Technical Research Report

Physics-Driven Machine Learning for Event Classification

Shaker Awni Shaker Alrawashdeh

February 2026

Abstract

This report presents a diagnostic, physics-driven machine learning study for event classification in a collider-like dataset. The analysis emphasizes interpretability, weighted evaluation, and physics-motivated operating points rather than black-box metric optimization. A linear baseline (logistic regression), an unsupervised variance analysis (PCA), and supervised classifiers (linear and RBF-kernel SVMs) are evaluated under a unified preprocessing and evaluation protocol. The results show that linear models capture most of the discriminative power, that variance-dominant directions are misaligned with class separation, and that physics-optimal thresholds can differ markedly from ROC-optimal choices due to event weights. These findings align with best practices in high-energy physics (HEP) analyses, where model transparency and sensitivity-driven decisions are essential.

Contents

1 Introduction

Machine learning has become a central tool in modern high-energy physics analyses, particularly in signal–background discrimination tasks involving high-dimensional reconstructed observables. This report presents a physics-driven machine learning study for event classification in a collider-like dataset. The central question is not simply how to maximize a generic classification metric, but rather how the structure of the physical observables constrains achievable performance and how model decisions translate to physics sensitivity.

The project is guided by three principles that are standard in HEP analyses:

- **Interpretability is critical.** A model should provide insight into which observables drive discrimination and how robust that discrimination is to variations in the data.
- **Weighted evaluation reflects physics reality.** Event weights encode expected yields or cross sections and must be included in training and performance metrics.
- **The final operating point is physics-motivated.** Global metrics (ROC-AUC, PR-AUC) are useful, but the final threshold should be selected to maximize a physics-relevant test statistic.

The workflow follows a structured pipeline: exploratory data analysis, a linear baseline (logistic regression), unsupervised variance analysis with PCA, and supervised classification with linear and kernel SVMs. The analysis emphasizes what these models reveal about the intrinsic information content of the observables rather than treating the modeling stage as a black-box optimization problem. In this sense, the project is a methodological study of how physics structure and statistical evaluation interact in real-world HEP classification tasks.

2 Dataset and Preprocessing

The dataset is stored in `Data - Events/Particle-Physics-Event-Classification.csv` and is accompanied by a descriptive README in `Data - Events/README.md`. The file contains 250,001 rows and 33 columns, including identifiers, a binary class label, and event weights. The class label in the CSV is `Label` with values `s` (signal) and `b` (background). The dataset documentation refers to the label as `Target`, so it is important to note that the actual column name in the file is `Label`. Event weights are stored in `Weight` and represent per-event importance for physical yield estimation.

Key dataset characteristics:

- Each row corresponds to a reconstructed event with engineered kinematic observables.
- Many jet-related features include a placeholder value of `-999` to represent undefined quantities in events with missing jets.
- The class distribution is not physically representative in raw counts; physical rarity is encoded by the weights rather than the sample size.

Preprocessing is centralized in `Preprocessing/data_preprocessing.py` to ensure consistent handling across notebooks. The steps are:

- Replace `-999` with `NaN` to mark missing values.
- Map labels to numerical values (`s` $\rightarrow$ `1`, `b` $\rightarrow$ `0`).
- Extract the feature matrix by dropping `EventId`, `Label`, and `Weight`.
- Apply median imputation to handle missing values.
- Standardize features using z-score scaling.

The data is split into training, validation, and test sets using stratified sampling to preserve class proportions. The default configuration uses a holdout split with the holdout further divided into validation and test subsets. This split design supports model selection on validation data and unbiased reporting on the test set, which is standard in physics workflows.

This preprocessing pipeline is applied consistently in both linear and SVM analyses. No feature engineering beyond standardized preprocessing was introduced, in order to isolate model-dependent effects. It is also used implicitly before PCA to ensure that variance structure is not dominated by differences in feature scales.

