Task 1

Problem 1

```
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
import pandas as pd
import numpy as np
```

Reading the dataset

```
data = pd.read_csv("data.txt", sep=" ", header=None)
data = data.to_numpy()
Y = data[:,-1].reshape(len(data),1)
X = data[:,:-1]
```

1. Computing the classification accuracy for the first dataset

```
classification_acc = []
for i in range(10):
    X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.4)
    svc = SVC(kernel ='linear', C = 1)
    svc.fit(X_train,Y_train)
    classification_acc.append(accuracy_score(Y_test,svc.predict(X_test)))

print(classification_acc)
```

```
[0.771875, 0.753125, 0.7625, 0.734375, 0.784375, 0.7625, 0.746875, 0.75625,
0.74375, 0.74375]
```

2. Computing the classification accuracy for the first dataset

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
```

```
scaler.fit(X)
X_scaled = scaler.transform(X)
```

```
classification_acc_scaled = []
for i in range(10):
    X_train, X_test, Y_train, Y_test = train_test_split(X_scaled, Y,
test_size=0.4)
    svc_scaled = SVC(kernel ='linear', C = 1)
    svc_scaled.fit(X_train,Y_train)

classification_acc_scaled.append(accuracy_score(Y_test,svc_scaled.predict(X_test))
)

print(classification_acc_scaled)
```

```
[0.721875, 0.7375, 0.803125, 0.775, 0.75, 0.746875, 0.778125, 0.740625, 0.76875,
0.73125]
```

A. Report the Difference between the dataset used in (1) and those used in (2).

The Second dataset have been scaled by removing the mean and scaling to unit variance by StandardScaler method.

The scaled value of a sample x is calculated as:

z = (x - u) / s

where u is the mean of the training samples, and s is the standard deviation of the training samples

B. Report the averaged accuracy over the ten trails. Note: each time the difference is in the data that is randomly chosen for testing and training.

```
print("The averaged accuracy for dataset 1")
print(np.average(classification_acc))
```

```
The averaged accuracy for dataset 1
0.7559375
```

C. Discuss the difference in the averaged accuracy of (1) and (2).

```
print("The averaged accuracy for dataset 2")
print(np.average(classification_acc_scaled))
```

```
The averaged accuracy for dataset 2
0.7553124999999999
```

the averaged accuracy of (1) and (2) are almost the same

D. Report all the preprocessing steps you did to the data.

The data have been read into dataframe using `pandas` package, then I have converted the dataframe into 2d NumPy array, I have separated the features to variable X and the output to variable Y