**Expt 2: Aim: To split the dataset into train, validation, and test datasets.**

1. **Iris dataset**
2. **digits dataset**
3. **diabetes dataset**

**2(a) Program to split the iris dataset into train, validation, and test datasets.**

**In this program, 30 % of the total samples are reserved for testing. Now 70% of data samples are divided in the ratio of 70:30 as training set and validation set.**

**import pandas as pd**

**from sklearn import datasets**

**from sklearn.model\_selection import train\_test\_split**

**# load the iris dataset**

**iris = datasets.load\_iris()**

**df = pd.DataFrame(iris.data)**

**label = pd.DataFrame(iris.target)**

**# Let us split 70% for Training and 30% for Testing**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(df, label,test\_size=0.3,shuffle = True)**

**X\_train, X\_val, y\_train, y\_val = train\_test\_split(X\_train, y\_train, test\_size=0.3,shuffle= True)**

**print(" Totoal number of samples in the Dataset  ", df.shape[0])**

**print(" Total number of labels in the Dataset ", label.shape[0])**

**print(" Total number of samples in training set X\_train ", X\_train.shape[0])**

**print(" Total number of labels in training set y\_train " , y\_train.shape[0])**

**print(" Total number of samples in testing set X\_test ", X\_test.shape[0])**

**print(" Total number of labels in testing set y\_test " , y\_test.shape[0])**

**print(" Total number of samples in validation set X\_test ", X\_val.shape[0])**

**print(" Total number of labels in validation set  y\_test " , y\_val.shape[0])**

**print('The 3 samples of training set')**

**print(X\_train.head(3))**

**print('The random 3 samples of training set')**

**print(X\_train.sample(3))**

**print('The 3 samples of testing set')**

**print(X\_test.head(3))**

**print('The random 3 samples of testing set')**

**print(X\_test.sample(3))**

**print('The 3 samples of validation set')**

**print(X\_val.head(3))**

**print('The random 3 samples of validation set')**

**print(X\_val.sample(3))**

**OUTPUT**

**Total number of samples in the Dataset 150**

**Total number of labels in the Dataset 150**

**Total number of samples in training set X\_train 73**

**Total number of labels in training set y\_train 73**

**Total number of samples in testing set X\_test 45**

**Total number of labels in testing set y\_test 45**

**Total number of samples in validation set X\_test 32**

**Total number of labels in validation set y\_test 32**

**The 3 samples of training set**

**0 1 2 3**

**100 6.3 3.3 6.0 2.5**

**65 6.7 3.1 4.4 1.4**

**86 6.7 3.1 4.7 1.5**

**The random 3 samples of training set**

**0 1 2 3**

**29 4.7 3.2 1.6 0.2**

**63 6.1 2.9 4.7 1.4**

**67 5.8 2.7 4.1 1.0**

**The 3 samples of testing set**

**0 1 2 3**

**61 5.9 3.0 4.2 1.5**

**53 5.5 2.3 4.0 1.3**

**55 5.7 2.8 4.5 1.3**

**The random 3 samples of testing set**

**0 1 2 3**

**53 5.5 2.3 4.0 1.3**

**42 4.4 3.2 1.3 0.2**

**55 5.7 2.8 4.5 1.3**

**The 3 samples of validation set**

**0 1 2 3**

**82 5.8 2.7 3.9 1.2**

**20 5.4 3.4 1.7 0.2**

**120 6.9 3.2 5.7 2.3**

**The random 3 samples of validation set**

**0 1 2 3**

**108 6.7 2.5 5.8 1.8**

**20 5.4 3.4 1.7 0.2**

**81 5.5 2.4 3.7 1.0**

**Assignment**

1. **Obtain the results for train = 65% validation = 15% test =20% and justify.**
2. **Obtain the results for train = 75% validation = 15% test =15% and justify.**
3. **Obtain the results for train = 85% validation = 5% test =10% and justify.**

