Machine Learning: Lab 1 – Dataframe manipulation with Pandas

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

Download the CSV files of cast.csv

Dataframe creation and manipulation:

1. Import the data from the cast.csv file. Display the top 10 and bottom 10 rows and the total number of rows and columns.

2. Display the column names and the datatypes of the columns.

3. How many rows are have at least one NA / NULL value? What percentage of the dataset has missing values?

4. Provide a strategy to handle missing values for each column. Implement one such strategy and give a reason for your choice.

5. Display the summary statistics (max, min, mean, median etc.) for the numeric columns, and display top 5 most frequent columns for non-numeric columns.

6. Find how many unique movie titles are present in the dataframe.

7. Find how many unique movies are there which have release years between 2000 to 2010.

8. How many character are there in the “Star Wars” movie series, and which of the character(s) have the highest rating.

9. Split the dataset into following partitions. Note that the partitions should not have any missing values and all the columns should have a common datatype.

    Training set: All movies with release year <=2010

    Test Set: All movies with release year > 2010

10. Identify which is a categorical variable/column. Give the count of the number of actors and actresses in the training set.

11. Do a random split of the training set into 80% training and rest 20% validation set (approximately) based on the unique movies, i.e., 80% of the movies in the with be randomly separated to train the models.

Plotting (Use Maplotlib / Seaborn):

1. Find the number of movies released each year and plot a line graph showing the trend of movie releases over time.

2. Create a bar plot which shows the number of characters for a particular rating in the training set. X axis has the ratings and Y- axis has the count of characters.

3. Draw a time series line plot, with years on the x-axis, and the average rating of the actors/actresses on the y-axis. This will be to see if there has been a pattern in the “likeness” or “dislikeness” of movies over the years.