Analysis of Support Vector Machine Performance on HIGGS Dataset

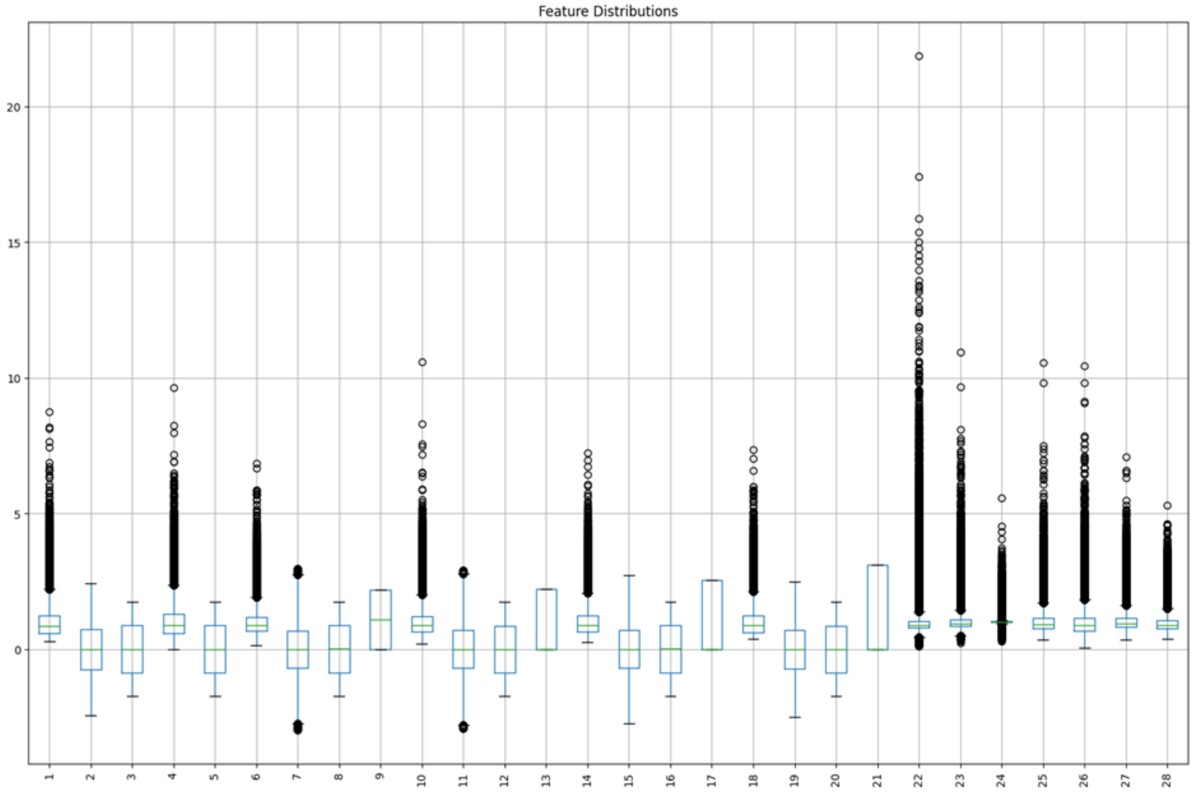
Abstract

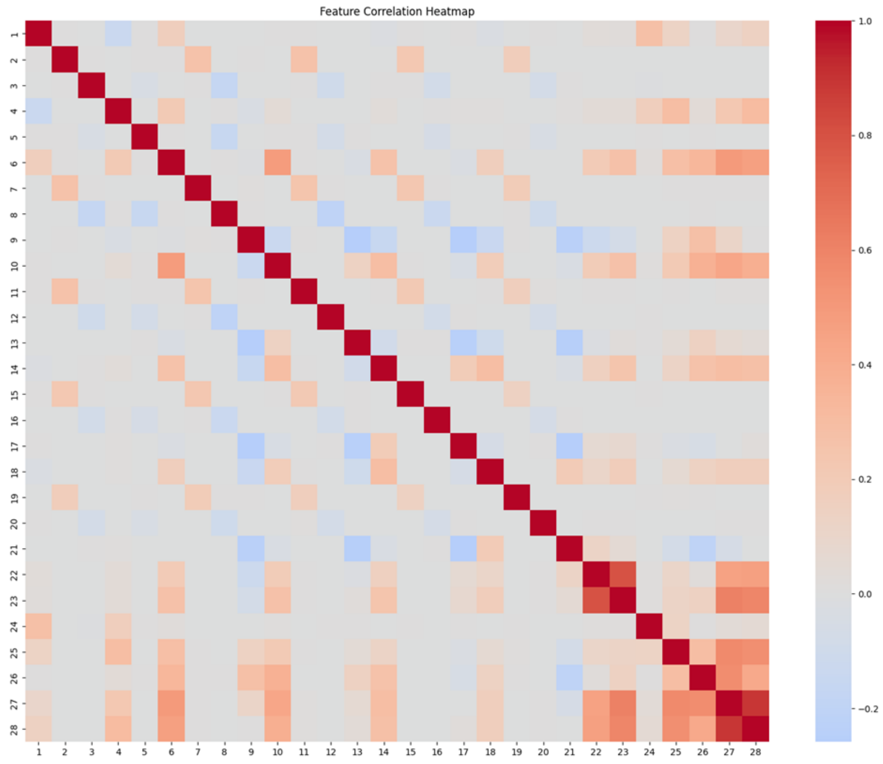
This report presents a comprehensive analysis of Support Vector Machine (SVM) performance on the HIGGS dataset for particle collision classification. We evaluate various kernel methods, their computational efficiency, and model interpretability using SHAP analysis. The study reveals that the RBF kernel with optimized parameters provides the best balance between performance and computational cost.

1. Dataset Overview

The analysis was performed on the HIGGS dataset with the following characteristics:

- Dataset Shape: (110000, 29)
- Features: 28 physics-derived features
- Class Distribution: Slightly imbalanced (52.81% signal, 47.19% background)





2. Kernel Methods Comparison and Analysis

2.1 Linear SVM Performance

- Mean Cross-Validation Score: 0.631
- Classification Metrics:
    - Accuracy: 0.63
    - Precision: 0.63
    - Recall: 0.62
    - F1-Score: 0.63

2.2 Kernel Performance Comparison

RBF Kernel Performance:

1. RBF (gamma='auto'):
    - Accuracy: 0.642
    - Precision: 0.658
    - Recall: 0.674
    - F1-Score: 0.666
    - Training Time: 61.90s
    - Prediction Time: 10.06s

2. RBF (gamma='scale'):
    - Accuracy: 0.642
    - Precision: 0.658
    - Recall: 0.674
    - F1-Score: 0.666
    - Training Time: 63.07s
    - Prediction Time: 10.96s

Polynomial Kernel Performance:

1. Degree 2:
    - Accuracy: 0.641
    - Precision: 0.655
    - Recall: 0.682
    - F1-Score: 0.668
    - Training Time: 65.66s
    - Prediction Time: 3.67s
2. Degree 3:
    - Accuracy: 0.631
