CS60050 Machine Learning Assignment 3

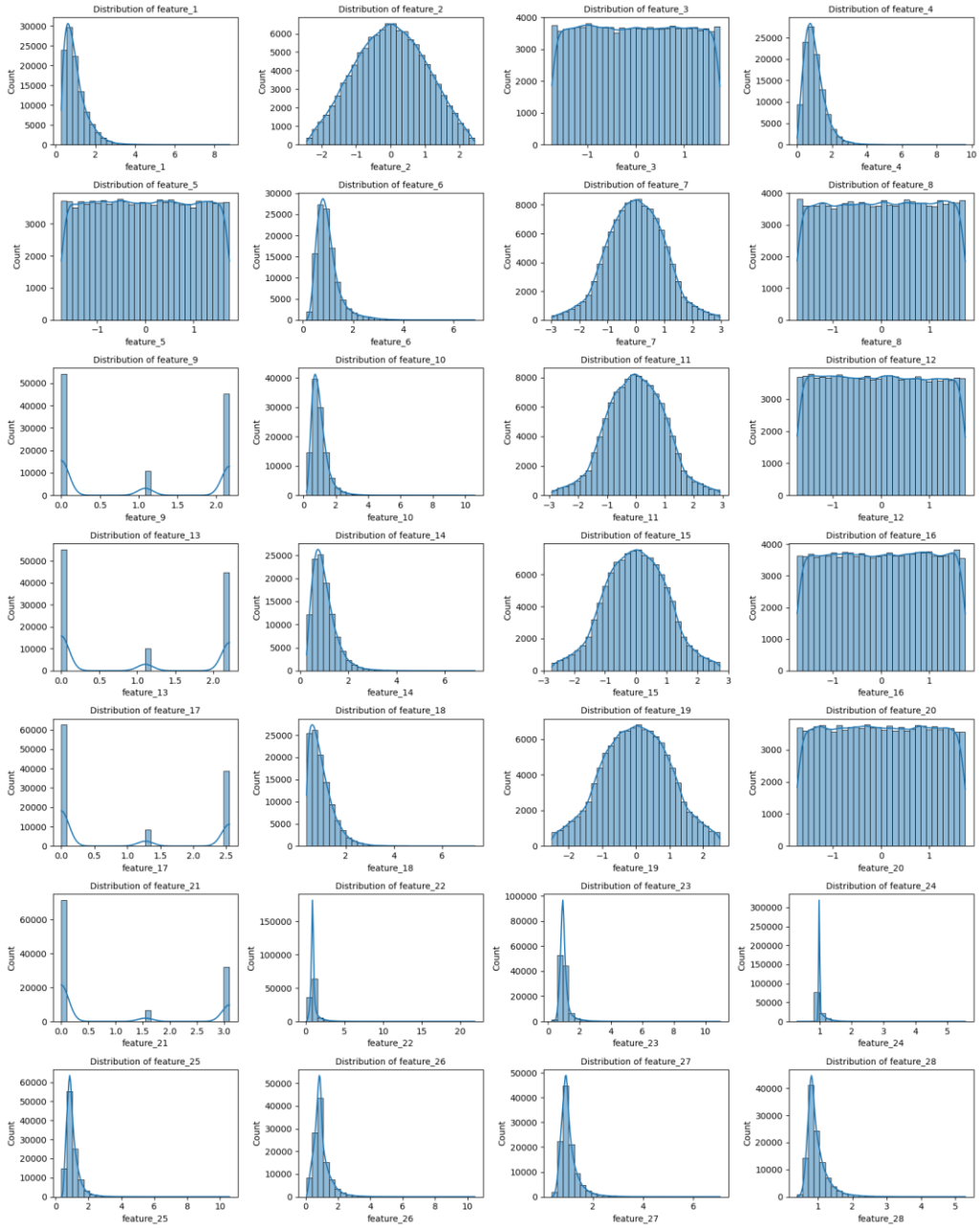
24CS60R77 – Ishan Rai

Part A (SVM)

