Suggested list of assignments for AI&ML and Data Science Honors Course

(Honours* in Artificial Intelligence and Machine Learning
Third Year of Engineering (Semester V) 310302:
Computational Programming Laboratory

1. Compute Estimators of the main statistical measures like Mean, Variance, Standard Deviation, Covariance, Correlation and Standard error with respect to any example. Display graphically the distribution of samples.
2. Plot the Normal Distribution for class test result of a particular subject. Identify the Skewness and Kurtosis
3. Load the dataset: birthwt Risk Factors Associated with Low Infant Birth Weight at https://raw.github.com/neurospin/pystatsml/master/datasets/birthwt.csv 1. Test the association of mother's (bwt) age and birth weight using the correlation test and linear regeression. 2. Test the association of mother's weight (lwt) and birth weight using the correlation testand linear regeression. 3. Produce two scatter plot of: (i) age by birth weight; (ii) mother's weight by birth weight. Elaborate the Conclusion
4. Apply Basic PCA on the iris dataset. The data set is available at: https://raw.github.com/neurospin/pystatsml/master/datasets/iris.csv • Describe the data set. Should the dataset been standardized? • Describe the structure of correlations among variables. • Compute a PCA with the maximum number of components .• Compute the cumulative explained variance ratio. Determine the number of components$K$by your computed values. • Print the$K$principal components directions and correlations of the$K$principal components with the original variables. Interpret the contribution of the original variables into the PC. • Plot the samples projected into the$K$first PCs. • Color samples by their species.
5. Perform clustering of the iris dataset based on all variables using Gaussian mixture models. Use PCA to visualize clusters.

Honors* in Data Science, Artificial Intelligence and Machine Learning Fourth year of Engineering (Semester VII)
Machine learning Laboratory

1. Creating & Visualizing Neural Network for the given data. (Use python) Note: download dataset using Kaggal. Keras, ANN visualizer, graph viz libraries are required.
2. Recognize optical character using ANN
3. Implement basic logic gates using Hebbnet neural networks
4. Exploratory analysis on Twitter text data Perform text pre-processing, Apply Zips and heaps law, Identify topics
5. Text classification for Sentimental analysis using KNN Note: Use twitter data
6. Write a program to recognize a document is positive or negative based on polarity words using suitable classification method

Honours* in Data Science Third year of Engineering (Semester V) 310502:
Data Science and Visualization Laboratory

1. Access an open source dataset "Titanic". Apply pre-processing techniques on the raw dataset.
2. Build training and testing dataset of assignment 1 to predict the probability of a survival of a person based on gender, age and passenger-class
3. Download Abalone dataset. (URL: http://archive.ics.uci.edu/ml/datasets/Abalone) Data set has total 8 Number of Attributes. Sex nominal M, F, and I (infant) Length continuous mm Longest shell measurement Diameter continuous mm perpendicular to length Height continuous mm with meat in shell Whole weight continuous grams whole abalone Shucked weight continuous grams weight of meat Viscera weight continuous grams gut weight (after bleeding) Shell weight continuous grams after being dried Rings (age/class of abalone)

   Load the data from data file and split it into training and test datasets. Summarize the properties in the training dataset. The number of rings is the value to predict: either as a continuous value or as a classification problem. Predict the age of abalone from physical measurements using linear regression or predict ring class as classification problem
4. Use Netflix Movies and TV Shows dataset from Kaggle and perform following operation :
   1. Make a visualization showing the total number of movies watched by children
   2. Make a visualization showing the total number of standup comedies
   3. Make a visualization showing most watched shows.
   4. Make a visualization showing highest rated show Make a dashboard (DASHBOARD A) containing all of these above visualizations.

TE Computer Data Science Assignments-
DS-I
Perform the following operations using Python on given dataset
1. Import all the required Python Libraries.
2. Load the Dataset into pandas dataframe.
3. Data Preprocessing: check for missing values in the data using pandas isnull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.
4. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.
5. Turn categorical variables into quantitative variables in Python.

DS-II
Create an "Academic performance" dataset of students and perform the following operations using Python.
1. Scan all variables for missing values and inconsistencies. If there are missing values and/or Inconsistencies, use any of the suitable techniques to deal with them.
2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them.

3. Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness and convert the distribution into a normal distribution.

DS-III

Descriptive Statistics - Measures of Central Tendency and variability

Perform the following operations on any open source dataset (e.g., data.csv)

1. Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc.) with numeric variables grouped by one of the qualitative (categorical) variable. For example, if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categorical variable.
2. Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa', 'Iris-versicolor' and 'Iris-versicolor' of iris.csv dataset.

DS-IV

Data Analytics I

Create a Linear Regression Model using Python/R to predict home prices using Boston Housing Dataset,The Boston Housing dataset contains information about various houses in Boston through different parameters. There are 506 samples and 14 feature variables in this dataset. The objective is to predict the value of prices of the house using the given features.

DS-V

Data Analytics II

1. Implement logistic regression using Python/R to perform classification on Social_Network_Ads.csv dataset.
2. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.

DS-VI

Data Analytics-III

1. Implement Simple Naïve Bayes classification algorithm using Python/R on iris.csv dataset.
2. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.

DS-VII

Data Visualization I

1. Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship. Use the Seaborn library to see if we can find any patterns in the data.
2. Write a code to check how the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram.

DS-VIII

Data Visualization II

1. Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column names : 'sex' and 'age')
2. Write observations on the inference from the above statistics.

DS-IX

Data Visualization III

Use the Iris flower dataset or any other dataset into a DataFrame. Scan the dataset and give the inference as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.
2. Create a histogram for each feature in the dataset to illustrate the feature distributions.
3. Create a boxplot for each feature in the dataset.
4. Compare distributions and identify outliers.