**2(b) Program to split the digits dataset into train, validation, and test datasets.**

**In this program, 20 % of the total samples are reserved for testing. Now 80% of data samples are divided in the ratio of 80:20 as training set and validation set.**

**#Program 2b**

**import pandas as pd**

**from sklearn import datasets**

**from sklearn.model\_selection import train\_test\_split**

**# load the digits dataset**

**digit = datasets.load\_digits()**

**df = pd.DataFrame(digit.data)**

**label = pd.DataFrame(digit.target)**

**#  Let us split 80% for Training and 20% for Testing**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(df, label,test\_size=0.2, shuffle = True)**

**# Of the 80% Traing Set , Now 20% is reserved for validation set and remaining 80% is reserverd for Training**

**X\_train, X\_val, y\_train, y\_val = train\_test\_split(X\_train, y\_train, test\_size=0.2, shuffle = True)**

**print(" Total number of samples in the Dataset  ", df.shape[0])**

**print(" Total number of labels in the Dataset ", label.shape[0])**

**print(" Total number of samples in training set X\_train ", X\_train.shape[0])**

**print(" Total number of labels in training set y\_train " , y\_train.shape[0])**

**print(" Total number of samples in testing set X\_test ", X\_test.shape[0])**

**print(" Total number of labels in testing set y\_test " , y\_test.shape[0])**

**print(" Total number of samples in validation set X\_test ", X\_val.shape[0])**

**print(" Total number of labels in validation set  y\_test " , y\_val.shape[0])**

**print('The 3 samples of training set')**

**print(X\_train.head(3))**

**print('The 3 random  samples of training set')**

**print(X\_train.sample(3))**

**print('The 3 samples of validation set')**

**print(X\_val.head(3))**

**print('The 3 random  samples of validation set')**

**print(X\_val.sample(3))**

**print('The 3 samples of testing set')**

**print(X\_test.head(3))**

**print('The 3 random  samples of testing set')**

**print(X\_test.sample(3))**

**OUTPUT**

**Total number of samples in the Dataset 1797**

**Total number of labels in the Dataset 1797**

**Total number of samples in training set X\_train 1149**

**Total number of labels in training set y\_train 1149**

**Total number of samples in testing set X\_test 360**

**Total number of labels in testing set y\_test 360**

**Total number of samples in validation set X\_test 288**

**Total number of labels in validation set y\_test 288**

**The 3 samples of training set**

**0 1 2 3 4 5 6 7 8 9 ... 54 55 \**

**146 0.0 0.0 0.0 6.0 14.0 1.0 0.0 0.0 0.0 0.0 ... 12.0 0.0**

**1124 0.0 0.0 0.0 8.0 12.0 0.0 0.0 0.0 0.0 0.0 ... 0.0 0.0**

**1113 0.0 0.0 11.0 8.0 4.0 13.0 16.0 3.0 0.0 2.0 ... 