1. Data Preprocessing and Exploration

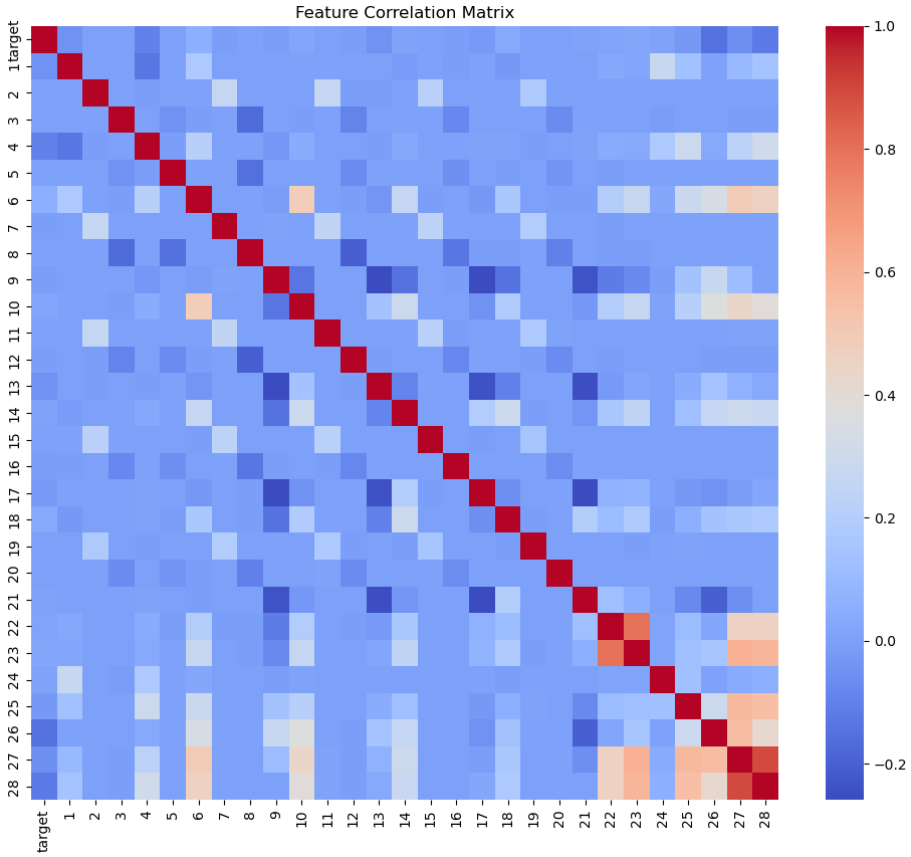
The first 21 features (columns 2-22) are kinematic properties measured by the particle detectors in the accelerator. The last seven features are functions of the first 21 features; these are high-level features derived by physicists to help discriminate between the two classes. The first column is the class label (1 for signal, 0 for background), followed by the 28 features (21 low-level features then 7 high-level features).

Histogram:



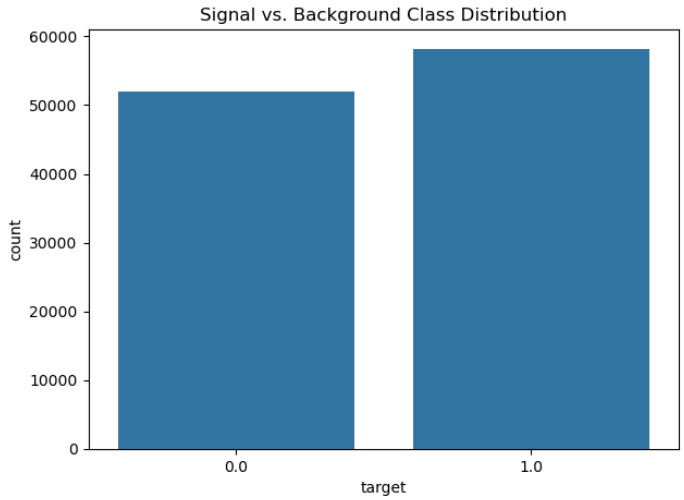
The distributions of the features vary widely, ranging from approximately normal to highly skewed. This suggests that the dataset contains a mix of continuous and potentially discrete features

Correlation Matrix:



