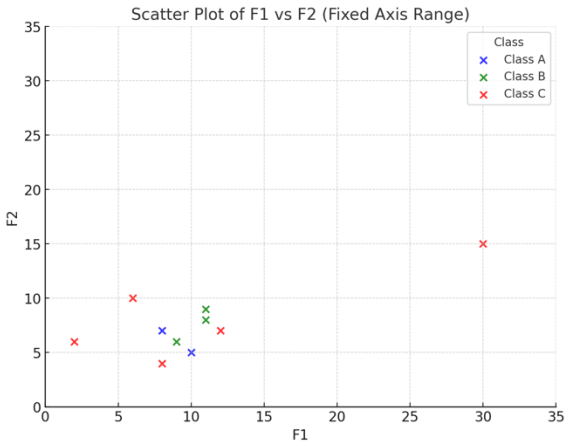
Training Exercises

1. Consider the following list of examples with 2 features and the corresponding class.

Sample	F1	F2	Class
A1	10	5	A
A2	8	7	A
A3	11	8	B
A4	9	6	B
A5	11	9	B
A6	6	10	C
A7	2	6	C
A8	30	15	C
A9	12	7	C
A10	8	4	C



1.1 Assuming a normal distribution of the feature F1, identify possible outliers using the z-score method (threshold k=2.5)

1.2 Compute the Fisher Score of each feature.

1.3 Let us apply an approach similar to the SMOTE method to generate two new samples of class B. We will use the A4 example as starting point and generate a new point at half the distance to the example A3, and another point at ¼ of the distance to A5.

1.4 Using a k-NearestNeighbour classifier with k=3 (majority voting), classify the new example (F1=10,F2=10). In case of a tie, the class of the closest neighbour should be selected.

2. Consider the following PCA eigenvectors W and eigenvalues λ. Project the dataset into the first principal component (the one that explains the most variance of the dataset).
   W = [0.2, 0.6; 0.8, 0.3] λ = [10, 40]

3. Compute the dice coefficient between the following registers (ignore punctuation and case):
   S1 = "my name is Yoda, a powerful jedi"
   S2 = "Yoda, my name is. Powerful jedi, I am"

4. A classifier of incoming emails sorts the emails into the classes {B – Business, A – Advertising, M – Miscellaneous}. Given the following table of real and predicted values, compute:

Id	Real	Predicted	Id	Real	Predicted
1	B	M	11	M	M
2	B	B	12	M	B
3	B	M	13	M	A
4	B	B	14	M	B
5	B	B	15	A	M
6	B	M	16	A	A
7	B	B	17	A	A
8	B	B	18	A	A
9	M	M	19	A	A
10	M	M	20	A	M

   4.1 Confusion Matrix
   4.2 Class-specific Precision, Recall and f1-score
   4.3 Global (weighted) Precision, Recall and f1-score

5. Explain the concept of overfitting and how it can be detected in a simple train-test split evaluation strategy.