3 Linear Baseline: Logistic Regression

Logistic regression is used as the primary linear baseline, following the philosophy that interpretable models are essential early in the workflow. This model provides a transparent view of which standardized features contribute to discrimination and serves as a reference point for evaluating more complex classifiers. For input vector $\mathbf{x}$, the model estimates

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-\left(\mathbf{w}^\top \mathbf{x} + b\right)}}. \tag{1}$$

Since the inputs are standardized, the coefficients can be interpreted as relative contributions of each observable to the log-odds of signal.

3.1 ROC

The logistic regression notebook reports weighted ROC curves and an AUC value that is in the low to mid 0.8 range. The ROC curve exhibits strong separation compared to random guessing, indicating that a linear decision boundary captures substantial discriminative structure in the data. This result supports the hypothesis that many of the relevant correlations are approximately linear in the engineered feature space.

A key point is that ROC-AUC is threshold-independent and therefore useful for comparing models, but it does not define the optimal selection threshold for physics. This is explicitly emphasized in the notebook through later threshold scanning and significance analysis.

3.2 PR

Precision–Recall curves are included to contextualize the classifier in the presence of class imbalance. In HEP contexts, PR curves are informative because they reveal how quickly precision deteriorates as recall increases, especially when the signal fraction is small. The PR curve is therefore used to complement ROC-AUC rather than to replace it.

The analysis highlights that PR performance is strongly influenced by class imbalance and event weighting. A model can achieve a strong ROC-AUC while still yielding modest precision at high recall values. This motivates the later focus on physics-driven threshold selection rather than reliance on a single global metric.

3.3 Threshold optimization

The notebook performs threshold scanning on the logistic regression output to compute physics-motivated yields. The procedure computes weighted signal and background yields as a function of the decision threshold and then evaluates the significance metric:

$$\mathcal{Z} = \frac{S}{\sqrt{S+B}}. \tag{2}$$

This analysis reveals a nontrivial dependence of physics significance on the threshold. In particular, the best threshold in the logistic regression study is near the lower end of the probability scale. The reason is physically meaningful: a strict threshold can reject a small number of high-weight signal events, which disproportionately harms the weighted signal yield even if the overall classification performance remains good.

This result reinforces a key lesson in HEP: a high ROC-AUC does not guarantee optimal physics sensitivity. The operating point must be chosen based on weighted yields, not purely on classification accuracy or ROC performance.

4 Variance Structure Analysis via PCA

PCA is used to understand the variance structure of the feature space and to diagnose whether dominant directions of variation correspond to discriminative directions. PCA is applied after imputation and scaling. Because PCA is unsupervised, the analysis uses the full dataset rather than restricting to a training split.

4.1 Explained variance

The explained variance spectrum shows that a relatively small number of principal components capture a large fraction of total variance. This indicates significant correlations among the input observables, which is consistent with the physics of reconstructed objects and derived kinematic variables.

However, high explained variance alone does not imply good class separation. The analysis emphasizes that PCA maximizes variance, not discriminative power. Therefore, the explained variance curve is used for representation insight, not for defining a reduced feature set for classification.

4.2 2D / 3D projections

The notebook provides both 2D and 3D PCA projections, with points colored by class label. The PC1–PC2 projection exhibits substantial overlap between signal and background. The 3D projection similarly shows overlap, with any apparent ordering being view-dependent and not robust under rotation.

An additional projection using PC2–PC4 (excluding the leading component) yields even stronger overlap, suggesting that the dominant variance direction does not carry

useful class separation. This is consistent with the unsupervised nature of PCA and the expectation that many high-variance directions represent shared physics structure rather than signal-specific features.

