NAME: JOSHUA ROBERT JOHN KENNEDY
MATRICULATION NUMBER : G2305691H

EE6407 – GENETIC ALGORITHM & MACHINE LEARNING

ASSIGNMENT 2

1. Introduction

This assignment uses Support Vector Machine (SVM) classification to analyse a multi-feature dataset. Two datasets are used in the study: a test set (TestData.xlsx) with 30 samples without labels and a training set (TrainingData.xlsx) with four feature columns and a class label column. The following are the main goals of this study:

1. Identify and address missing values and outliers in the training data.
2. Develop and train an SVM classifier using the pre-processed training data.
3. Apply the trained classifier to predict class labels for the test data.

2. Data Analysis and Pre-processing

2.1 Missing Values

Initial examination of the training dataset revealed the presence of missing values across the feature set:

- Feature1: 2 missing values
- Feature2: 0 missing values
- Feature3: 0 missing values
- Feature4: 1 missing value

I used mean imputation, a technique that substitutes the mean value of the corresponding feature for missing data points, to deal with these missing values. This method was chosen because it could reasonably approximate the missing values while preserving the data's general distribution. By using this approach, the dataset's integrity is maintained without the underlying patterns being noticeably distorted.

2.2 Outlier Detection and Treatment

For every feature, box plot visualisation was used to identify outliers. As seen in Fig. 1, this analysis indicated the existence of outliers. I used the Interquartile Range (IQR) approach, which entails the following actions, to lessen the impact of these outliers:

1. Calculation of Q1 (25th percentile) and Q3 (75th percentile) for each feature.
2. Computation of the IQR (IQR = Q3 - Q1).
3. Definition of outliers as data points falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR.
4. Replacement of identified outliers with the mean of the respective feature.

The resilience of this approach in detecting extreme values while preserving the dataset's general structure led to its selection. This method aids in lessening the impact of unusual data points on the analysis and model training that follows.

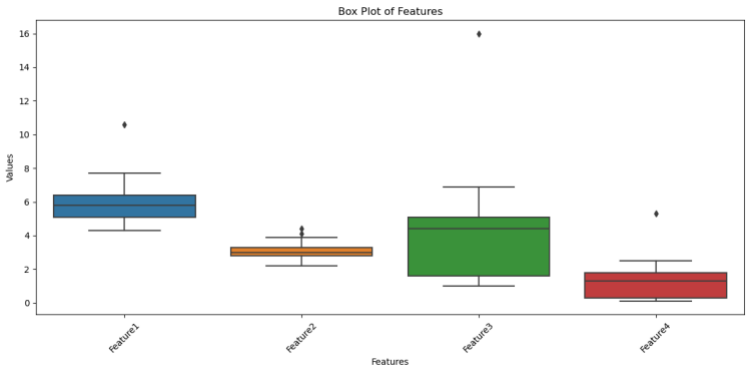

Figure :1 Box plot of Training dataset before pre-processing

2.3 Pre-processing Results

Following the implementation of missing value imputation and outlier treatment, the dataset exhibited the following characteristics:

1. Complete elimination of missing values across all features.
2. Reduction in the range of values for Features indicating successful mitigation of extreme outliers as shown in Fig:2.
3. Slight alterations in feature correlations, with the most notable correlation observed between Feature3 and Feature4 (0.949121).

The summary statistics for the pre-processed features are presented in Table 1.

Table 1: Summary Statistics of Pre-processed Features

	Feature1	**Feature2**	**Feature3**	**Feature4**
count	120.000000	120.000000	120.000000	120.000000
mean	5.839470	3.032542	3.749590	1.203116
std	0.797379	0.386086	1.758025	0.761200
min	4.300000	2.200000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.325000	3.300000	5.100000	1.800000
max	7.700000	3.900000	6.900000	2.500000

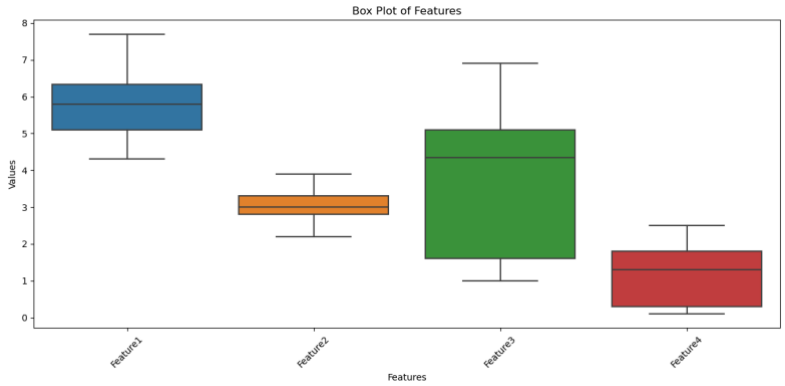

Figure :2 Box plot of Training dataset After pre-processing

3. SVM Classifier Training

A Support Vector Machine classifier was selected for its effectiveness in handling multi-class classification problems and its ability to identify optimal decision boundaries in high-dimensional feature spaces.

3.1 Classifier Parameters

The SVM classifier was configured with the following parameters:

- Kernel: Linear

Because of its ease of use, interpretability, and efficiency in situations involving linearly separable data, the linear kernel was selected. This choice strikes a mix between model performance and complexity, especially considering how short the training dataset is.

4. Test Data Prediction

The test dataset underwent pre-processing steps consistent with those applied to the training data to ensure uniformity in data treatment. These steps included:

1. Replacement of '?' values with NaN (Not a Number).
2. Imputation of NaN values using means calculated from the training data.

Following pre-processing, the trained SVM classifier was applied to predict class labels for the 30 samples in the test dataset.

4.1 Prediction Results

The SVM classifier generated the following class label predictions for the test dataset:
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3]
This prediction set comprises:

- 10 samples classified as Class 1
- 11 samples classified as Class 2
- 9 samples classified as Class 3

5. Discussion

The analysis and classification process revealed several notable points:

1. **Data Quality**: The training data's outliers and missing values highlight how crucial comprehensive data pre-processing is for machine learning applications. A more reliable dataset for model training was produced thanks to the application of mean imputation and IQR-based outlier management techniques.
2. **Feature Correlations**: The high correlation observed between Feature3 and Feature4 (0.949121) suggests potential multicollinearity. This finding may warrant further investigation into feature selection or dimensionality reduction techniques in future iterations of the model.
3. **Class Distribution**: The test data shows a well-balanced distribution of classes, as indicated by the prediction results, indicating that the classifier does not show appreciable bias towards any certain class. This balance is a good measure of how well the model performs in various class areas.
4. **Model Simplicity**: There is a trade-off between interpretability and model complexity when using a linear kernel SVM. With the training dataset being very short (120 samples), this decision was made to avoid overfitting.

6. Conclusion

In summary, this investigation showed how to effectively use an SVM classifier to predict class labels using a dataset of four features. Through methodical handling of outliers and missing values, the data was ready to produce the best possible outcomes. The test samples might be divided into three groups by the SVM model using a linear kernel.