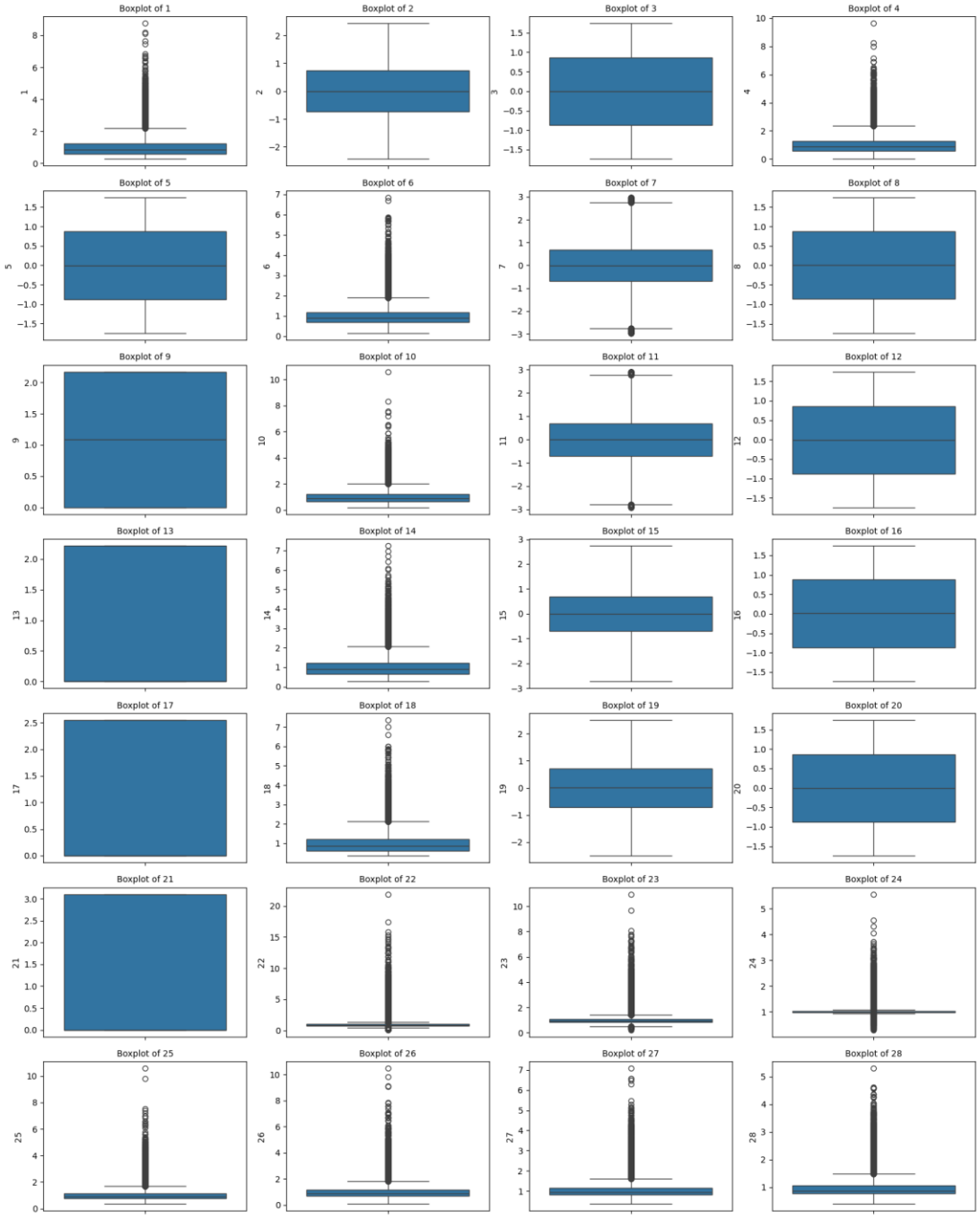
In this matrix, we can observe that there are several features that are highly correlated with each other, forming distinct blocks. This suggests that some features might be redundant or provide similar information, which could be considered for feature selection or dimensionality reduction techniques.

Bar Graph:



The bar chart shows the distribution of two classes, signal and background. The height of each bar represents the number of instances in each class. In this case, the background class has a slightly higher count than the signal class.

Box Plot:



The boxplots show the distribution of each feature in the dataset. We can observe that some features have a wider range of values and more outliers than others. Some features are also skewed, with the median not being centered within the box. Overall, the boxplots provide a visual summary of the distribution and variability of each feature.

Feature Engineering:
A pipeline was implemented tha does the following things:

1. **Imputation:** Handles missing values using median imputation.
2. **Scaling:** Standardizes the features for better model performance.
3. **Feature Engineering:** Creates new features by combining existing ones using polynomial features with interaction terms.

The transformed dataset, including the target label, is then ready for further analysis and modeling.