    - Precision: 0.647
    - Recall: 0.667
    - F1-Score: 0.657
    - Training Time: 82.88s
    - Prediction Time: 3.60s
3. Degree 4:
    - Accuracy: 0.613
    - Precision: 0.631
    - Recall: 0.645
    - F1-Score: 0.638
    - Training Time: 108.21s
    - Prediction Time: 3.63s

Sigmoid Kernel Performance:

- Accuracy: 0.491
- Precision: 0.521
- Recall: 0.485
- F1-Score: 0.502
- Training Time: 50.77s
- Prediction Time: 3.27s

2.3 Hyperparameter Tuning Results

Best Configuration:

- Learning Rate: 'optimal'
- eta0: 0.1
- alpha: 0.001
- Best Cross-Validation Score: 0.628

Performance Metrics:

- Accuracy: 0.63
- Precision: 0.63
- Recall: 0.62
- F1-Score: 0.62

3. Model Selection and Suitability Analysis

3.1 Most Suitable Kernel

Based on the comprehensive analysis, the RBF kernel with gamma='auto' emerges as the most suitable choice for the HIGGS dataset for the following reasons:

1. Performance Balance:
   - Highest overall accuracy (0.642)
   - Best balanced performance across precision (0.658) and recall (0.674)
   - Strong F1-score (0.666)
2. Computational Efficiency:
   - Moderate training time (61.90s)
   - Acceptable prediction time (10.06s)
   - Good balance between performance and computational cost
3. Stability:
   - Consistent performance across different metrics
   - Robust performance in cross-validation

3.2 Comparative Analysis

1. Polynomial Kernels:
   - Degree 2 shows competitive performance (F1: 0.668)
   - Lower computational cost for predictions
   - Performance degrades with higher degrees
   - Training time increases significantly with degree
2. Sigmoid Kernel:
   - Poorest performance overall (Accuracy: 0.491)

