Question 1)

In machine learning taking the large inputs can cause the performance. For overcoming the poor performance we will use dimensionality reduction for reducing the number of inputs.

Question 2)

Given data set is : df = pd.read\_csv(‘file.csv’)

* To check whether the given column has duplicate values:

df.duplicated()

* This returns the duplicate records.

df.duplicated([‘column name’])

* If any column of the dataset has the duplicated values in that column then it returns True from that row, if not then it returns false.
* To remove those duplicate values from the dataset:

df.drop\_duplicates()

* To remove the duplicates from a particular column:

df.drop\_duplicates([‘column\_name’])

* This sorting happens using the first value in the dataset of the column. To use some other value of the column for sorting we must use ‘keep’. If we are using ‘keep=last’ so that sorting happens using last column.

df.drop\_duplicates([‘column\_name’],keep=’first’)

or

df.drop\_duplicates([‘column\_name’],keep=’last’)