Support Vector Machine Classifier on the HIGGS Dataset

1. Introduction

The Support Vector Machine (SVM) classifier was applied to a subset of the HIGGS dataset to explore its predictive capabilities. Using a random sample of 5,500 data points (0.005% of the dataset), the analysis focused on improving model accuracy through data preprocessing, feature engineering, and kernel optimization.

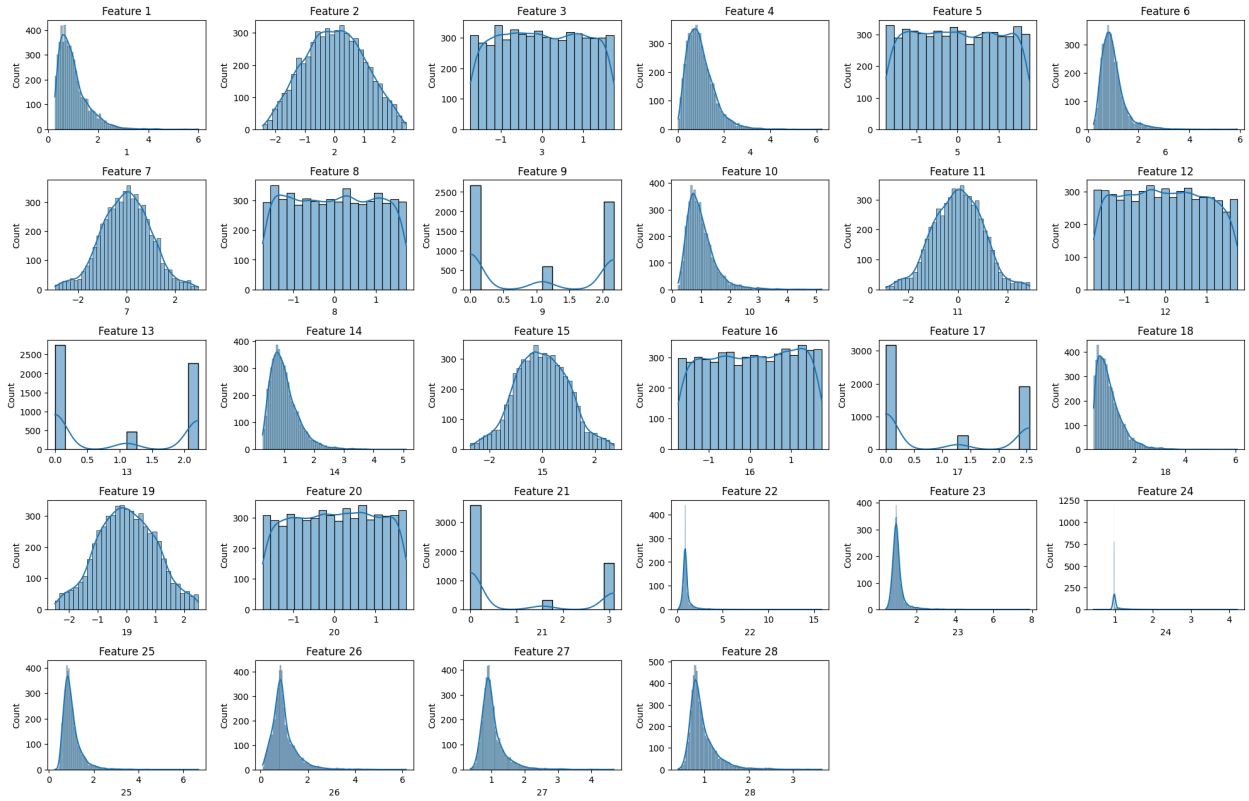
2. Libraries Used

- Data Manipulation and Exploration: pandas, numpy
- Visualization: seaborn, matplotlib
- Machine Learning: scikit-learn
- Model Interpretation: lime

