Programing Assignment 1

(Programing)

Please paste code, produced tables and plots on your solution.

1. NumPy is a package which provides convenient matrix/vector computations: (10%)
   a. Please generate a 8×8 matrix A and find the minimum, mean, maximum values of each row and column using NumPy. (3%)
   b. Please generate another 8×8 matrix B and find the transpose and inverse of B. (3%)
   c. Please compute the element-wise multiplication and matrix multiplication of A and B. (4%)

2. Pandas DataFrame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns). A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas DataFrame consists of three principal components, the data, rows, and columns. (10%)
   a. Given a table of NBA players' stats as follows, please generate a Pandas DataFrame based on the table. (3%)

Player	GP	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%
James Harden	11	38.5	31.6	9.9	24	41.3	4.4	12.5	35
Kawhi Leonard	24	39.1	30.5	10.1	20.7	49	2.3	6	37.9
Paul George	5	40.8	28.6	8.8	20.2	43.6	3	9.4	31.9
Stephen Curry	22	38.5	28.2	8.6	19.6	44.1	4.2	11.1	37.7
Damian Lillard	16	40.6	26.9	8.6	20.6	41.8		9.9	37.3
Giannis Antetokounmpo	15	34.3	25.5	8.6	17.4	49.4	1.2	3.7	32.7
Nikola Jokic	14	39.7	25.1	9.4	18.6	50.6	1.6	4	39.3
CJ McCollum	16	39.7	24.7		21.9	44	2.9	7.3	39.3
Russell Westbrook	5	39.4	22.8	8	22.2	36	2.2	6.8	32.4
DeMar DeRozan	7	35.9	22	8.3	17	48.7	0	0.1	0
James Harden	11	38.5	31.6	9.9	24	41.3	4.4	12.5	35

   b. Please check how many data are missing and fill the missing data with the average of other players. (4%)
   c. Now, we get the stats of another player as follows, please add his information into our DataFrame. (3%)

Player	GP	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%
Lou Williams	6	29.3	21.7	7.5	17.3	43.3	1	3	33.3

3. Parkinson Dataset with replicated acoustic features Data Set (*http://archive.ics.uci.edu/ml/datasets/Parkinson+Dataset+with+replicated+acoustic+features+*) contains acoustic features extracted from 3 voice recording replications of the sustained /a/

phonation for each one of the 80 subjects (Some of them with Parkinson's Disease, i.e., status=1). Please find the data as Parkinson.csv file. (Hint: columns 'ID' and 'Recording' can not be considered as the features.) (40%)

a. As we discussed in class, given a dataset to analyze, before designing supervised learning model or unsupervised model, we need to understand the structure and statistics of the data, i.e., distribution of class labels, distribution of each feature, etc. Please implement such data analysis using Python. (10%)
b. Considering each record as an individual sample, please train a decision tree classifier (max_depth = 3) to predict the status of each sample. Please plot your decision tree. (15%)
c. As discussed in class, Grid Search can help us to tune the model parameters to find the optimal solution. Please tune your decision tree classifier to improve the predictive performance. (15%)

4. Indian Liver Patient Dataset (*https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29*, please find the data as the ILPD.csv file.) provides the age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos of patients. Please train a KNN classifier and a Logistic Regression classifier to predict class label of the patient. (for KNN classifier please refer to: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html) (40%)

Note: some data are missing.