Experiment:1

Title: Data Pre-processing

Aim: To Collect, Clean, Integrate and Transform Healthcare Data based on specific disease < specify the name of the dataset here> using python.

Theory:

<Explanation about Data Pre processing tasks with example. >

Procedure:

Perform the following Pre -processing tasks on the chosen dataset

1. Identify duplicate rows
2. Find the missing values and remove/replace
3. Replace the missing values of nominal attribute (if any) to NULL
4. Identify columns with very few values
5. Rename the nominal values of nominal attribute
6. Identify columns that contain single value
7. Perform transformation as per data attributes

Output:

<Provide the snapshots for the above pre-processing steps>

Conclusion:

References:

https://www.javatpoint.com/data-preprocessing-machine-learning
https://serokell.io/blog/data-preprocessing
https://vitalflux.com/data-preprocessing-steps-in-machine-learning/
https://archive.ics.uci.edu/datasets

Sample pre-processing tasks solved using python

1. Load any dataset , identify the irrelevant columns for specific analysis and drop them.

   Code:

   ```
   import pandas as pd
   df = pd.read_csv('hepatitis.csv')
   irrelevant_cols = ['histology',]
   df = df.drop(columns=irrelevant_cols)df.head(10)
   ```

2. Find the rows with missing values and replace with specific string or value

   Code:

   ```
   df = df.fillna(value=-1)
   df.head(10)
   ```

3. Add any 2 dummy attributes in your dataset

   Code:

   ```
   import numpy as np
   df['dummy1'] = np.random.rand(len(df))
   df['dummy2'] = np.random.randint(0, 2, size=len(df))
   df.head(10)
   ```

4. Apply any 2 standard normalization techniques on numeric attributes of the sampledataset chosen.

   Code:

   ```
   from sklearn.preprocessing import StandardScaler, MinMaxScaler
   numeric_cols = ['alk_phosphate', 'sgot']
   scaler1 = StandardScaler()
   df[numeric_cols] = scaler1.fit_transform(df[numeric_cols])
   print(df.head())
   ```