**Exercise 1b - Choose and Prepare Same Dataset for SVM and NN(MLB) Assignments**

1. Choose one of these 3 data-sets. The following Python code loads each data set and provides details on it:

from sklearn.datasets import fetch\_openml

#load credit-g dataset from OpenML

#https://www.openml.org/d/31

#German Credit dataset

#This dataset classifies people described by a set of attributes as good or bad credit risks.

creditg = fetch\_openml(name='credit-g', as\_frame=True)

print(creditg.frame.columns)

print(creditg.frame.describe())

print(creditg.details)

#load diabetes dataset from OpenML

#https://www.openml.org/d/37

#Pima Indians Diabetes Database

diabetes = fetch\_openml(name='diabetes', as\_frame=True)

print(diabetes.frame.columns)

print(diabetes.frame.describe())

print(diabetes.details)

#load spambase dataset from OpenML

#https://www.openml.org/d/44

#SPAM E-mail Database

spambase = fetch\_openml(name='spambase', as\_frame=True)

print(spambase.frame.columns)

print(spambase.frame.describe())

print(spambase.details)

Describe your chosen dataset.

1. Read the following article on encoding categorical features and provide a summary: <https://towardsdatascience.com/guide-to-encoding-categorical-features-using-scikit-learn-for-machine-learning-5048997a5c79>
2. For any categorical input variables, use one-hot encoding:  Read and find the examples in <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html>
3. For categorical output variables as strings use LabelEncoder. Read and find the examples in:<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>

**Grading**

|  |  |  |
| --- | --- | --- |
| **Question** | **Maximum Points** | **Student Score** |
| 1 Choose and describe dataset | 40 |  |
| 2 Read and summarize article on encoding features | 20 |  |
| 3 I need be use one-hot encoding for categorical input variables in your dataset; show your code and summarize your results. Use LabelEncoder for categorical output variables; show your code and summarize your results | 40 |  |
| **Total:** | 100 |  |