**Applied Machine Learning**

**Assignment 1**

|  |  |
| --- | --- |
| **Name** | YAP LI XEN |
| **Student ID** | P7414389 |
| **Lecturer** | MR. PAUL |

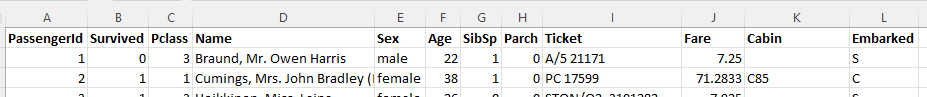
**PART A: CLASSIFICATION**

1. **How is your prediction task defined? And what is the meaning of the output variable?**

* The prediction task is to predict categorical data “Survived”
* The output variables “Survived” consists of 2 values (1=“Survived”, 0=”Not Survived”)
* This is 2-class variables

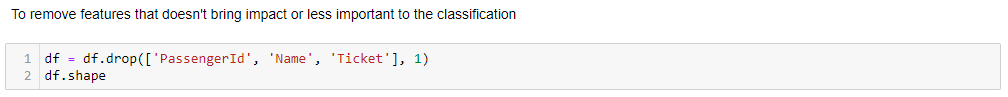
1. **How do you represent your data as features?**

* Target: Survived
* Features: All columns except Survived



1. **Did you process the features in any way?**

* Feature Selection



* Column: Fare

Graphical user interface, text, application

Description automatically generated with medium confidence

* Column: Cabin

Text

Description automatically generated

* Column: Age

Graphical user interface, text, email

Description automatically generated

* Column: Embarked

Graphical user interface, text, application

Description automatically generated

* Encoding

Table

Description automatically generated

1. **Did you bring in any additional sources of data?**

* Column: hasFamilyAboard

Graphical user interface

Description automatically generated with low confidence

1. **How did you select which learning algorithms to use?**

* Referring to scikit-learn algorithm cheat-sheet recommendation

A picture containing diagram

Description automatically generated

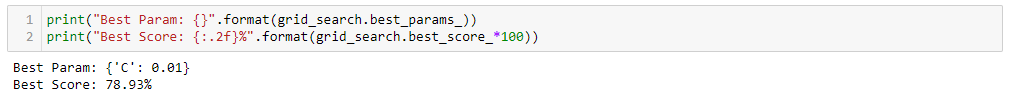
1. **Did you try to tune the hyperparameters of the learning algorithm, and in that case how?**

* Hyperparameter Tuning: GridSearchCV
* To include a range of values in chosen parameter and assign to param grid
* To set cv to 5 (split the train-validation data into 5 folds)

A picture containing text

Description automatically generated

* Then it will generate the best param and best score



1. **How do you evaluate the quality of your system?**

* Score the trained model using both train data and test data, then compare the result
* We can see how well the trained model can predict the test data
* The comparison can also help us to determine if this is under-fitting, appropriate-fitting or over-fitting

Text

Description automatically generated

1. **How well does your system compare to a stupid baseline?**
2. **Can you say anything about the errors that the system makes? For a classification task, you may consider a confusion matrix.**

* Using confusion matrix, it will show us the total number of
  + True Positive
  + False Positive
  + True Negative
  + False Negative

Graphical user interface, text, application, email

Description automatically generated

* Using classification report
* With the true positive (TP) and and false positive (FP), we will be able to calculate the precision
* With the true positive (TP) and and false negative (FN), we will be able to calculate the recall

Graphical user interface, application

Description automatically generated

1. **Is it possible to say something about which features the model considers important? (Whether this is possible depends on the type of classifier you are using)**

* Once all data has transformed into numerical, we are able to determine which features are correlation to the target.

Graphical user interface, text, application

Description automatically generated