LOGBOOK

CP60032E

Lecturer: Massoud Zolgharni

Module: Module Natural Language Interfaces

Pouya Shahverdi Moghaddam
21245645

Table of Contents

Week 2

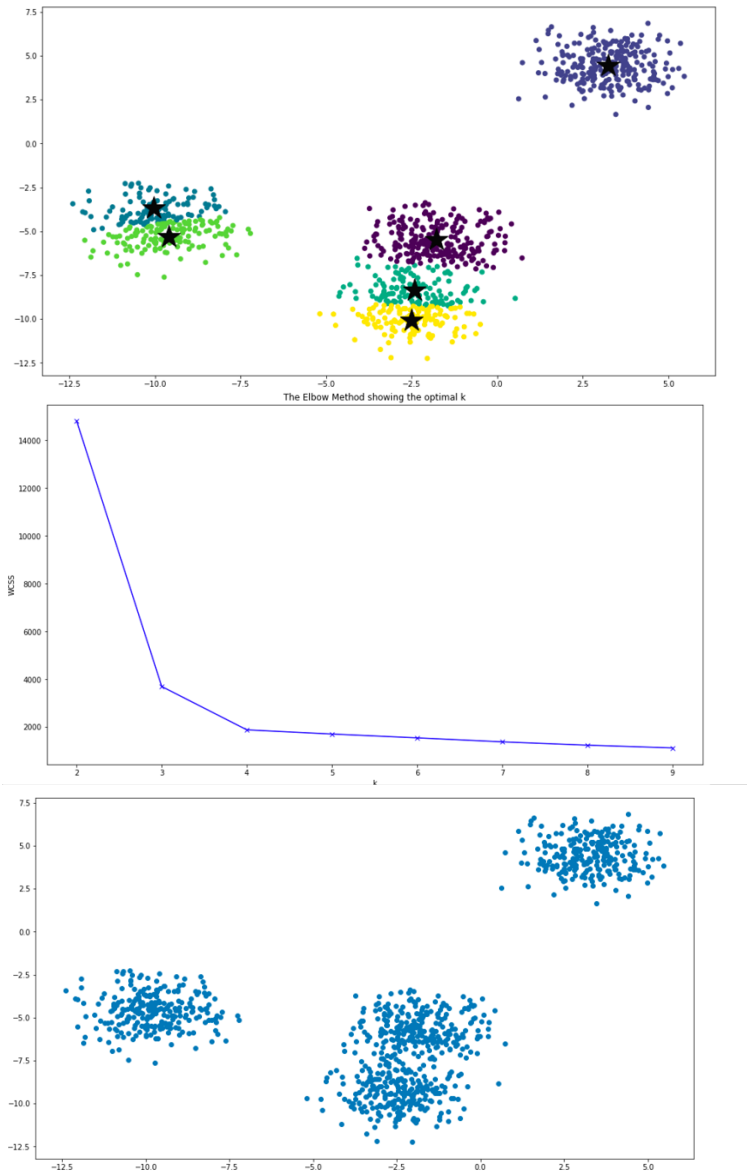
Task 1

Describe the issues associated with K-means algorithm, and explain the mitigating options.

Even though K-means is a faster algorithm compared to the others it does have issues. With the K-means algorithm, it is mostly useful when you know how many clusters you want, we then have the issue of choosing the “wrong” K, which can lead to strange and unreliable results, furthermore results can depend on the initial centroids which fall under the mitigating options. For this you would need to try multiple set of randomly chosen initial centroids and then select the best result.

Task 2

The dataset we are going to use has 1000 entries with unknown number of clusters (download data_Kclusters.txt from Blackboard). Use the Elbow Method to find the optimum value for K and show the results for clustering using the optimum k. You can use the following lines to plot the elbow.



Week 3

Task 1-

Explain why accuracy alone might not be a good performance measure for classification models?

When there is a class imbalance accuracy is not a meaningful measure, for example we can have a 99% chance of accuracy in a matter such as a cancer screening machine where patients are sent for screening based on the machine. We can say 99% accuracy can be classed as very high under normal circumstances and this can be seen as a success but the 1% can be fatal as 1% error when dealing with human lives where a patient can die will make the machine dangerous. Even taking into accounts the money side of things, the cost of trying to fix the 1% error can end up costing more than the money saved from the 99%.

