**PRACTICAL 3**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name:** | Harsh Shah | **Semester:** | VII | **Division:** | 6 |
| **Roll No.:** | 21BCP359 | **Date:** | 06-08-24 | **Batch:** | G11 |
| **Aim:** | Understanding Pre-Processing in Datasets. | | | | |

**Question 1**

**Dataset:** diabetes.csv

import numpy as np

import pandas as pd

from sklearn.preprocessing import MinMaxScaler, Binarizer, StandardScaler

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

***# Dataset without label/class***

df1 = df.drop(['Outcome'], *axis*=1)

***# Scaling***

min\_max\_scaler = MinMaxScaler(*feature\_range*=(0,1))

scaled\_features = min\_max\_scaler.fit\_transform(df1)

scaled\_df = pd.DataFrame(scaled\_features, *columns*=df1.columns)



Figure 1: Scaled df

***# Binarization***

binarizer = Binarizer(*threshold*=0.0)

binarized\_data = binarizer.fit\_transform(scaled\_df)

binarized\_df = pd.DataFrame(binarized\_data, *columns*=scaled\_df.columns)

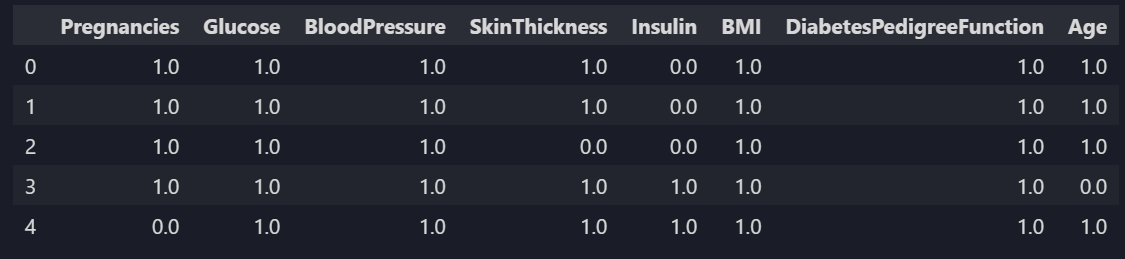


Figure 2: Binarized df.head()

***# Standardization***

scaler = StandardScaler()

standardized\_data = scaler.fit\_transform(binarized\_df)

standardized\_df = pd.DataFrame(standardized\_data, *columns*=binarized\_df.columns)



Figure 3: Standardized df.head()

**Question 2**

**Dataset:** spam.csv

import re

import nltk

import pandas as pd

from nltk.corpus import stopwords

nltk.download("stopwords")

df = pd.read\_csv("spam.csv", *encoding*="latin-1")

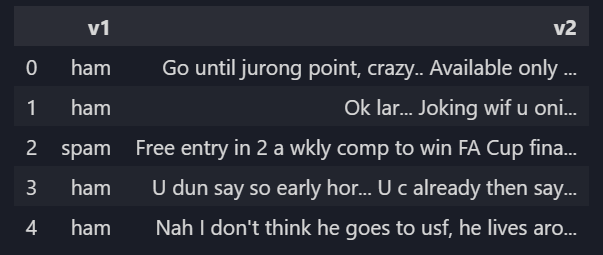


Figure 4: df.head()

**# *Remove Puntuation and Stopwords***

*def* remove\_punctuations(*text*):

    return re.sub(*r*"[^\w\s]", "", text)

*def* remove\_stopwords(*text*):

    stop\_words = *set*(stopwords.words("english"))

    return " ".join([word for word in text.split() if word.lower() not in stop\_words])

df["v2"] = df["v2"].apply(remove\_punctuations)

df["v2"] = df["v2"].apply(remove\_stopwords)

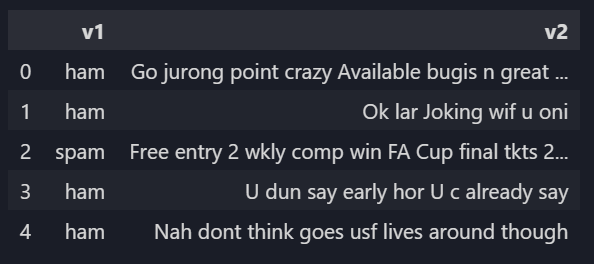


Figure 5: df.head()