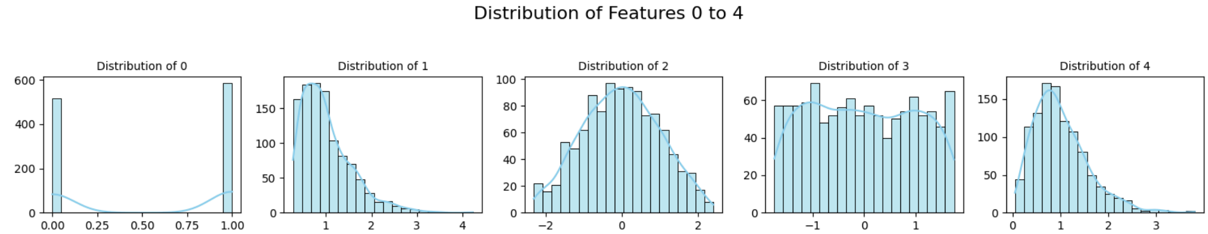
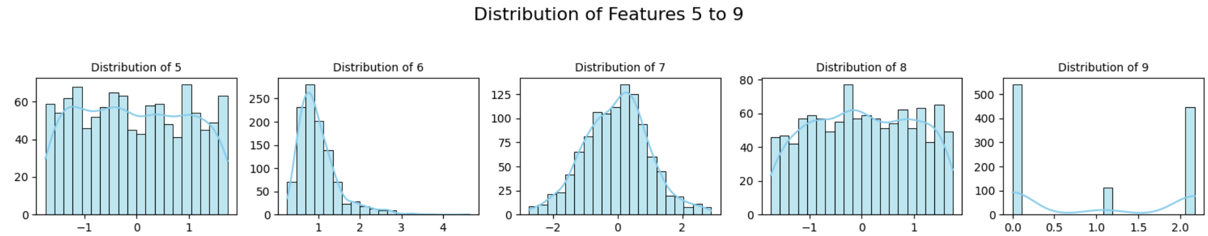
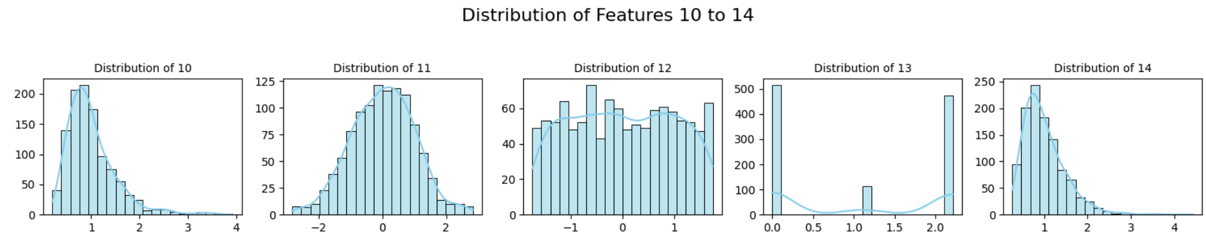
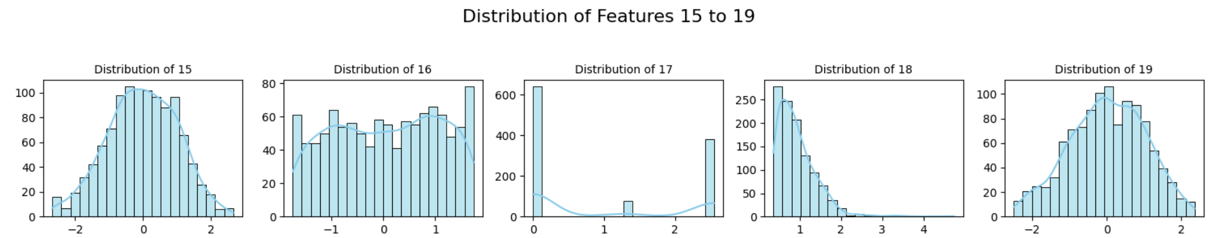
REPORT : ASSIGNMENT 3
PART 1 : SVM