Task 2-

Let us assume we developed a machine learning model to predict which patients on dialysis will be admitted to the hospital in the next week. Given 100 patients, we have the following break-down:

True Positives: people that are hospitalized that you predict will be hospitalized
True Negatives: people that are NOT hospitalized that you predict will NOT be hospitalized
False Positives: people that are NOT hospitalized that you predict will be hospitalized
False Negatives: people that are hospitalized that you predict will NOT be hospitalized Calculate the performance metrics in the table below:

Calculate the performance metrics in the table below:

Confusion Matrix		Actual	
		Hospitalized	Not Hospitalized
Predicted	Hospitalized	33	10
	Not Hospitalized	17	40

Recall (Sensitivity)	
Specificity	
Accuracy	
Precision	

Recall (Sensitivity)	0.66
Specificity	0.8
Accuracy	0.73
Precision	0.7674

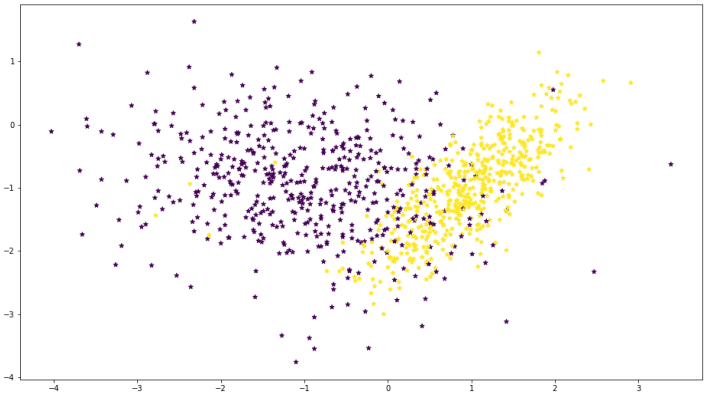
Task 3 Try K-NN classifier for dataset provided in for different numbers of neighbours (K = 3 and K = 7). Which K provides better results? Justify your answer by using AUC of ROC curve.