0.0 0.0**

**56 57 58 59 60 61 62 63**

**146 0.0 0.0 0.0 9.0 15.0 14.0 6.0 0.0**

**1124 0.0 0.0 0.0 12.0 10.0 0.0 0.0 0.0**

**1113 0.0 0.0 16.0 7.0 0.0 0.0 0.0 0.0**

**[3 rows x 64 columns]**

**The 3 random samples of training set**

**0 1 2 3 4 5 6 7 8 9 ... 54 55 56 \**

**982 0.0 0.0 8.0 11.0 0.0 0.0 0.0 0.0 0.0 0.0 ... 8.0 5.0 0.0**

**1080 0.0 0.0 9.0 16.0 6.0 0.0 0.0 0.0 0.0 3.0 ... 0.0 0.0 0.0**

**703 0.0 0.0 1.0 13.0 16.0 10.0 0.0 0.0 0.0 1.0 ... 9.0 0.0 0.0**

**57 58 59 60 61 62 63**

**982 0.0 4.0 15.0 16.0 16.0 16.0 16.0**

**1080 0.0 12.0 14.0 3.0 0.0 0.0 0.0**

**703 0.0 0.0 14.0 16.0 11.0 0.0 0.0**

**[3 rows x 64 columns]**

**The 3 samples of validation set**

**0 1 2 3 4 5 6 7 8 9 ... 54 55 \**

**1031 0.0 1.0 15.0 16.0 4.0 0.0 0.0 0.0 0.0 9.0 ... 12.0 1.0**

**807 0.0 0.0 2.0 10.0 10.0 11.0 0.0 0.0 0.0 0.0 ... 1.0 0.0**

**1589 0.0 4.0 16.0 16.0 16.0 15.0 3.0 0.0 0.0 11.0 ... 0.0 0.0**

**56 57 58 59 60 61 62 63**

**1031 0.0 0.0 13.0 16.0 16.0 15.0 11.0 1.0**

**807 0.0 0.0 2.0 14.0 16.0 13.0 0.0 0.0**

**1589 0.0 5.0 16.0 3.0 0.0 0.0 0.0 0.0**

**[3 rows x 64 columns]**

**The 3 random samples of validation set**

**0 1 2 3 4 5 6 7 8 9 ... 54 55 56 \**

**1328 0.0 0.0 0.0 2.0 15.0 8.0 0.0 0.0 0.0 0.0 ... 0.0 0.0 0.0**

**438 0.0 0.0 2.0 12.0 12.0 12.0 9.0 2.0 0.0 0.0 ... 0.0 0.0 0.0**

**1082 0.0 0.0 6.0 16.0 15.0 5.0 0.0 0.0 0.0 1.0 ... 8.0 0.0 0.0**

**57 58 59 60 61 62 63**

**1328 0.0 0.0 2.0 15.0 5.0 0.0 0.0**

**438 0.0 3.0 15.0 3.0 0.0 0.0 0.0**

**1082 0.0 8.0 16.0 16.0 12.0 1.0 0.0**

**[3 rows x 64 columns]**

**The 3 samples of testing set**

**0 1 2 3 4 5 6 7 8 9 ... 54 55 56 \**

**154 0.0 0.0 0.0 3.0 16.0 3.0 0.0 0.0 0.0 0.0 ... 0.0 0.0 0.0**

**867 0.0 0.0 10.0 16.0 16.0 4.0 0.0 0.0 0.0 0.0 ... 2.0 0.0 0.0**

**862 0.0 0.0 5.0 16.0 16.0 16.0 9.0 0.0 0.0 0.0 ... 0.0 0.0 0.0**

**57 58 59 60 61 62 63**

**154 0.0 0.0 4.0 12.0 0.0 0.0 0.0**

**867 0.0 11.0 16.0 16.0 10.0 0.0 0.0**

**862 0.0 11.0 10.0 0.0 0.0 0.0 0.0**

**[3 rows x 64 columns]**

**The 3 random samples of testing set**

**0 1 2 3 4 5 6 7 8 9 ... 54 55 56 \**

**398 0.0 0.0 7.0 13.0 3.0 0.0 0.0 0.0 0.0 0.0 ... 16.0 3.0 0.0**

**287 0.0 0.0 3.0 11.0 13.0 5.0 0.0 0.0 0.0 0.0 ... 11.0 0.0 0.0**

**1115 0.0 0.0 1.0 13.0 2.0 0.0 0.0 0.0 0.0 0.0 ... 14.0 5.0 0.0**

**57 58 59 60 61 62 63**

**398 0.0 6.0 15.0 6.0 9.0 9.0 1.0**

**287 0.0 3.0 9.0 16.0 16.0 6.0 0.0**

**1115 0.0 1.0 11.0 16.0 16.0 13.0 1.0**

**[3 rows x 64 columns]**

**Assignment**

1. **In the above program replace shuffle = True by shuffle = False. Execute two times, note the result and justify the result.**

**2) Split the diabetes dataset in training , testing and validation sets. 25% of the total dataset is reserved for testing set. Out of 75% of the remaining samples, training and validation sets are divided in the ratio 70:30**