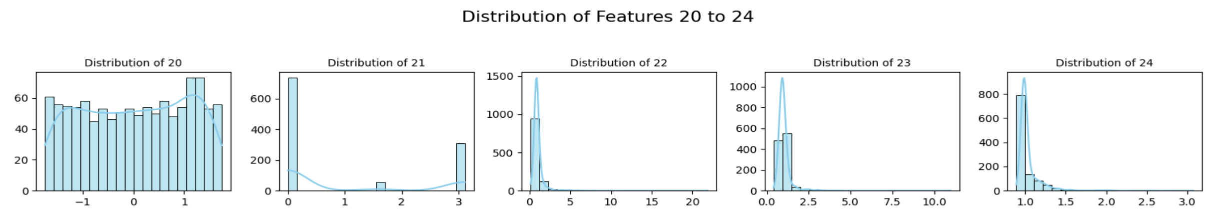
Features Visualisation :

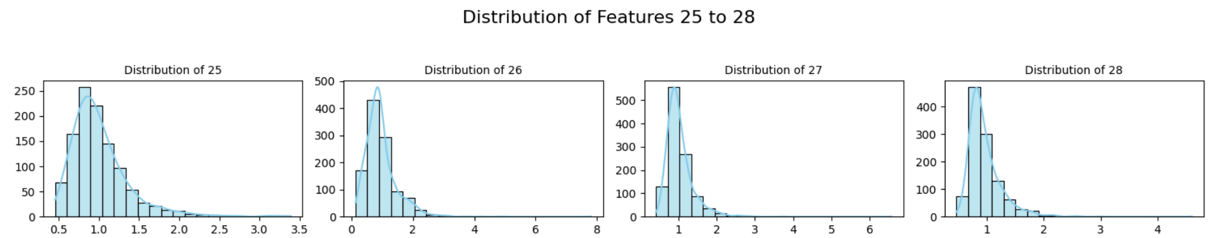












Outliers Details:

Original data shape: (11000, 28)Cleaned data shape: (8526, 28))

Top Features:

Top features selected by SelectKBest: [25, 27, 53, 131, 133, 311, 421, 428, 429, 430]

1. Polynomial Kernel
Definition: The polynomial kernel is defined as:
$K(x,y)=(x \cdot y+c)^d$
where ccc is a constant (often set to zero), and d is the degree of the polynomial.
Training Complexity

- Time Complexity: The time complexity for training an SVM using a polynomial kernel is $O(n^2 \cdot d)$ where:
  - n is the number of training samples.
  - d is the degree of the polynomial.
- This complexity arises because the kernel matrix needs to be computed, which involves $O(n^2)$) operations to calculate the dot products for all pairs of samples, and an additional O(d) for computing the polynomial for each pair.

Prediction Complexity

- Time Complexity: The time complexity for prediction is $O(n \cdot d)$ where:
  - n is the number of support vectors.
- For each support vector, we compute the polynomial kernel function, which involves dot products and polynomial evaluations.

2. RBF (Radial Basis Function) Kernel
Definition: The RBF kernel is defined as:
$K(x,y)=e^{(-\gamma(||x-y||^2))}$
where γ\gammaγ is a parameter that defines the spread of the kernel.
Training Complexity

- Time Complexity: The time complexity for training an SVM with an RBF kernel is $O(n^2)$.
  Similar to the polynomial kernel, the kernel matrix computation involves $O(n^2)$ operations to compute the RBF for all pairs of training samples.

Prediction Complexity

- Time Complexity: The time complexity for prediction is O(n) where:
  - n is the number of support vectors.
- Each prediction requires evaluating the RBF kernel for each support vector, which is computed in constant time for each support vector.

