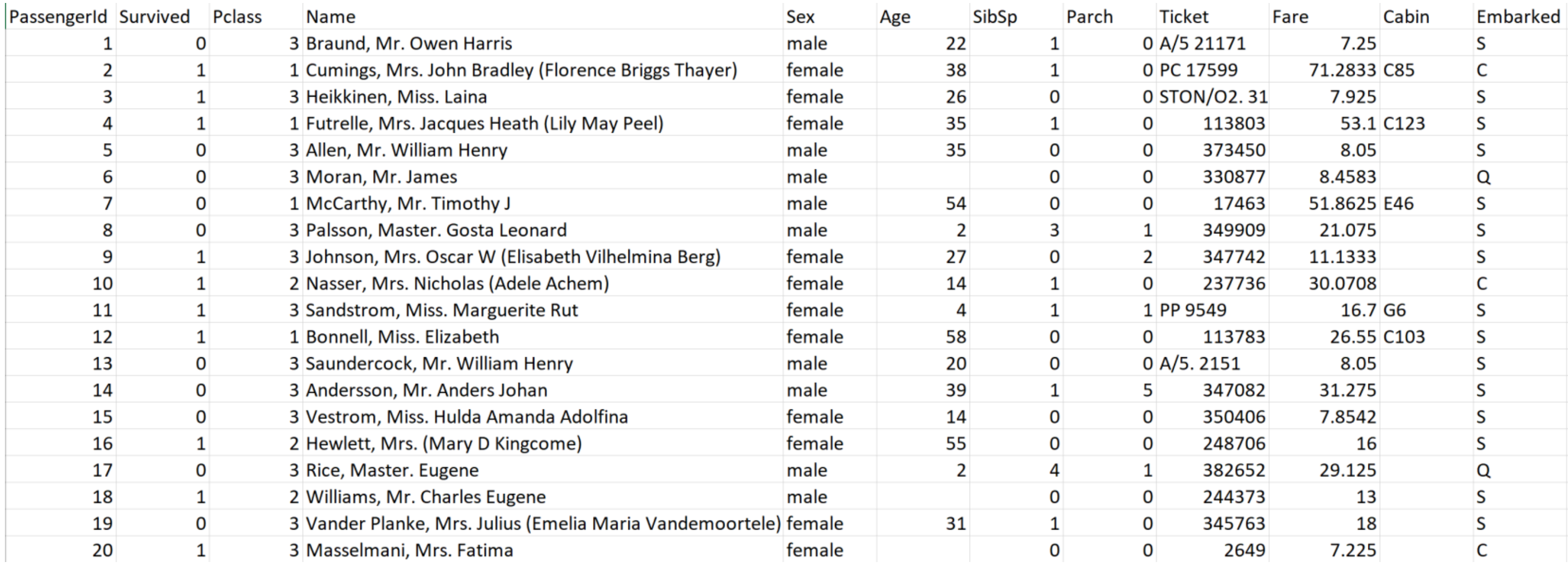
Challenge 2: Goal

- In this challenge we will realize a machine learning system that applies standard techniques to a dataset with missing values and categorical features
- The goal is to learn the basics of Scikit-learn library focusing on:
    - feature engineering for categorical data and missing values
    - training and testing standard classifiers trying to optimize parameters and hyperparameters
    - assessing the performance of different models to find the best one

Challenge 2: Material

- We use a dataset available at:
  - Kaggle repository: https://www.kaggle.com/c/titanic/data
  - the file train.csv is in the Google Drive folder



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 31	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C

Challenge 2: Dataset Overview

- The training set has to be used to build the machine learning models, and also to evaluate the performance using cross validation. In the training set, you find the outcome (i.e. the label, also known as the “ground truth”) for each passenger
- The model can be trained with the given “features”, but you can also use feature engineering to create new features
- There is also a test set, but there are no labels in this case. Thus, if you want to evaluate the performance on the test set, you should submit to the Kaggle competition the outcome for each passenger obtained with the trained model

Challenge 2: Method and evaluation

- Study the features and apply the appropriate transformations for missing data and categorical features
- If you think it is necessary, apply feature transformations, dimensionality reduction and feature selection
- Train and test four classifier models (see on the Scikit-learn user guide which parameters we can optimize):
  - Naive Bayes
  - kNN
  - Decision Tree
  - Logistic Regression
- Which one is the best in terms of accuracy when using a 10-fold cross validation on the training set (i.e. using a fold in each run as test set)?