To do:

* Moeten we de trainingset splitten in train en validation set?
* Validation methods en beoordeling kwaliteit
* Refs voor verantwoorden welke algorithms en welke validation methods
* Statistics over de initial data
* Statistics over de uitkomsten
* Standaard format checken voor data mining reports

Preprocessing/exploratory data analysis:

* There was no person in the data set for whom more than 2 attributes were not filled in. (Excel)
* For the combined dataset (training and test, survival not included) 1014 of 1306 values for Cabin were empty. (Excel) We added a letter U instead of a deck indication letter, as well as a letter u instead of a even or odd room indication.
* For the combined dataset (training and test, survival not included) 236 of 1306 values for Age were empty. (Excel) We attempt to predict missing values for Age, based on the median for people with comparable features regarding Sex, Class and Prefix.
* For the combined dataset (training and test, survival not included) 1 of 1306 values for Fare was empty. (Excel)
* For the combined dataset (training and test, survival not included) 2 of 1306 values for Embarked were empty. (Excel) For these two passengers, the most frequent location was used; S.
* Paid fares did not exactly correspond with class. (Excel)
* Females have waaaaay bigger chance of survival (0.7+ vs 0.2-) (Jupyter)
* First class has higher chance of survival, females 2nd higher than 3rd (Jupyter)
* Embarked in C higher for both female and male, males Q lower than S (Jupyter)
* Based on Age (**groups: unknown, 0-4, 4-12,** 12…) Females not much difference (except 4-12 = lower) Males up to 12 roughly equal to females, after that sort of tiny normal distribution (Jupyter)
* Ticket fare the more expensive the ticket, the higher the chance of survival (effect bigger fpr males than for females) (Jupyter)
* Divided people into: alone, parch, sibsp and both. Alone lower chance, parch best. (Jupyter)
* Children under 4 did not travel alone, elderly only travelled alone or with spouse, the latter guaranteed survival, the former gave very low chances. All between 18-35 has increased chances when parch compared to alone, sibsp or both. (Jupyter)
* We subtracted the prefixes from the names, and divided the group into: Mr, Master, Mrs, Miss, Royalty and Officers. As expected because of the results for Sex and Age, Miss and Mrs had high chances of survival, also, Master had greater chances than Mr. In accordance with the results for Class, Royalty had relatively good chances too. Furthermore, Officers had low chances, probably because most of them were men. (Jupyter)
* We split Cabin indications into letters (representing the deck), even and odd numbers (representing ship side) and low to high numbers (representing ship front and back).

College dataset:

* Lage dagen en maanden: misschien veel 01/01 ingevuld