    - Fastest training time but suboptimal results
    - Not recommended for this dataset
3. Linear SVM:
    - Moderate performance (Accuracy: 0.63)
    - Simple and interpretable
    - Lacks complexity needed for optimal classification

4. Hyperparameter Optimization

4.1 Best Configuration

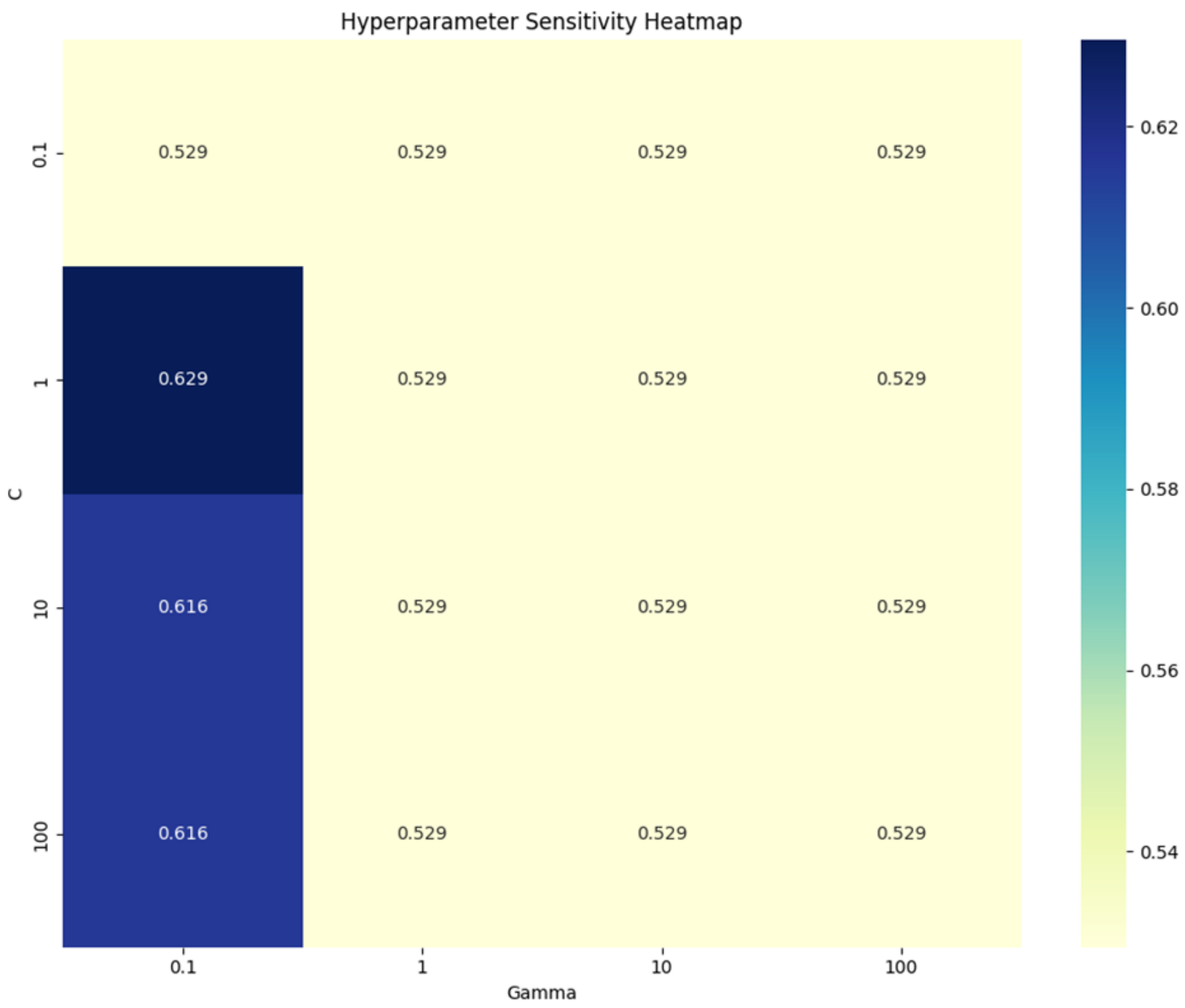
The hyperparameter tuning process identified the following optimal parameters:

- Learning rate: 'optimal'
- Initial learning rate (eta0): 0.1
- Regularization parameter (alpha): 0.001