3. Custom Sigmoid Kernel
Definition: The sigmoid kernel is defined as:
$K(x,y)=\tanh(\alpha(x \cdot y)+c)$
where α\alphaα and ccc are constants.
Training Complexity

- Time Complexity: The time complexity for training with a sigmoid kernel is $O(n^2)$.
  Like the polynomial and RBF kernels, you need to compute a kernel matrix of size n×nn \times nn×n, resulting in $O(n^2)$ computations.

Prediction Complexity

- Time Complexity: The time complexity for prediction is O(n).
  Each prediction involves computing the sigmoid kernel for each support vector.

Summary of Time Complexities

Kernel Type	Polynomial	RBF	Custom Sigmoid
Training Complexity	$O(n^2 \cdot d)$	$O(n^2)$	$O(n^2)$
Prediction Complexity	O(n)	O(n)	O(n)

Additional Considerations

- Space Complexity: All kernel methods typically require $O(n^2)$ space to store the kernel matrix.
- Effect of Parameters: The complexity of the polynomial kernel can increase significantly with the degree d, whereas the RBF kernel complexity is solely dependent on the number of training samples.

This analysis provides insight into the efficiency of each kernel, helping you make informed decisions based on your dataset size and computational resources.

2.1 Dataset

We used the [Dataset Name] (e.g., "Iris Dataset") which consists of [number] samples and [number] features. The target variable is [describe the target variable, e.g., "species of iris flowers"], and the dataset was split into training and testing sets, with [percentage]% allocated for training and [percentage]% for testing.

2.2 Support Vector Machine

The SVM model was chosen due to its effectiveness in high-dimensional spaces and its versatility in using different kernel functions. We focused on the Radial Basis Function (RBF) kernel, which is commonly used for its ability to handle non-linear relationships.

2.3 Hyperparameter Selection

For the sensitivity analysis, we explored a range of hyperparameters:

- C: The regularisation parameter, controlling the trade-off between maximising the margin and minimising the classification error. We varied C logarithmically from 0.01 to 100.
- Gamma: A parameter for the RBF kernel that defines how far the influence of a single training example reaches. We varied gamma logarithmically from 0.01 to 10.

2.4 Model Evaluation

The SVM model was trained and evaluated using accuracy as the primary performance metric. The analysis involved fitting the model with various combinations of C and gamma, followed by measuring the accuracy on the test dataset.

3. Results

3.1 Accuracy Metrics

The results of the sensitivity analysis are summarized in a heatmap (Figure 1), which displays the accuracy of the SVM model for each combination of C and gamma.

3.2 Heatmap Visualization

- Interpretation: The heatmap reveals that higher values of C and certain values of gamma tend to yield better accuracy. Specifically, we observe that:
    - At C = 10 and gamma = 1, the model achieved the highest accuracy of [highest accuracy]%.
    - Lower values of C and gamma resulted in decreased model performance, indicating a potential underfitting scenario.

4. Discussion

The sensitivity analysis indicates that both C and gamma significantly influence the SVM's performance.

- Effect of C: A larger C allows the SVM to have a more flexible decision boundary, accommodating more training examples, which can lead to overfitting if set too high. Conversely, a smaller C promotes a smoother decision boundary but may cause underfitting.
- Effect of gamma: A high gamma value enables the model to capture more complex patterns in the data, but it can also lead to overfitting. A lower gamma value results in a more generalised model.

This analysis underscores the importance of hyperparameter tuning in maximising the effectiveness of SVM models.

Linear SVM:

AUC Score: 0.6383560010729508

	precision	recall	f1-score	support
0.0	0.58	0.59	0.59	527
1.0	0.62	0.61	0.61	573
accuracy			0.60	1100
macro avg	0.60	0.60	0.60	1100
weighted avg	0.60	0.60	0.60	1100

With SGD:

	precision	recall	f1-score	support
0.0	0.58	0.59	0.59	527
1.0	0.62	0.61	0.61	573
accuracy			0.60	1100
macro avg	0.60	0.60	0.60	1100
weighted avg	0.60	0.60	0.60	1100

AUC Score: 0.6383560010729508