Assignment- VI

(Based on Data Preprocessing)

Q1: Consider the following dataset:

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	78	false	yes
rain	70	96	false	yes
rain	68	80	false	yes
rain	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rain	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rain	71	80	true	no

(a) Encode the Temperature and Humidity as binary column wit values True/False using a threshold being decided by Gain Ratio.

(b) Train a C4.5 Decision Tree Classifier on the transformed dataset.

(c) Using Gain Ratio, find the best two attributes that decide the Play attribute.

(d) Using χ^2 test, check whether Play label depends upon Outlook feature or not at 95% confidence level.

Q2: Refer to Q4 of Assignment II, find word embeddings of the PPMI matrix of dimensionality 500 using SVD decomposition (using step-by-step implementation)

Q3: Implement Linear Discriminant Analysis (LDA) step-by-step on Iris dataset (present in sklearn.datasets).

Q4: Download the dataset regarding Car Price Prediction from the following link:
https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data

1. Load the dataset with following column names ["symboling", "normalized_losses", "make", "fuel_type", "aspiration","num_doors", "body_style", "drive_wheels", "engine_location", "wheel_base", "length", "width", "height", "curb_weight",

"engine_type", "num_cylinders", "engine_size", "fuel_system", "bore", "stroke", "compression_ratio", "horsepower", "peak_rpm", "city_mpg", "highway_mpg", "price"] and replace all ? values with NaN

2. Replace all NaN values with central tendency imputation. Drop the rows with NaN values in price column
3. Using isolation forest technique (present in sklearn.ensemble) identify outliers present in length column.
4. There are 10 columns in the dataset with non-numeric values. Convert these values to numeric values using following scheme:
   (i) For “num_doors” and “num_cylinders”: convert words (number names) to figures for e.g., two to 2
   (ii) For "body_style", "drive_wheels": use dummy encoding scheme
   (iii) For “make”, “aspiration”, “engine_location”,fuel_type: use label encoding scheme
   (iv) For fuel_system: replace values containing string *pfi* to 1 else all values to 0.
   (v) For engine_type: replace values containing string *ohc* to 1 else all values to 0.
5. Divide the dataset into input features (all columns except price) and output variable (price). Scale all input features.