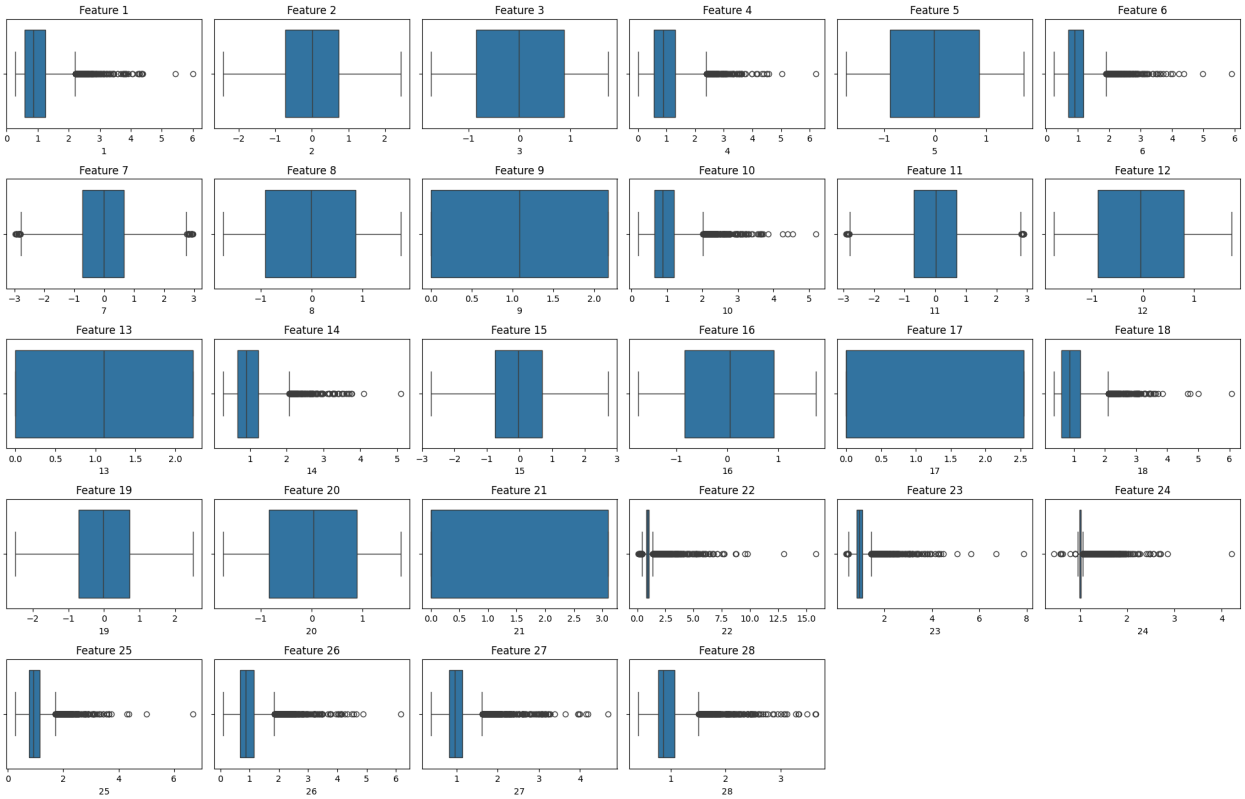
3. Data Preprocessing and Exploration

- Data Shape: After sampling, the dataset had a shape of **(5500, 29)**.
- Feature Distribution: Visualized using histograms and boxplots.

Feature Distribution using Histogram -



Outliers Distribution using Boxplot-



- Outliers Detection:

- Method: Z-score analysis identified **1,234** outliers.

- Feature Engineering:

- Original features: **(5500, 28)**.
- Polynomial features: Expanded to **(5500, 406)**.
- Selected 15 best features using **SelectKBest**, reducing to a shape of **(5500, 15)**.

4. Linear SVM Implementation

A linear SVM was initially implemented to benchmark performance before exploring more complex kernels.

- Linear SVM Results:

- Accuracy: 0.6285
- Precision: 0.6071
- Recall: 0.8209
- F1 Score: 0.6980
- AUC: 0.6805

- Stochastic Gradient Descent (SGD):
  - Utilized to handle large datasets and improve training speed.
  - **SGD SVM Results:**
    - Accuracy: 0.6160
    - Precision: 0.5921
    - Recall: 0.8557
    - F1 Score: 0.6997
    - AUC: 0.6703

5. SVM with Polynomial, RBF, and Custom Kernels

Three kernels were evaluated to determine the most effective for this dataset.

