

Module 9 Live Coding Assignment: Machine Learning

Throughout this assignment, please remember to import all the necessary libraries at the beginning of every exercise.

In this assignment you'll be working with the "Titanic" dataset. You can find more information about the data and download it from here:

<https://www.kaggle.com/c/titanic/data>

Data Exploration

Question 1

Read the datasets "train.csv" and "test.csv" and save them in two variables named "df_train" and "df_test" respectively. Use the basic functions to get information about the data.

Question 2

Build the following plots using seaborn. Use only the "df_train":

- Bar plot for the "Survived" column
- Bar plot for the "Survived" column grouped by "Sex"
- Bar plots for the "Pclass" column, first for only the variable, and then grouped by "Survived". Use subplots.
- Factorplot for "Pclass" vs. "Survived" grouped by "Sex".
- Violin plot for "Age" vs. "Pclass" and "Age" vs. "Sex" grouped by "Survived".
- Factorplot for "Embarked" vs. "Survived".
- Bar plot for "Embarked" vs. "Pclass"
- Factorplot for "Survived" vs. "Pclass" grouped by "Sex".
- Histograms for "Fare" for each "Pclass". Use subplots.

Question 3

Plot the correlation matrix for the “df_train” dataset.

Data cleaning and feature engineering

Question 4

Divide the range for the “Age” column from 0-80 into 5 bins. Call the new variable “Age_band”.

Question 5

Divide the range for the “Fare” column into 4 bins. Call the new variable “Fare_band”.

Question 6

Transform the “Sex”, “Embarked” and “Initial” columns from string to numeric.

Question 7

Drop unnecessary columns: “Name”, “Age”, “Ticket”, “Fare”, “Cabin”, “PassengerId”. Then, plot the correlation matrix for the final “df_train” dataset.

Modeling

Important note: We will only be using the “df_train” dataset for modeling. You should transform your “df_test” to be able to use it as a separate testing dataset.

Question 8

Separate your data into a train and test datasets using “scikit-learn”, and stratify by “Survived”.

Question 9

Build the following models using “scikit-learn”:

- Linear-SVM
- Logistic Regression
- Decision Tree
- K-Nearest Neighbours(KNN)
- Gaussian Naive Bayes
- Random Forest

Get the accuracy for each model.

Question 10

Run a K-fold cross-validation for each one of the models above. Save the mean accuracy and standard deviation of each CV in a dataframe called "models_df". Then, see your dataframe, and do a boxplot for the accuracy for each model.

Question 11

Plot the confusion matrix for each of your models.