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

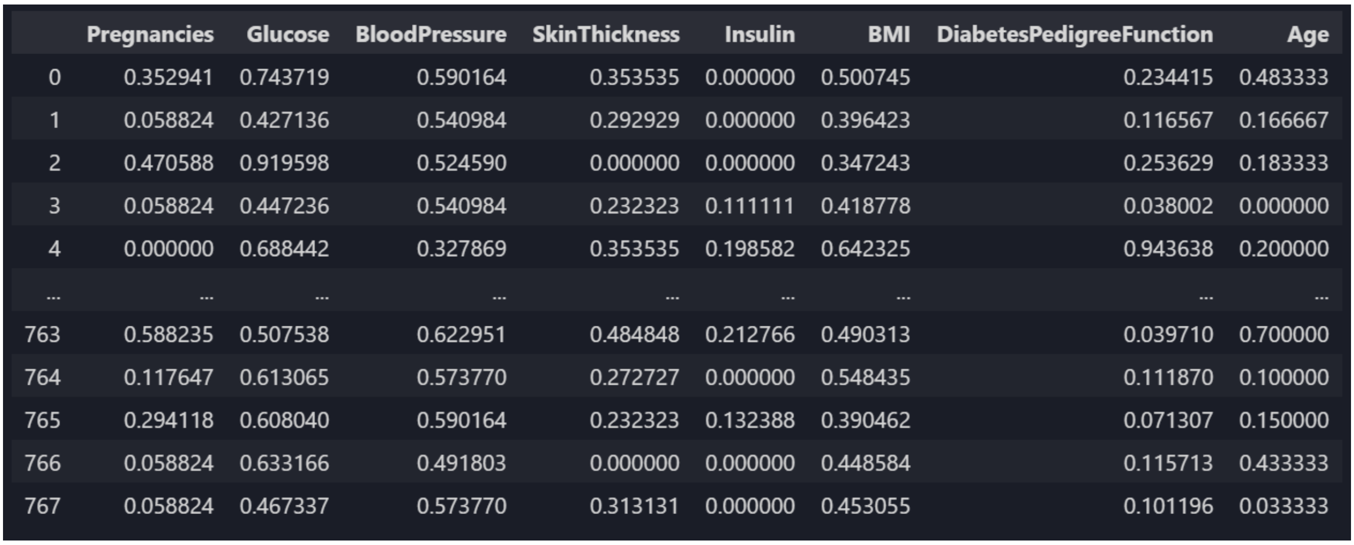
df1 = df.drop(['Outcome'], *axis*=1)

# Scaling

min_max_scaler = MinMaxScaler(*feature_range*=(0,1))

scaled_features = min_max_scaler.fit_transform(df1)

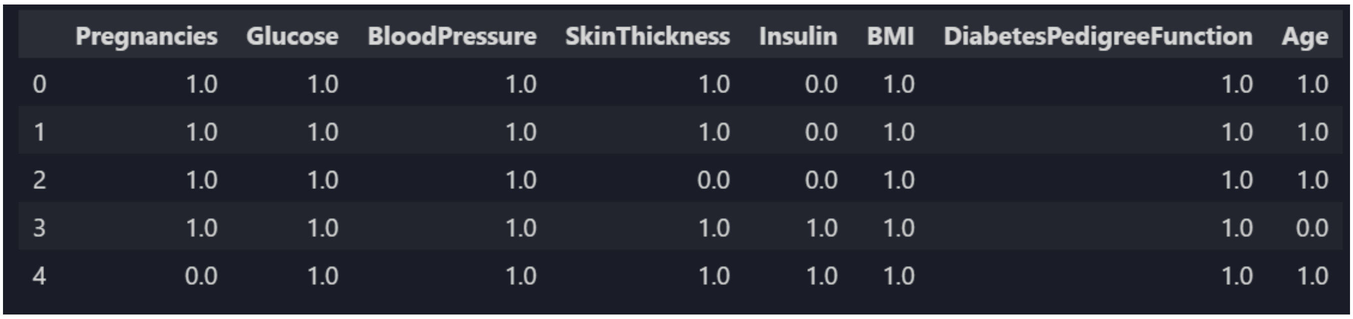
scaled_df = pd.DataFrame(scaled_features, *columns*=df1.columns)



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.352941	0.743719	0.590164	0.353535	0.000000	0.500745	0.234415	0.483333
1	0.058824	0.427136	0.540984	0.292929	0.000000	0.396423	0.116567	0.166667
2	0.470588	0.919598	0.524590	0.000000	0.000000	0.347243	0.253629	0.183333
3	0.058824	0.447236	0.540984	0.232323	0.111111	0.418778	0.038002	0.000000
4	0.000000	0.688442	0.327869	0.353535	0.198582	0.642325	0.943638	0.200000
...	...	...	...	...	...	...	...	...
763	0.588235	0.507538	0.622951	0.484848	0.212766	0.490313	0.039710	0.700000
764	0.117647	0.613065	0.573770	0.272727	0.000000	0.548435	0.111870	0.100000
765	0.294118	0.608040	0.590164	0.232323	0.132388	0.390462	0.071307	0.150000
766	0.058824	0.633166	0.491803	0.000000	0.000000	0.448584	0.115713	0.433333
767	0.058824	0.467337	0.573770	0.313131	0.000000	0.453055	0.101196	0.033333