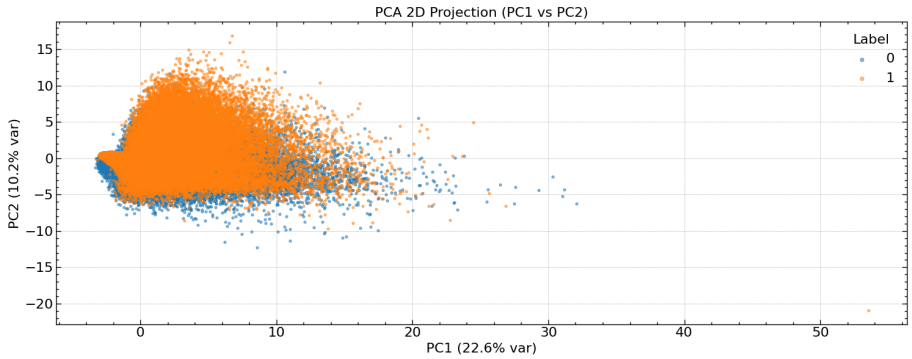

Figure 1: Two-dimensional PCA projection (PC1 vs PC2) of the dataset, with points colored by class label. Despite capturing the largest fractions of total variance, the leading principal components do not exhibit clear signal–background separation, indicating that variance-dominant directions are misaligned with discriminative structure.

4.3 Misalignment with class separation

The core conclusion from PCA is that variance-dominant directions are not aligned with the class separation direction. This provides empirical justification for focusing on supervised classifiers and for not applying PCA as a dimensionality reduction step before classification in this project. The PCA analysis serves as a diagnostic that motivates why linear or moderately nonlinear supervised models should be used to uncover discriminative structure that is not captured by variance alone.

5 Support Vector Machines

SVMs are introduced to test whether a margin-based classifier and nonlinear decision boundaries can extract additional discriminative information beyond the linear baseline. The analysis uses the same preprocessing pipeline and weighted evaluation metrics to ensure comparability.

5.1 Linear SVM

A linear SVM is trained with a grid search over the regularization parameter C. Model selection is performed on the validation set using weighted ROC-AUC. The final linear model is trained on the combined training and validation data and evaluated on the test set. The linear SVM yields performance similar to logistic regression and serves as a strong, stable baseline with an interpretable linear decision function.

The linear decision function provides a direct mapping from standardized features to a separating hyperplane, which supports physics interpretability and helps identify which observables contribute most strongly to discrimination.

5.2 RBF Kernel

A nonlinear SVM with an RBF kernel is used to test whether weak nonlinear correlations improve classification. Because kernel SVMs scale poorly with dataset size, the hyperparameter scan is intentionally limited to a small grid. Weighted ROC and PR curves are computed for the nonlinear model and compared to the linear baseline.

The RBF model provides a modest improvement in ROC-AUC relative to the linear model. The improvement is consistently described in the notebook as incremental rather than transformative, suggesting that only weak nonlinear structure remains after the engineered features are considered.

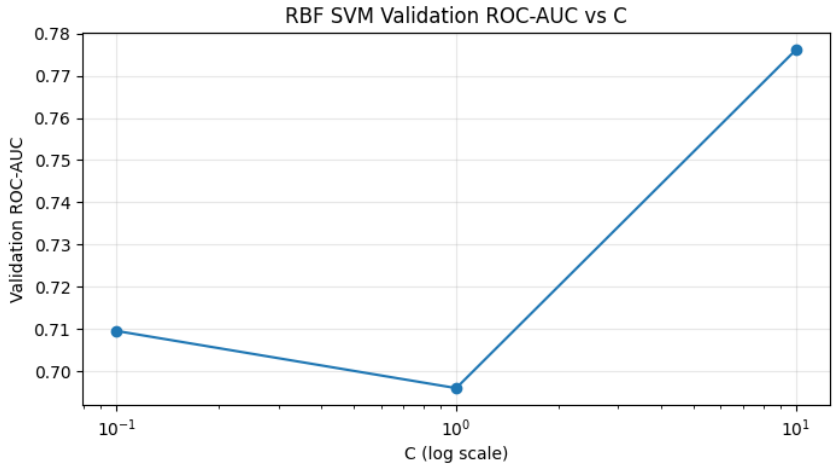

Figure 2: Validation ROC-AUC as a function of the regularization parameter C for the RBF SVM. Model selection is performed on the validation set using weighted ROC-AUC. The modest variation across C values indicates limited sensitivity to hyperparameter tuning.