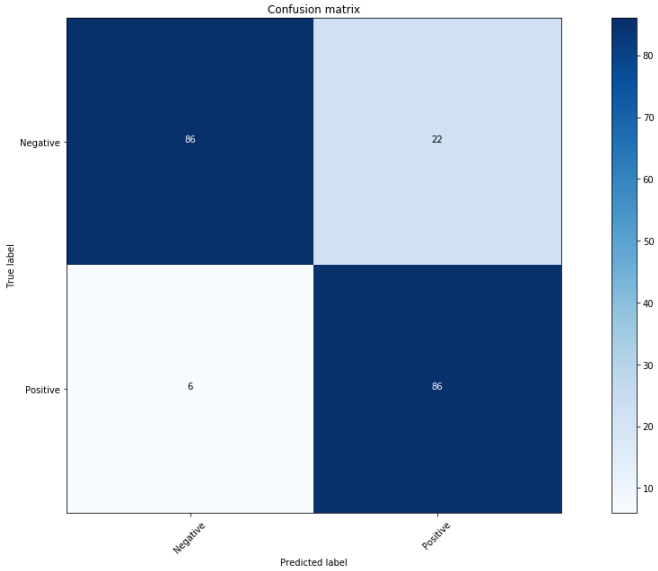
K=3

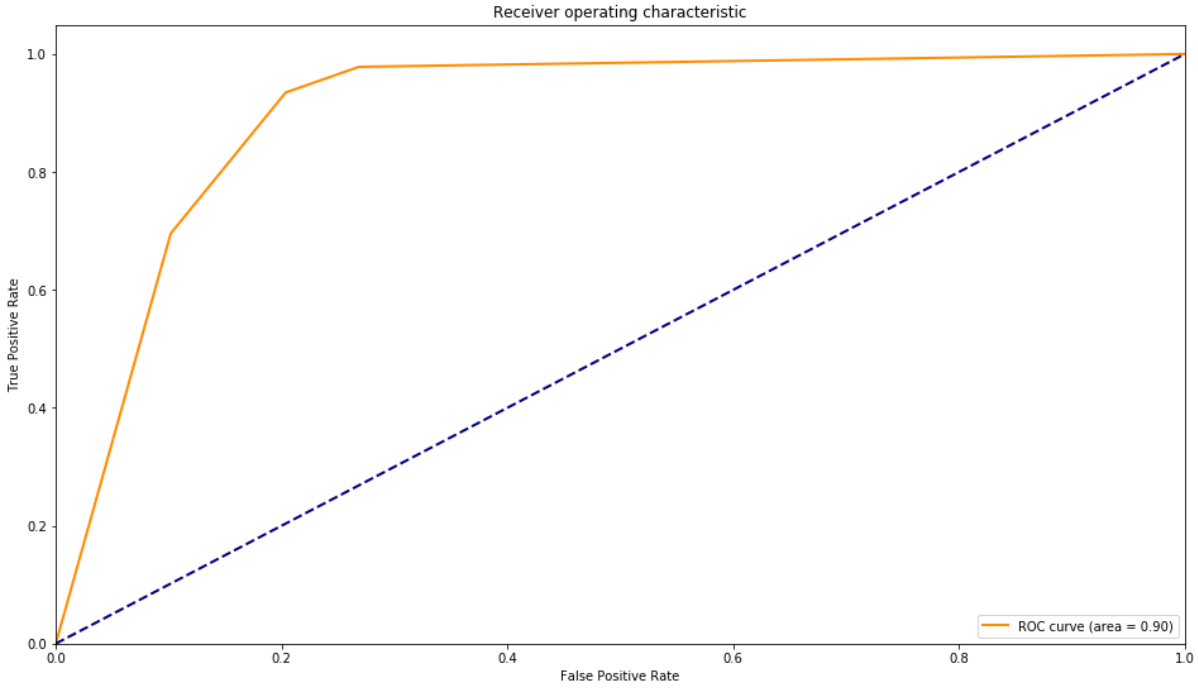
```
TP =  86
FP =  22
FN =  6
TN =  86

Performance Metrics
TPR = Sensitivity = recal = 0.93
TNR = Specificity = 0.80
PPV = Precision = 0.80
ACC = Accuracy = 0.86

AUC: 0.904
```





