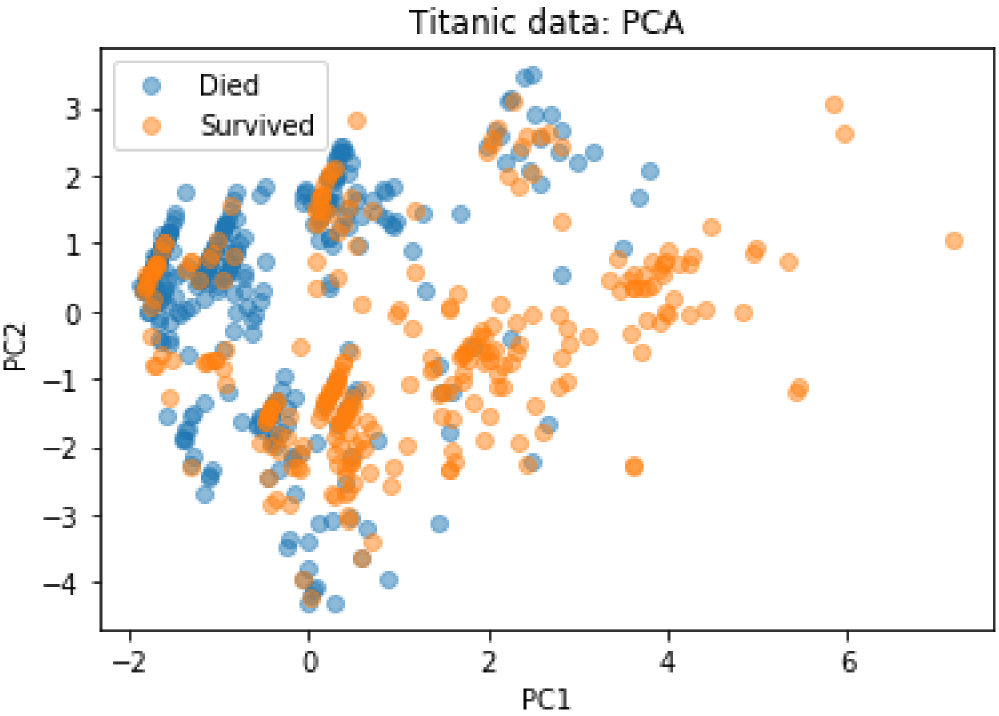
PC2: [0.26 -0.09 -0.15 -0.48 0.48 0.36 -0.33 -0.37 0.05 -0.16 -0.04 0.19]

PC3: [-0.08 -0.5 0.5 -0.11 0.11 -0.25 0.16 0.07 -0.03 -0.43 0.31 0.31]

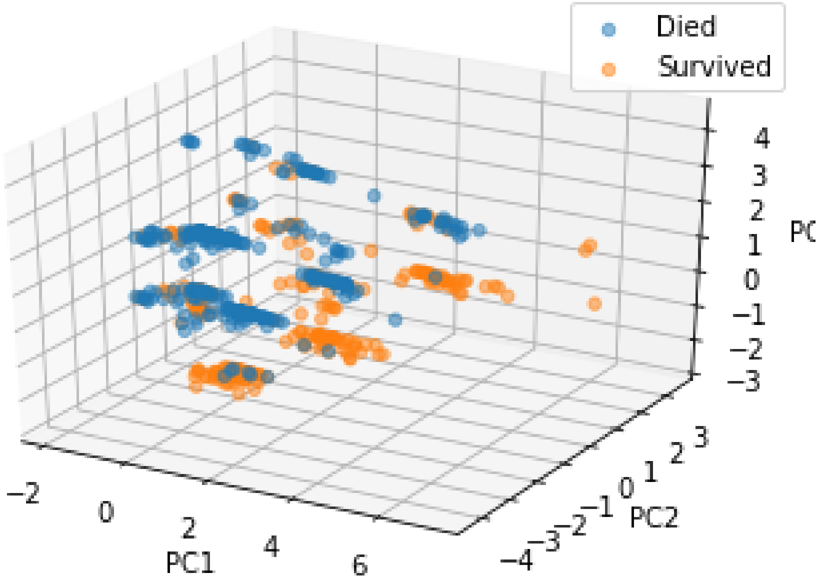
PC1 mainly captures variation of the attributes: Pclass, Sex, and Fare.
PC2 mainly captures variation of the attributes: Sex, Age, SibSp and Parch.
PC3 mainly captures variation of the attributes: Pclass and Embarked.

For an observation to have a large positive projection on the PC1 axis, the passenger would need to have high values for Age, SibSp, Parch, Fare, be a female, on 1st class and embarked in Cherbourg, and low values for the other attributes.

Data projection on principal components It is not possible to represent the first 7 principal components on a graph. But with 2 principal components, we can already have an overview of the data.



To visualize even more of the variation, we can plot the data on the first 3 principal components, since they explain more than 50% of the variation in the data. From these 3 axes, it is possible to see that the passengers are mostly located differently whether they survived or not. This phenomenon would certainly be more obvious if we were able to represent the data according to the 7 main principal components.



4 Discussion

While processing our data, we encountered many anomalies. By looking into these anomalies we gained a better understanding of our data set. After removing erroneous values, we ended up with 712 observations. These should be extensive enough to make for a good training set for our classification algorithm. Since it is hard to reduce the number of dimensions of the data set (we still need 7 principal components to explain the majority of the variance), we can conclude that every attribute will be important in order to predict the survival of a passenger. We have prepared the data and should be set to achieve our primary machine learning aim.

5 Attribution Table

Section	Responsibility
Introduction	Jens (100%)
Description of dataset	Jens (60%) / Maud (40%)
Explanation of the attributes of the data	Jens (100%)
Data visualizations and PCA	Maud (70%) / Jens (30%)
Discussion	Maud (60%) / Jens (40%)

References

[1] Encylopedia Titanica: Titanic People Explorer,
`https://www.encyclopedia-titanica.org/explorer/`

[2] Encylopedia Titanica: Harland & Wolff : Titanic Guarantee Group,
`https://www.encyclopedia-titanica.org/titanic-guarantee-group/`

[3] Encylopedia Titanica: Thomas Drake Martinez Cardeza,
`https://www.encyclopedia-titanica.org/titanic-survivor/thomas-cardeza.html`

[4] Statistical Consultants Ltd: Titanic Fare Data,
`http://www.statisticalconsultants.co.nz/blog/titanic-fare-data.html`

[5] Encyclopedia Britannica: shilling,
`https://www.britannica.com/topic/shilling`

[6] Encyclopedia Britannica: pound sterling,
`https://www.britannica.com/topic/pound-sterling`