Polynomial Kernel

The polynomial kernel was tested with degrees 2, 3, and 4.

Degree	C	Training Time(s)	Accuracy	Precision	Recall	F1 Score	AUC
2	10	216.35	0.6375	0.6058	0.8783	0.7170	0.7104
3	10	166.90	0.6482	0.6062	0.9339	0.7352	0.7503
4	10	170.68	0.6325	0.5902	0.9725	0.7346	0.7617

RBF Kernel

Explored with various hyperparameters for optimal performance.

C	Gamma	Training Time(s)	Accuracy	Precision	Recall	F1 Score	AUC
0.1	0.01	15.19	0.5998	0.5744	0.9061	0.7031	0.6780
1	0.001	13.87	0.5982	0.5718	0.9218	0.7058	0.6715
10	1	17.89	0.9753	0.9613	0.9927	0.9767	0.9943

Custom Kernel: Sigmoid Kernel

Evaluated using a custom kernel implementation with various hyperparameters.

C	Gamma	Training Time(s)	Accuracy	Precision	Recall	F1 Score	AUC
1	0.001	13.65	0.5807	0.5584	0.9471	0.7026	0.6631
10	0.01	8.33	0.5896	0.6070	0.6106	0.6088	0.5853
10	1	9.81	0.5158	0.5370	0.5372	0.5371	0.5092