5.3 Computational cost

The notebook explicitly discusses computational scaling. While linear SVMs train quickly, kernel SVMs are significantly more expensive. A single RBF configuration is reported to take on the order of tens of minutes (approximately 23 minutes) on the full dataset. This cost makes extensive hyperparameter optimization impractical and reinforces the trade-off between model complexity and feasibility in large HEP datasets.

5.4 Performance comparison

The performance comparison emphasizes that:

- Linear models already capture most of the discriminative power.
- The RBF kernel provides a measurable but modest gain.
- Improvements in ROC-AUC and PR-AUC do not necessarily translate to significantly better physics significance.

This comparison supports the conclusion that the dominant discriminative structure in the dataset is approximately linear, and that nonlinear models should be adopted only when their added complexity is justified by physics gains.

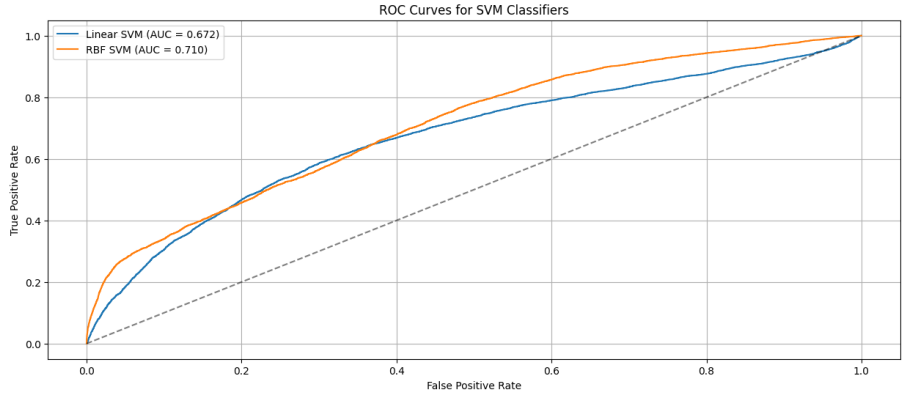

Figure 3: Weighted ROC curves for linear and RBF SVM classifiers evaluated on the independent test set. The RBF kernel provides a modest improvement in ROC-AUC, indicating the presence of weak nonlinear correlations but confirming that most discriminative power is already captured by linear decision boundaries.

6 Physics-Motivated Optimization

A central theme of the project is that the final decision threshold should maximize physics sensitivity rather than generic ML metrics. The notebooks implement threshold scanning and significance evaluation for both linear models and SVMs.

6.1 Significance

The physics-motivated metric used in the study is the approximate significance:

$$\mathcal{Z} = \frac{S}{\sqrt{S+B}}. \tag{3}$$

Here S and B are the weighted signal and background yields after applying a classification threshold. This statistic captures the balance between signal efficiency and background contamination and is commonly used in HEP for discovery sensitivity.

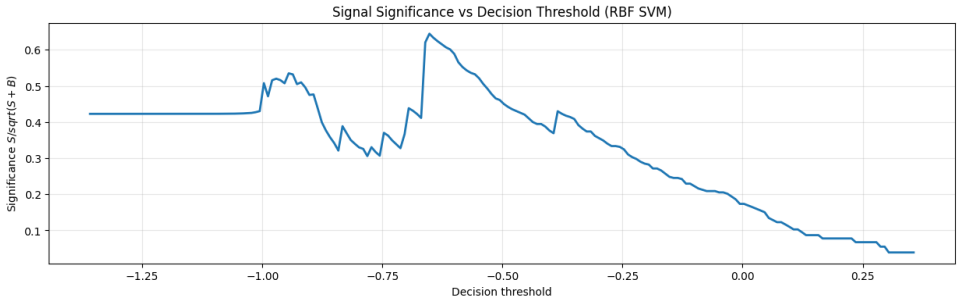

Figure 4: Signal significance $S/\sqrt{S+B}$ as a function of the RBF SVM decision threshold on the test set. The curve exhibits a clear maximum at an intermediate threshold, demonstrating that the physics-optimal operating point does not necessarily coincide with ROC-optimal or probability-based thresholds.