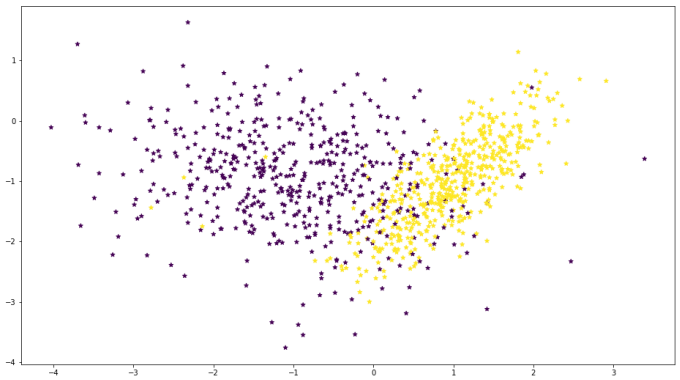


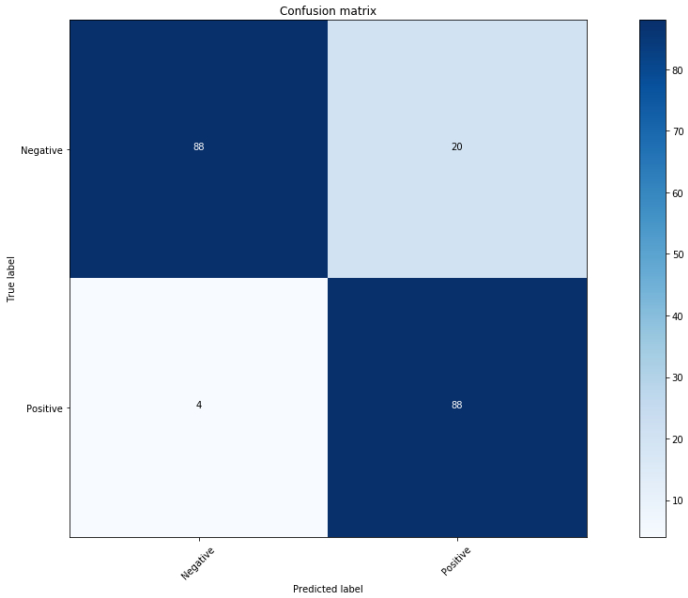
K = 7

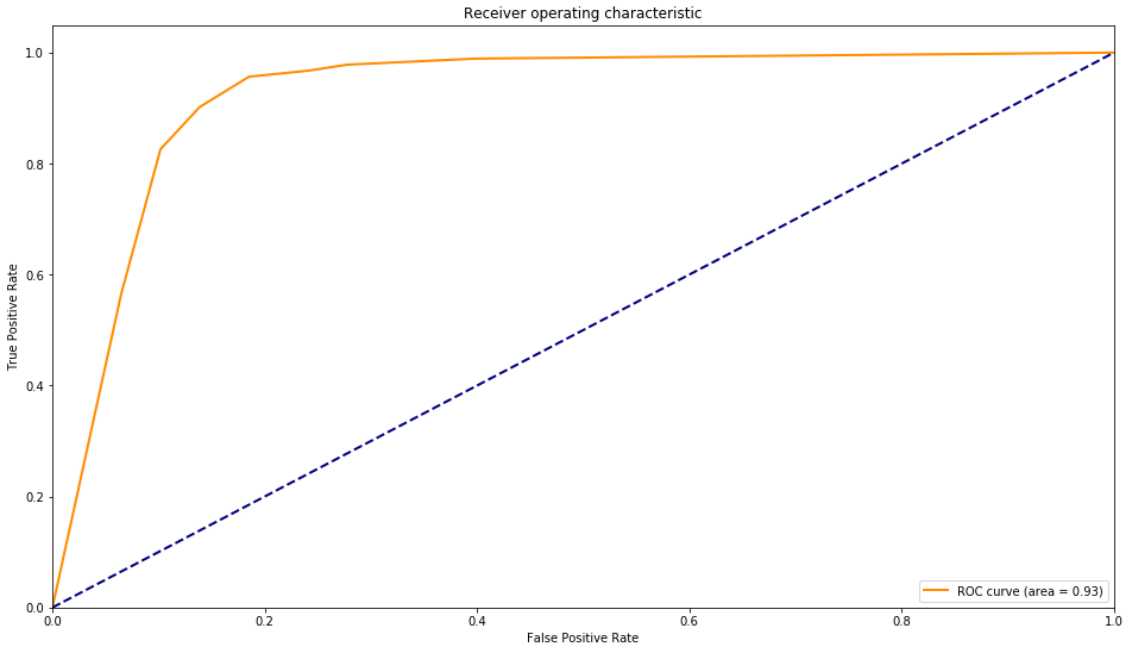
```
TP =  88
FP =  20
FN =  4
TN =  88

Performance Metrics
TPR = Sensitivity = recal = 0.96
TNR = Specificity = 0.81
PPV = Precision = 0.81
ACC = Accuracy = 0.88

AUC: 0.926
```







Based on the findings above k = 7 achieves better results

Week 4

Task 1

Explain the difference between the classification and regression problems in machine learning. Support you answer with one or two examples.

Regression and Classification are both under the same categorisation however where they differ is that one is used to predict a quantity and one is used to predict a classification.
Regression will predict value from a continues set unlike classification predicting the class data point which is normally referred to as a "label"

One example we can have for describing a regression system is a rain predicting machine. The machine will make predication based on a previous data/values. This would be the classic version of regression where as an example of a classification would be a spam filter machine learning program where the program will learn to flag spam by using certain keywords, this would be the data.

Task 2

For this task, download Advertising.csv from Blackboard. This is a simple data set that contains, in unit of £, the marketing spend along with the company sales for each month. Using the advertising data, model the linear relationship between the marketing budget and sales. Then answer the following questions:

- How much sale can we expect when the marketing budget is zero?
- For a marketing spend of £7,500, how much sale can we expect?

You can use Linear_regression.py for this task.

Using the Linear_regression.py means that we have to make some changes in order to accommodate the above.

Changing the dataset to "Advertising.csv
Changing the X label to Marketing Budget
Changing the Y label to Sales
Changing the plt.title to Dataset

- How much sale can we expect when the marketing budget is zero?

We need to use the line equation in order to work out the above.