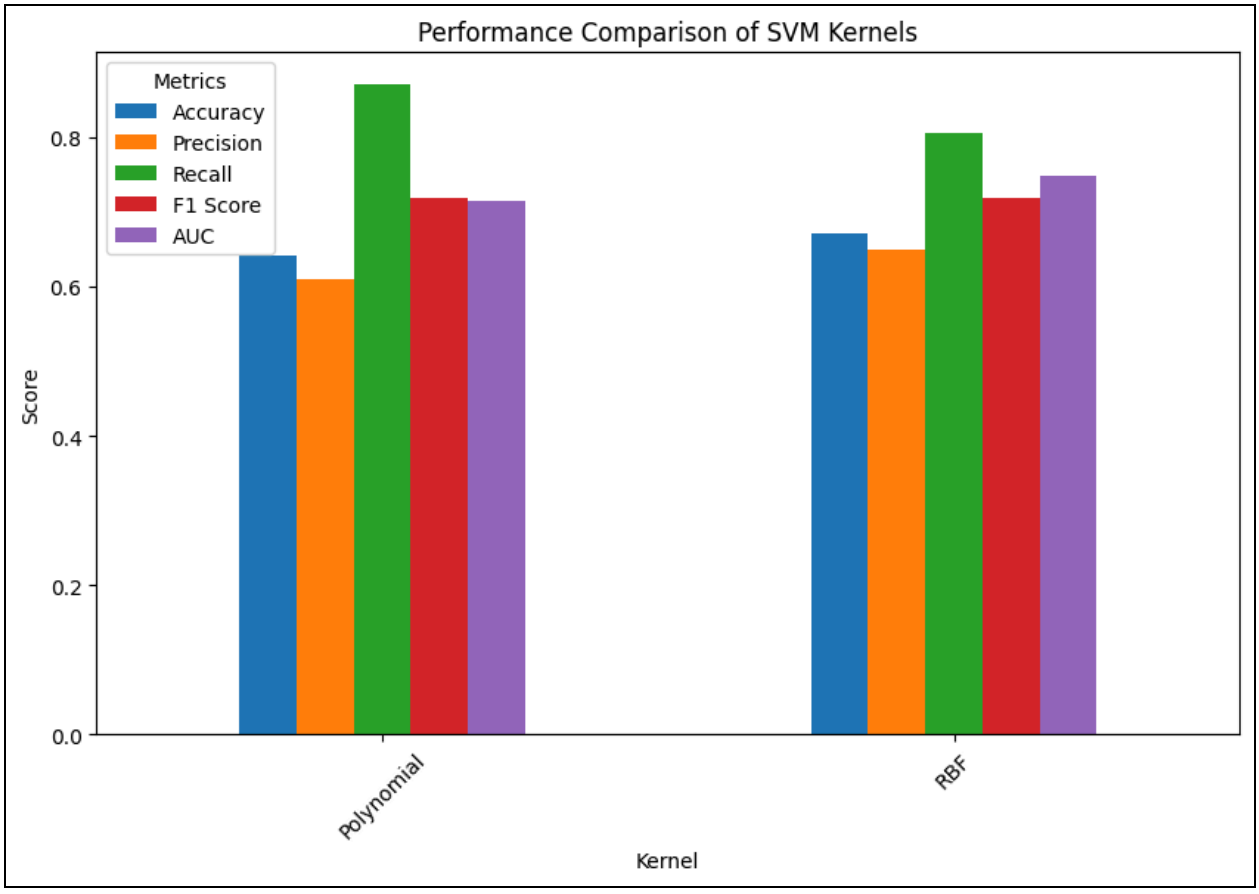
6. Hyperparameter Tuning

Utilized RandomizedSearchCV for hyperparameter optimization.

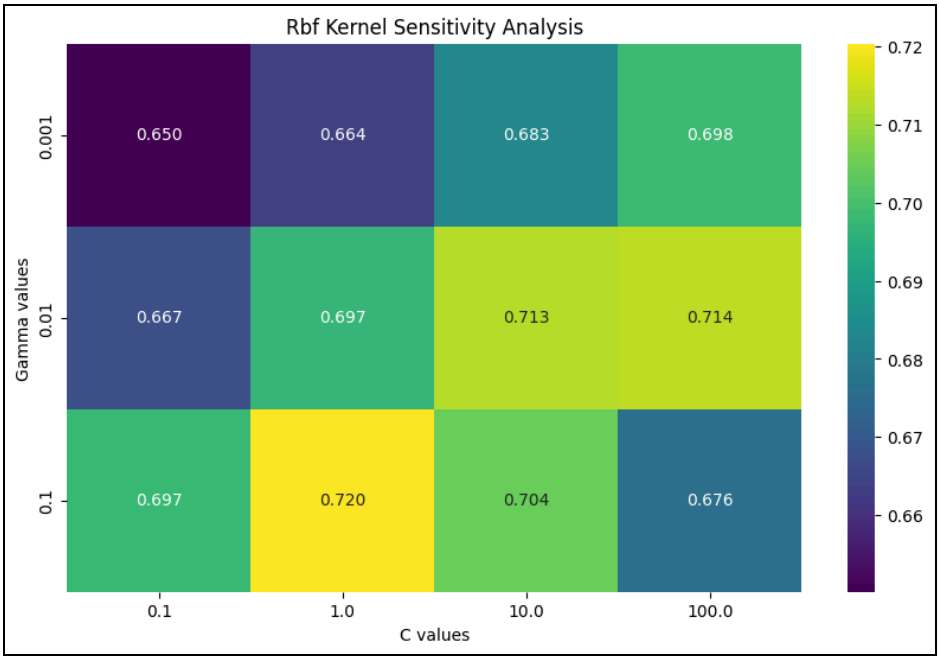
- Polynomial Kernel: Best parameters: {'gamma': 'auto', 'degree': 2, 'C': 10}
  - Best cross-validation AUC score: 0.6847
- RBF Kernel: Best parameters: {'gamma': 'scale', 'C': 1}
  - Best cross-validation AUC score: 0.7163