6.2 Threshold scan

Threshold scanning is performed by sweeping the classifier output across a grid of thresholds. At each threshold, the weighted sums of signal and background events are computed. The significance curve typically exhibits a clear maximum, demonstrating that the best physics operating point is not necessarily aligned with the threshold that optimizes ROC or PR metrics.

The logistic regression analysis reveals a strong dependence on threshold, with the optimal point occurring at a low threshold due to the presence of high-weight signal events that are otherwise rejected. The RBF SVM analysis shows an optimal threshold in an intermediate range of the decision score, illustrating that the scale and calibration of the classifier output affect the location of the physics-optimal point.

6.3 Interpretation

The threshold scan results demonstrate that physics sensitivity is dominated by how the classifier interacts with event weights. A classifier can look strong in ROC space but still perform poorly for discovery if it rejects a small number of high-weight signal events. The study therefore treats threshold optimization as a first-class component of the analysis rather than a post-hoc detail.

This approach mirrors real HEP workflows, where classifier outputs are used as continuous discriminants and operating points are chosen based on expected sensitivity rather than on purely ML metrics.

7 Discussion

This section synthesizes the results and frames them in the context of physics-driven modeling.

7.1 Why nonlinear gain is modest

The modest gain from the RBF kernel suggests that the engineered features already encode the dominant discriminative structure in approximately linear combinations. Many features in the dataset represent derived kinematic quantities, which often linearize physical relationships. In such cases, nonlinear kernels provide limited additional information and can only capture weak higher-order correlations.

Another factor is the presence of missing values and imputation. While imputation makes the dataset usable for standard ML pipelines, it can also smooth subtle nonlinear relationships, reducing the benefit of complex models. In this context, the linear decision boundary is a natural baseline and the nonlinear improvements are expected to be incremental.

7.2 Interpretability vs complexity

The project consistently highlights the trade-off between interpretability and complexity. Logistic regression and linear SVMs provide coefficient-based insights that are valuable in physics analyses, where the aim is to understand which observables drive signal–background separation. Nonlinear kernels sacrifice interpretability and incur heavy com-

putational cost. Given the limited performance gain, the project concludes that linear models are a strong and defensible choice for physics-driven studies.

This balance is not only a technical preference but also a methodological stance: interpretable models facilitate validation, systematic studies, and communication of results to a broader physics audience.

7.3 Implications for HEP analyses

Three implications stand out:

- **Weighted evaluation is essential.** Standard ML metrics can be misleading if weights are ignored or underemphasized.
- **The physics-optimal threshold can differ markedly from ROC-optimal thresholds.** Threshold selection must be tied to expected yields and significance.
- **Simple models can be competitive.** In many physics datasets, linear baselines capture most of the signal–background separation, making them robust, interpretable, and computationally efficient.

These conclusions align well with contemporary HEP best practices, where machine learning is used to augment, not replace, physics reasoning.

8 Conclusion

This project provides a rigorous, physics-motivated analysis of event classification using logistic regression, PCA, and SVMs. The study demonstrates that:

- The dataset contains meaningful discriminative structure, captured effectively by linear models.
- PCA reveals strong feature correlations but shows that variance-dominant directions do not align with class separation.
- Nonlinear SVMs yield only modest gains while incurring substantial computational cost.
- Physics-driven threshold optimization is critical and can alter the interpretation of model performance.

Overall, the work reflects strong methodological discipline and a clear understanding of how machine learning should be integrated into HEP analyses. The report emphasizes interpretability, robust evaluation, and physics-aware decision making, all of which are key qualities for graduate-level research in particle physics and machine learning.

The results illustrate that the integration of machine learning into HEP workflows must remain guided by physical interpretability, computational feasibility, and sensitivity-based decision criteria rather than by metric optimization alone.

This study is therefore best understood not as a benchmarking exercise, but as a structured methodological analysis of how modeling assumptions interact with the statistical and physical structure of collider observables.