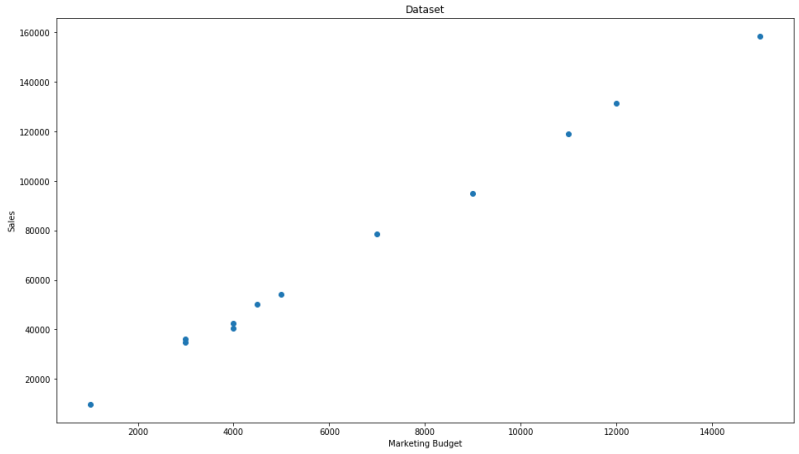
Y=mx+b

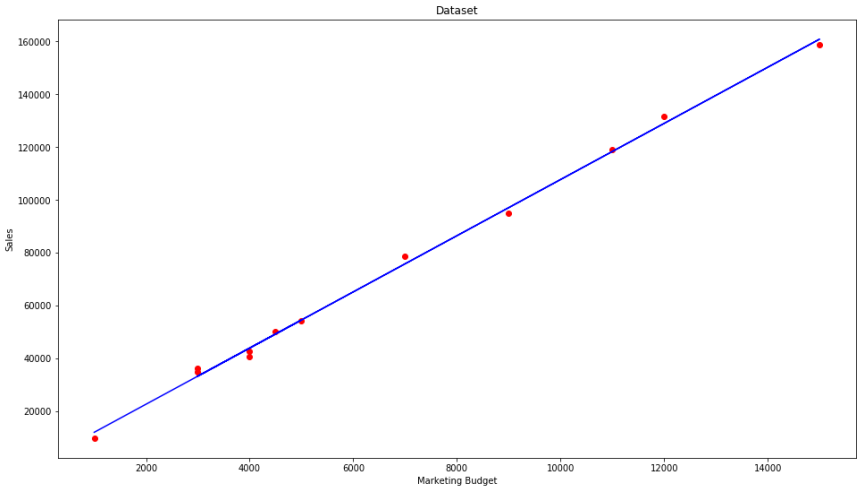
```
Line equation: y = mx + b
Optimized m: [[10.62219546]]
Optimized b: [1383.47138013]
```

Y = 10.62 * (0) + 1383.47 = 1383.47

• For a marketing spend of £7,500, how much sale can we expect?

Y = 10.62 * (7500) = 1383.47 = 81,033.47



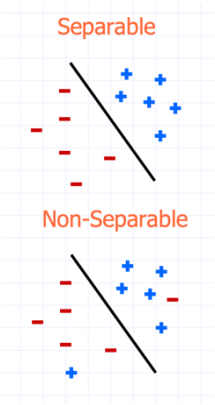


Week 5

Task 1

Explain why logistic regression might be a better classifier than Perceptron. Support your answer with examples.

One of the reasons why Perceptron is not as good as logistic regression is due to to a number of reasons. We have the issue with separability for when there is a linear separator, perceptron will find but this can cause issues because the operation can go on forever unless we put a maximum limit on number of irritations.



Task 2

Task 2

For this task, download data_Seminar_W5.txt from Blackboard. This is a simple data set that contains 1000 entries with 2 features and corresponding class labels (0 or 1) for each entry.

0.47	3.87	0.00
2.84	3.33	0.00
0.61	2.51	0.00
3.82	1.65	1.00
1.28	0.63	1.00
0.99	5.87	0.00
1.05	-0.10	1.00
0.91	4.20	0.00
0.88	3.64	0.00
0.91	-0.41	1.00
0.89	4.50	0.00
3.09	1.38	1.00

Snapshot of data (first two columns are features, and last column is the class label).

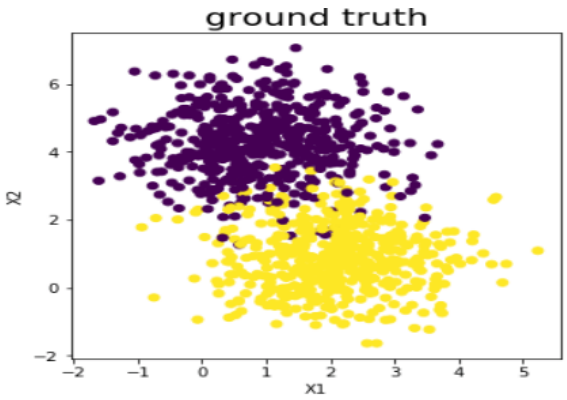
Apply the Perceptron algorithm to obtain a linear classifier and plot the results.