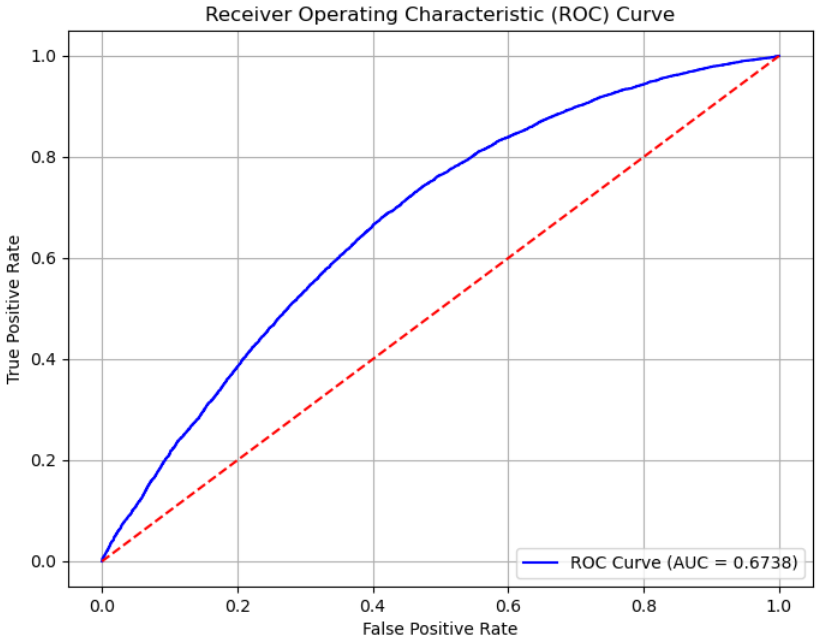
Feature selection:
The code employs the SelectKBest method to identify the top 10 most significant features from the transformed dataset. The f_classif scoring function is utilized to assess the predictive power of each feature in relation to the target variable (label). Subsequently, the selected features are displayed on the console.

```
Selected features by SelectKBest:
Index (['feature_4', 'feature_26', 'feature_28', 'feature_1 feature_6',
       'feature_1 feature_28', 'feature_6 feature_9', 'feature_6
feature_26',
       'feature_9 feature_10', 'feature_9 feature_14', 'feature_9
feature_26'],
      type='object')
```

2. Linear SVM Implementation

Key classification metrics:

```
Accuracy: 0.5941
Precision: 0.5800
Recall: 0.8161
F1-Score: 0.6781
AUC (Area Under the ROC Curve): 0.6738
```



After using mini-batch learning for SVM:

```
Accuracy: 0.4899
Precision: 0.5125
Recall: 0.5405
F1-Score: 0.5261
```

3. SVM with Polynomial, RBF, and Custom Kernels

Polynomial Kernel:
```
Best Parameters for Polynomial Kernel: {'C': 10, 'degree': 2}
```

```
Polynomial Kernel - Accuracy: 0.6143, Precision: 0.5892, Recall: 0.8711,
F1-Score: 0.7029, AUC: 0.6868
Polynomial Kernel - Training Time: 240.62 seconds
```

RBF – Kernel:

```
Best Parameters for RBF Kernel: {'C': 10, 'gamma': 0.13167456935454494}
RBF Kernel - Accuracy: 0.7328, Precision: 0.7192, Recall: 0.8039, F1-Score:
0.7592, AUC: 0.8146
RBF Kernel - Training Time: 528.67 seconds
```

Custom (Sigmoid) Kernel:

```
Custom Kernel - Accuracy: 0.4990, Precision: 0.5220, Recall: 0.5195, F1-
Score: 0.5207, AUC: 0.4991
Custom Kernel - Training Time: 89.46 seconds
```

4. Hyper parameter Tuning

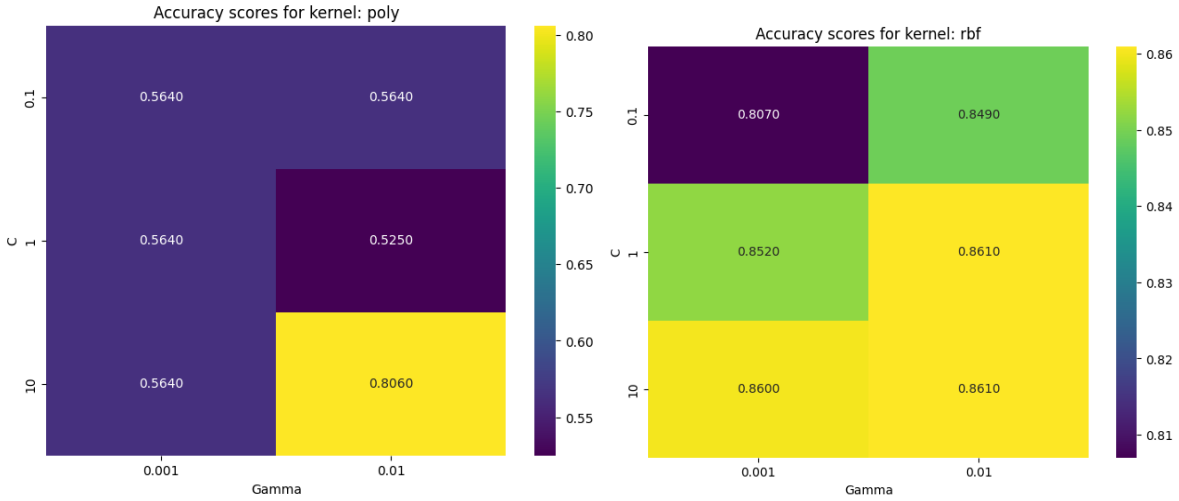
Bayesian Optimization for SVM:

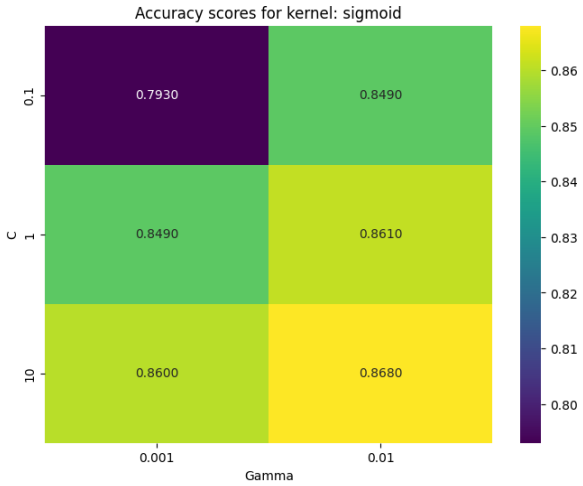
```
Best Parameters for RBF Kernel: OrderedDict[('C', 9.948719998234102),
('gamma', 0.05650888599199564)])

RBF Kernel - Accuracy: 0.6882, Precision: 0.6672, Recall: 0.8077, F1-Score:
0.7308, AUC: 0.7700
RBF Kernel - Training Time: 297.19 seconds
```

Hyper parameter Sensitivity Analysis:

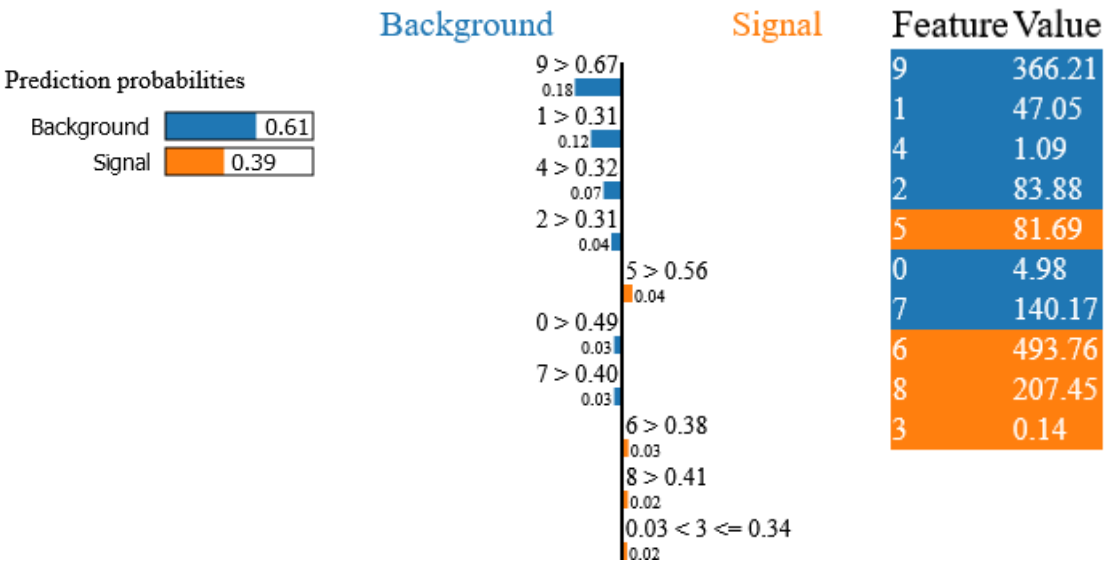
Heat map given below illustrates the accuracy of an SVM model using a polynomial kernel for different combinations of C (regularization parameter) and Gamma (kernel coefficient). The color intensity represents accuracy, with warmer colors indicating higher accuracy.





5. Analysis and Report

LIME (Local Interpretable Model-Agnostic Explanations)



The LIME explanations reveal the factors driving the model's predictions. For instance, feature 9 with a value of 366.21 significantly influences the prediction towards the "Background" class. Conversely, feature 5 with a value of 81.69 strongly supports the "Signal" class. This analysis highlights that the model's decisions are based on specific feature interactions, providing insights into its decision-making process.