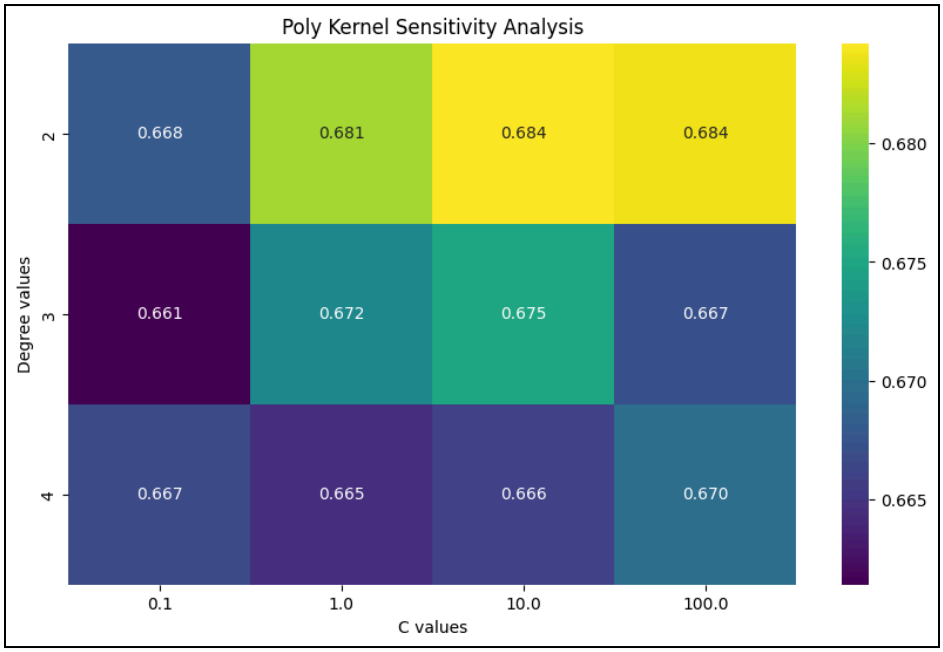
Results Summary

Kernel	C	Degree	Gamma	Accuracy	Precision	Recall	F1 Score	AUC
Polynomial	10	2	Auto	0.6422	0.6108	0.8703	0.7178	0.7147
RBF	1	-	Scale	0.6705	0.6488	0.8067	0.7192	0.7485


Performance Comparison of SVM Kernels
Metrics
Accuracy
Precision
Recall
F1 Score
AUC
Score
0.0
0.2
0.4
0.6
0.8
Polynomial
RBF
Kernel

