1. Explain dimension reduction in machine learning.

Dimension Reduction is the process of reducing the number of random variables under considering, by obtaining a set of principal variables. They are of two types one is feature selection, feature extraction.

1. How can you handle duplicate values in a dataset for a variable in Python using pandas?

df = pd.read\_csv('file.csv')

pandas have function called duplicated to tell that columns are having duplicated values or not

df.duplicated() (for checking the dataset containing duplicate columns)

df.duplicated([‘column name’]) (return the duplicated records)

df.drop\_duplicates() (removes the duplicated columns )

df.drop\_duplicates(subset="column name", keep="last") (sorting is done using the first value in the dataset of the column, keep function is being used)

df.drop\_duplicates(subset="column name", keep="last") (sorting is done using the last value in the dataset of the column)

df.duplicated(subset=["column name", "column name"]) (function to the dataframe directly and specify which column we want to check for duplicates with subset argument)