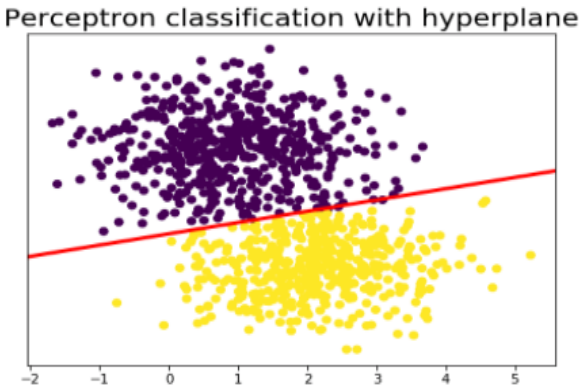
- Is the data linearly separable? Justify your answer.
- If linearly inseparable, what is the percentage of misclassified points using your linear classifier??

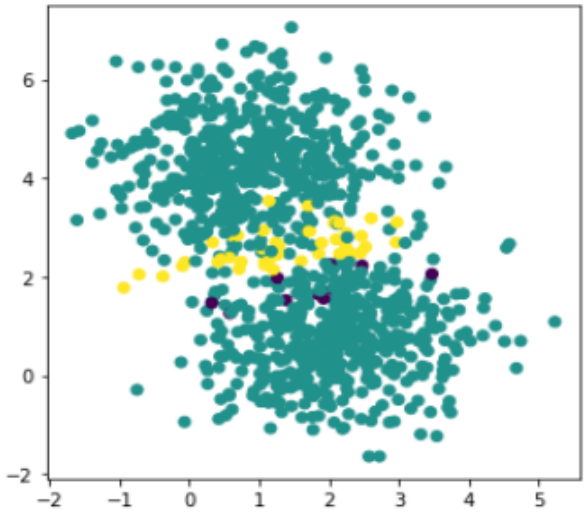
You can use Perceptron.py for this task, but it needs to be modified. You can load the data as follows:

```
DATA = np.loadtxt( "data_Seminar_W5.txt" )
X = DATA[:, 0:2]
Y = DATA[:, 2]
```

Tip: compute the total number of misclassified points (classified as 0 where it should be 1, or classified as 1 where it should be zero), and then divide by total number of points (i.e. 1000).

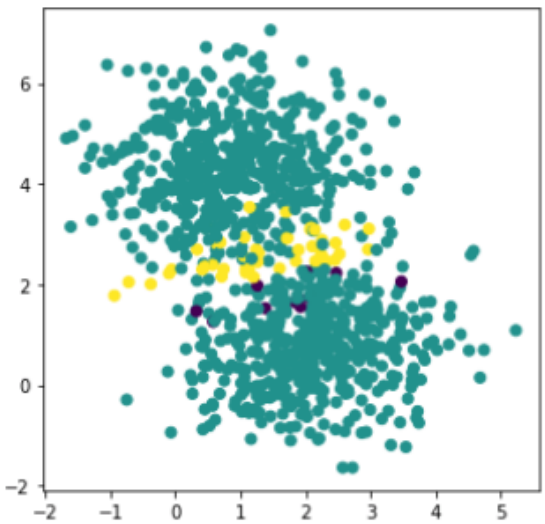






58.0

```
misclassified_point =  Y - result_class
a = np.absolute(misclassified_point)
a= sum (a)
print (a)
```



58.0

Percentage of misclassified points 5.8

Week 6

Task 1

Explain how a Multi-layer Perceptron can address the limitation of a single-layer Perceptron.

Multi-layer perceptron has one major advantage compared to the single-layer perceptron and that is MLP's contain one or more hidden layers, whereas a single-layer perceptron doesn't and thus will only end up learning linear functions also the multi-layer perceptron
will also learn non-leaner functions.

A Multi-layer perceptron can also be used for function approximation and regression.

Task 2

Explain what over-fitting is and how it can be avoided (only one technique to avoid over-fitting is required here, but it must be properly explained).

Overfitting occurs when a model fit the noise in the data instead of the relationship, this occurs when the model learns the data too well, to an extent that it has learn all the detail and noise in the training data, which ends up impacting the performance of the model in a negative manner. This is especially problematic because the new data the model is working on will not contain the same details and noise and therefore the impact of the learnt model is negative on the data with bad accuracy.

There are two main methods used that can prevent over-fitting. We can use a validation dataset and we can also use a resampling technique.

With the validation dataset, this is a subset of the training data used and this is used right at the end after the training and before the testing. Once the machine has been tuned and learned the validation dataset