Machine Learning: Lab 3 – Exploratory Analysis of Titanic Dataset

Download the titanic dataset from https://www.kaggle.com/competitions/titanic/data

Prerequisites: Python basics, numpy, pandas, matplotlib, etc.

Importing Data:

1. Create training and testing dataframes from the downloaded csv files. Find

    A) number of rows in training and test sets

    B) display the structure of the dataset along with the datatypes of the fields

Data Cleaning:

1. Analyse the data and identify which columns are not relevant for survivor prediction task. Drop those columns from the dataframes.

2. Check how many columns have missing values in them (NA) and how many have NaN values. Logically impute the dataset.

3. Identify any categorical valued columns (non-numeric) and convert them to numeric.

Exploratory Analysis (On training set):

1. Show how many passengers were male and female and plot using matplotlib. On the same plot depict the people who survived and who died. Make accurate axis and legend. Save the plot in a png file.

2. Show the histogram of the count of passengers who died (according to their age). Age ranges should be <10, 10 to <20, 20 to <30 and so on.

How many minor children died and how many of them survived (<16 years). Create a separate plot for the passengers who survived.

3. Show the distribution on the count of passengers who died (according to the fare they paid). Choose appropriate fare ranges.

Give the percentage of passengers who survived as had paid more than $100. Justify if there was any bias in the rescue operation towards the rich (Yes/No/not enough evidence).

4. Plot graphs showing correlation between different pairs of attributes. Infer if there is any significant correlation between survivors and any specific feature.

5. Find the number of passengers who survived belonging to each Pclass and plot the results.

Classification :

Split the training set into training and validation partitions in the ratio of 75% : 25% and train a linear classification model for survival prediction on the randomly shuffled training partition. Evaluate on the validation set. Make predictions on the test set.

Make a valid submission on Kaggle using your login account.