Figure 1: Scaled df

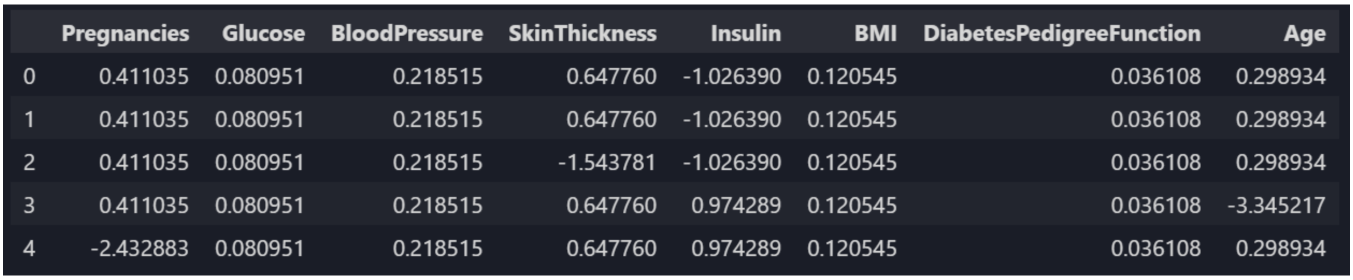
```
# Binarization
binarizer = Binarizer(threshold=0.0)
binarized_data = binarizer.fit_transform(scaled_df)
binarized_df = pd.DataFrame(binarized_data, columns=scaled_df.columns)
```



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0
1	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0
2	1.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0
4	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

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)
```



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.411035	0.080951	0.218515	0.647760	-1.026390	0.120545	0.036108	0.298934
1	0.411035	0.080951	0.218515	0.647760	-1.026390	0.120545	0.036108	0.298934
2	0.411035	0.080951	0.218515	-1.543781	-1.026390	0.120545	0.036108	0.298934
3	0.411035	0.080951	0.218515	0.647760	0.974289	0.120545	0.036108	-3.345217
4	-2.432883	0.080951	0.218515	0.647760	0.974289	0.120545	0.036108	0.298934

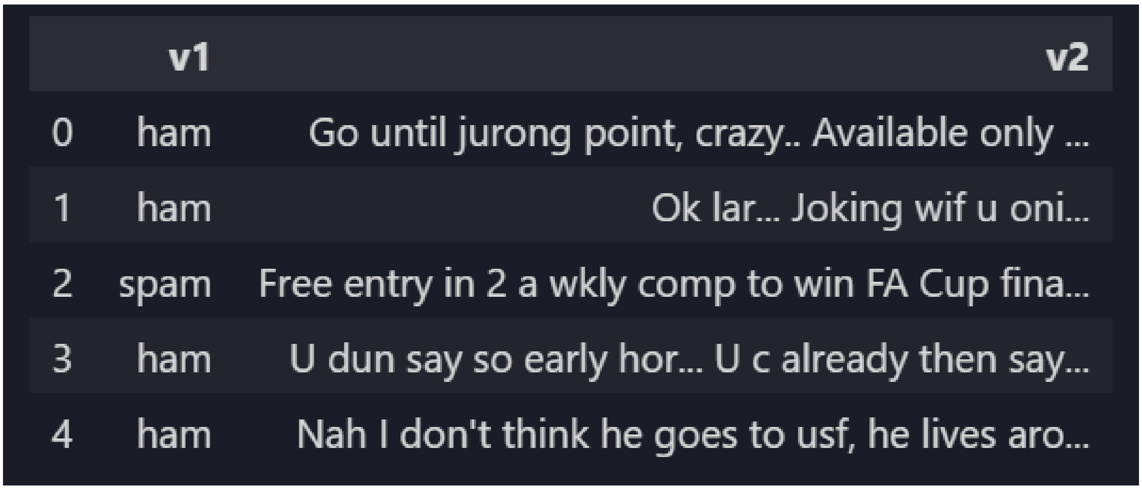
Figure 3: Standardized df.head()

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")
```



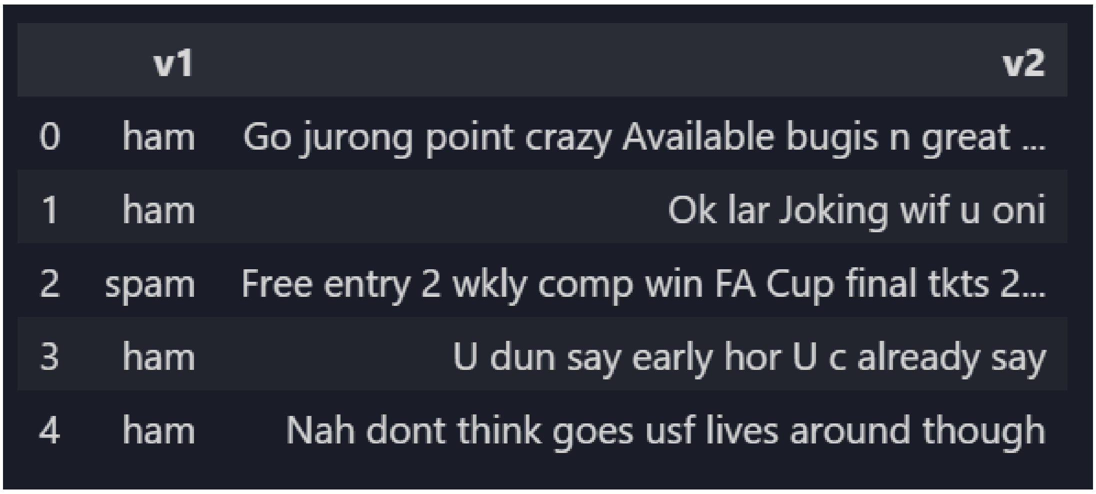
	v1	v2
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

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)
```



	v1	v2
0	ham	Go jurong point crazy Available bugis n great ...
1	ham	Ok lar Joking wif u oni
2	spam	Free entry 2 wkly comp win FA Cup final tkts 2...
3	ham	U dun say early hor U c already say
4	ham	Nah dont think goes usf lives around though

Figure 5: df.head()