Table 1: Titanic Dataset Definitions

Variable	**Definition**	**Key**
Survival	Survival	$0 =$ No, $1 =$, Yes
Pclass	Ticket class	$1 = 1$st, $2 = 2$nd, $3 = 3$rd
Name	Passenger name	
Sex	Sex	
Age	Age in years	
Sibsp	# of siblings/spouses aboard	
Parch	# of parents /children aboard	
Ticket	Ticket number	
Fare	Passenger fee	
Cabin	Cabin number	

Embarked	Port Embarked	C = Cherbourg,
		Q = Queenstown,
		S = Southampton

10 Assignment - Titanic Survival

The titanic dataset 'titanic.csv' contains the columns explained in Table 1.

Note that the this assignment allows you to invest different levels of effort and time matching the different degrees of interest in this course by using only a part of the features in the data table or by choosing to construct or not to construct new features. This means feel free to drop features to simplify the task. Further, note that thoroughly working on this assignment will help you in the written exam and I will randomly choose people in class presenting the group assignment work or will collect the results. This assignment is meant to be worked on individually to begin with and finally ends with a group work. Use tomorrow's lecture unit time to fully work on this assignment. Note that you will need to spend more time on this assignment than the four hours tomorrow. If you enjoy working in a team from the beginning on, feel free to do so. The assignment deadline is 27/04/2023.

1. Data Analysis and Preprocessing

a) Load the titanic 'titanic.csv' datasets into pandas dataframes.

b) Identify the label column.

c) Analyse the training dataset statistically and visually with respect to understanding

- which features are important to keep (correlate with survival) and which can we discard? Discard the not meaningful features using the pandas function ".drop".
- which features need preprocessing due to incomplete data? Either complete these columns in a reasonable way or alternatively, appropriately modify, or delete these rows (using the pandas function ".dropna").
- If we can create new features? If this is the case, construct new meaningful features.
- transform the categorical features to numerical inputs (revise what you learnt about encoding categorical features)

For this purpose make use of

- the pandas describe function (use it for numerical as well as categorical values)
- pandas functions "groupby", "mean" to summarize survival probabilities for categorical column data like "Pclass" and "Sex" (first find out what a groupby in combination with mean does) .
- seaborn violinplots (first find out what a violinplot shows) to show distributional data for age.

Furthermore

- look at the age and survival correlation by using histograms binning the age in age bands and use a facetgrid to show the data for age distribution over survival and pclass, use hue to differentiate between the Sex in each subplot of the facetgrid.
- use the pandas method "isna()" to identify missing data values.
- explore alternative methods to visualize the data appropriately (look at seaborn and matplotlib documentation - interested students might look at plotly)
- understand and interpret your statistical summaries and plots to decide which features to keep, which to discard, and which to modify.
- **Tip:** If you decide to not drop the "Name" feature and to extract titles from names you can use the following code snippet

```
df('Title') = df.Name.str.extract(' ((A-Za-z)+)', expand=False)
pd.crosstab(df('Title'), df('Sex'))
```

Afterwards you can use the pandas ”replace” function to summarize rare titles in a column named something like ”Rare”.

- For every challenge you encounter in this assignment seek help looking at the appropriate library documentations.
- If you enjoy this assignment feel free to improve your work by improving the suggested data analysis.

2. Building Machine Learning Models

a) Prepare training and testing appropriately (partitioning the data into training and test set and if required normalizing the data)

b) Recollect, which of the machine learning methods you learnt in class can be applied to predict the survival on the test set.

c) Train and test those machine learning models you identified as suitable. Compare the training and test accuracies for all these models. Which models performs best?

d) Visualize the performance results (training and test accuracies) appropriately (e.g. plotting them as bar-charts).

e) If you enjoy this assignment feel free to improve your work by improving the suggested models further.

3. Presentation Challenge

a) Individually summarize your assignment findings and results in a 10-15min presentation focusing more on the results not your code. Show your own results. If you want to complement those by other explanatory sources, make sure you highlight the original sources. Make sure that your presentation slide deck chooses font sizes that are not too small for the audience. All your figures should include appropriate axis labels, titles, and if reasonable captions. Rehearse to hold the presentation at least one time.

b) Organize yourself in groups of 5 students. Discuss and merge your work/results/presentations to potentially present in class. Individually rehearse the talk at least once.

4. Cheat Sheet

Revise your cheat sheet! This last task is due until the final last lecture unit.