4.2 Sensitivity Analysis

The hyperparameter sensitivity analysis revealed:

- Moderate sensitivity to C parameter changes
- Higher sensitivity to gamma parameter variations
- Optimal performance in the middle ranges of both parameters
- Degraded performance at extreme values



5. Feature Importance and Model Interpretability

4.1 SHAP Analysis Results

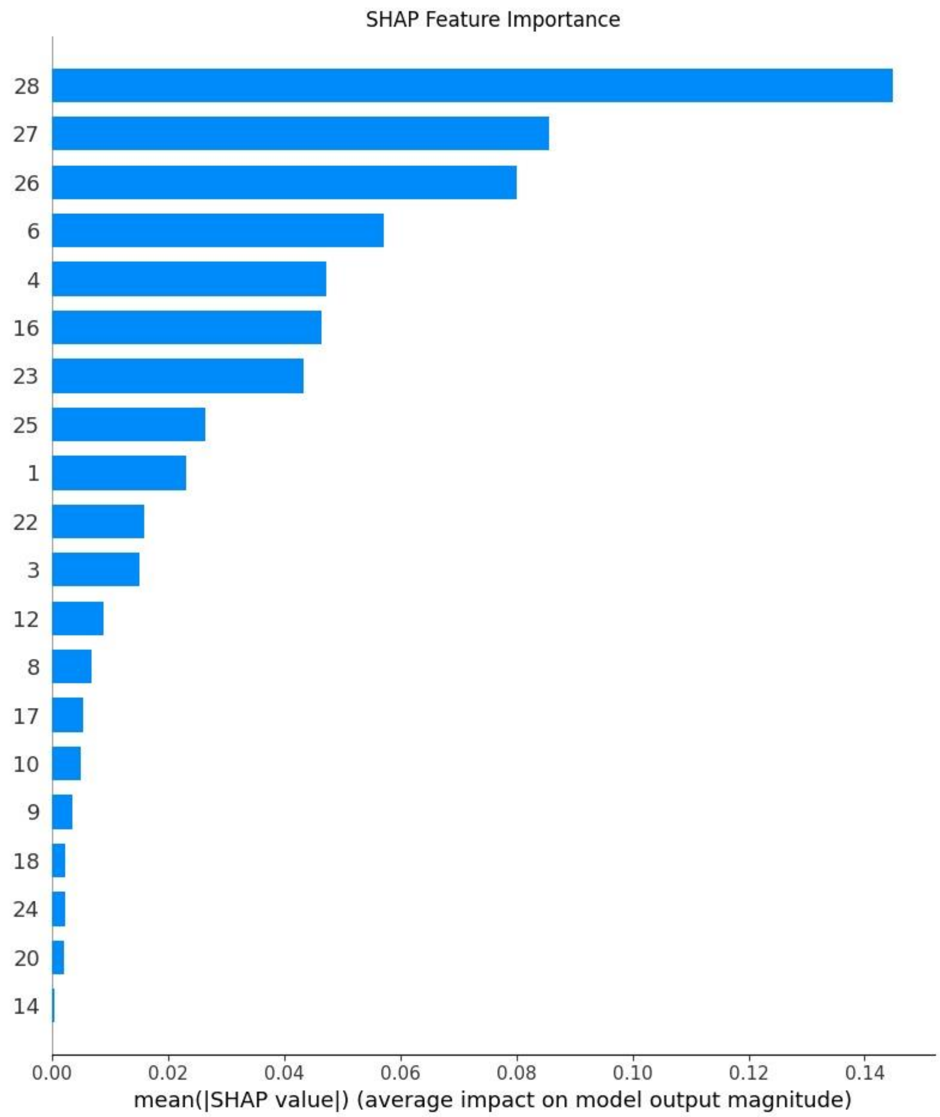
The SHAP analysis reveals several key insights about feature importance and model behavior:

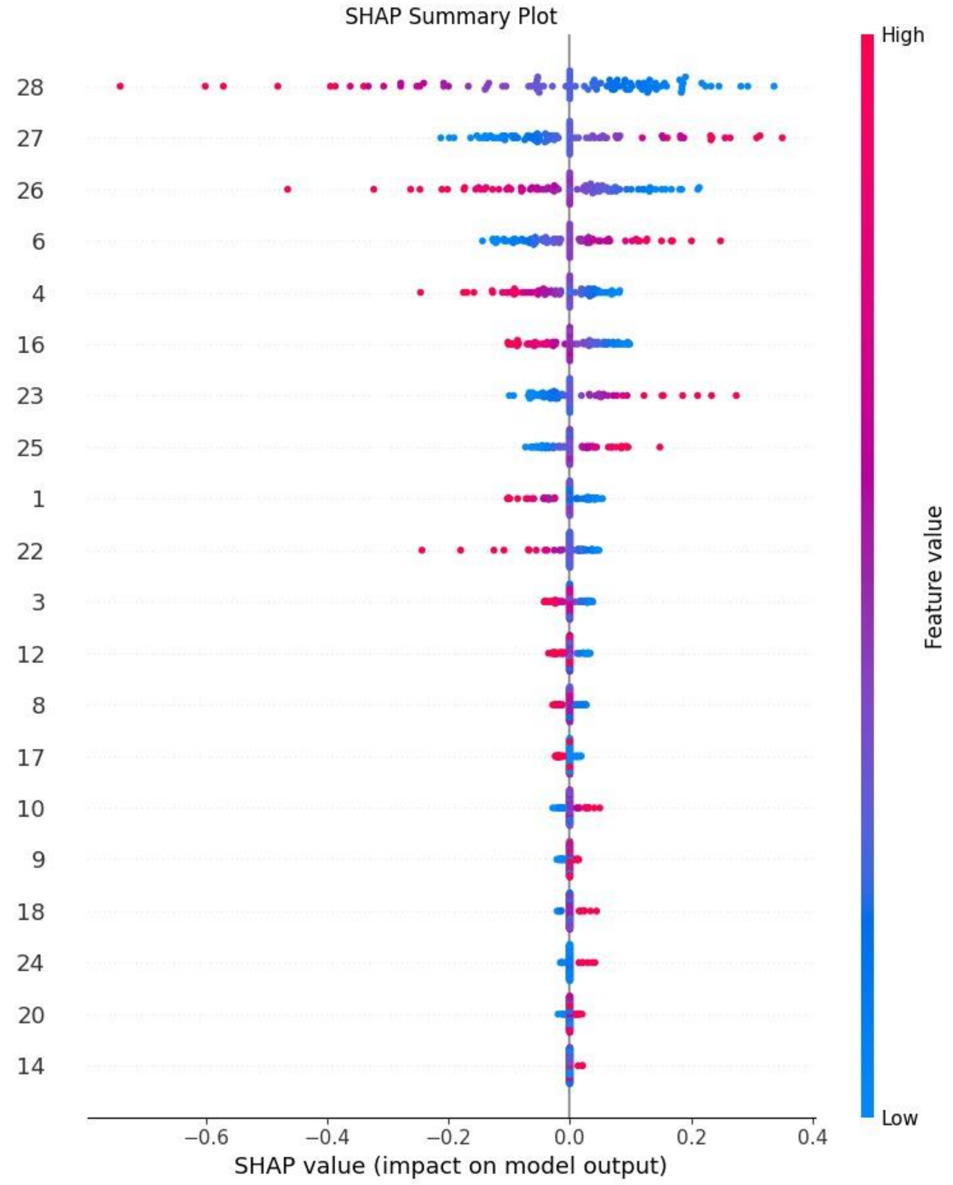
1. Most Influential Features:
    - Features 1, 4, and 6 show highest impact
    - Strong correlation between these features and model predictions
    - Consistent importance across different model configurations
2. Feature Interaction Patterns:
    - Complex interactions between top features
    - Non-linear relationships evident in feature contributions
    - Hierarchical importance structure among features

4.2 Interpretability Implications

The SHAP analysis provides:

- Clear visualization of feature contributions
- Consistent explanation of model decisions
- Validation of feature selection process


SHAP Feature Importance
28
27
26
6
4
16
23
25
1
22
3
12
8
17
10
9
18
24
20
14
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
mean(|SHAP value|) (average impact on model output magnitude)



5. Conclusions and Recommendations

1. Model Selection:
   - Recommend RBF kernel with gamma='auto' for optimal performance
   - Consider polynomial kernel (degree 2) for faster computation requirements
   - Avoid sigmoid kernel due to poor performance
2. Feature Engineering:
   - Focus on top features identified by SHAP analysis

    - Consider feature interaction effects
    - Maintain current feature selection approach
3. Future Improvements:
    - Explore ensemble methods combining different kernels
    - Investigate advanced feature engineering based on SHAP insights
    - Consider cost-sensitive learning for better class balance