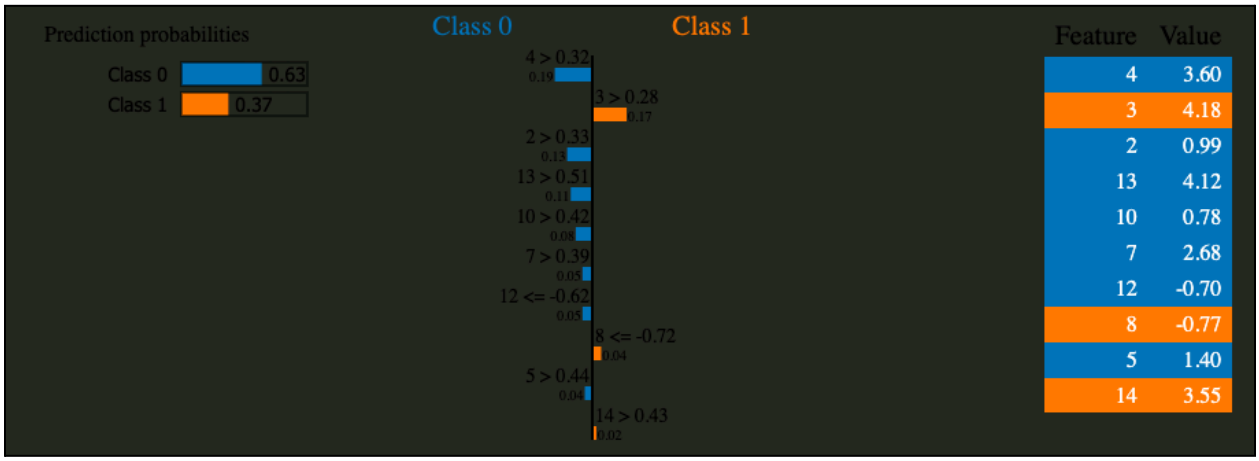
7. Hyperparameter Sensitivity Analysis

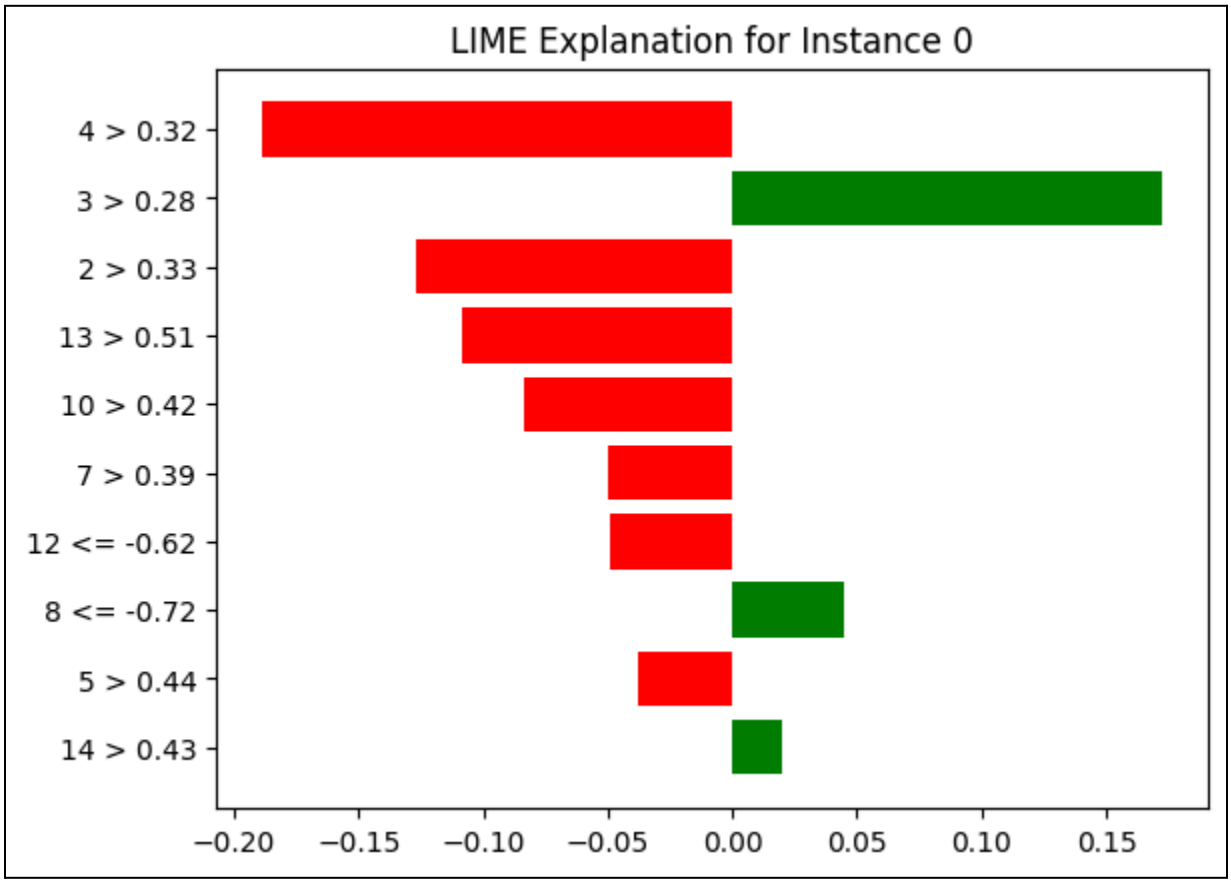




8. Explainability and Interpretability

Model interpretability was enhanced using lime to provide insights into feature importance and individual prediction influences. This explainability step assists in understanding how the SVM model decisions align with the dataset's intrinsic characteristics.





9. Conclusion

The RBF kernel with optimized parameters yielded the highest AUC and accuracy, suggesting it is most suitable for this dataset. Feature selection and kernel tuning proved vital, significantly improving model performance metrics across various SVM configurations.