Analysis of CMS Dataset to Identify Higgs Particle Decay

Name: Haroon Parvez

Student ID: 220475901

Department of Physics and Astronomy, School of Physical and Chemical Sciences, Queen Mary University of London, London, E1 4NS.

1 Introduction

This report summarises the analysis of a CMS dataset containing variables used to classify events as either background or signal (such as Higgs particle decays to bottom quarks). The analysis was conducted using Principal Component Analysis (PCA), three classification algorithms (Naive Bayes, Linear Discriminant Analysis, and Logistic Regression), and a custom threshold optimisation technique (Punzi significance). The aim was to determine which methods and input representations best separated signal from background events.

2 Data Quality and Characteristics

The dataset consists of 225,766 events with 28 columns. These consisted of 26 numerical predictor variables and two label columns: *isSignal* and *isBackground*. Through initial checks, a small number of fill values were present:

- Values of -1 appeared in two columns (*tau_vertexEnergyRatio_0* and *tau_vertexEnergyRatio_1*) and were replaced with NaN then dropped (resulting in the loss of 104 rows of data, this will not have any noticeable effect on analysis).
- Some columns had values of 0, but only the isSignal and isBackground labels had an excessive number of 0s (due to labelling). This did not need to be removed as these are not fill values.
- No obvious missing or corrupted data beyond the fill values. The format was clean and well-structured.

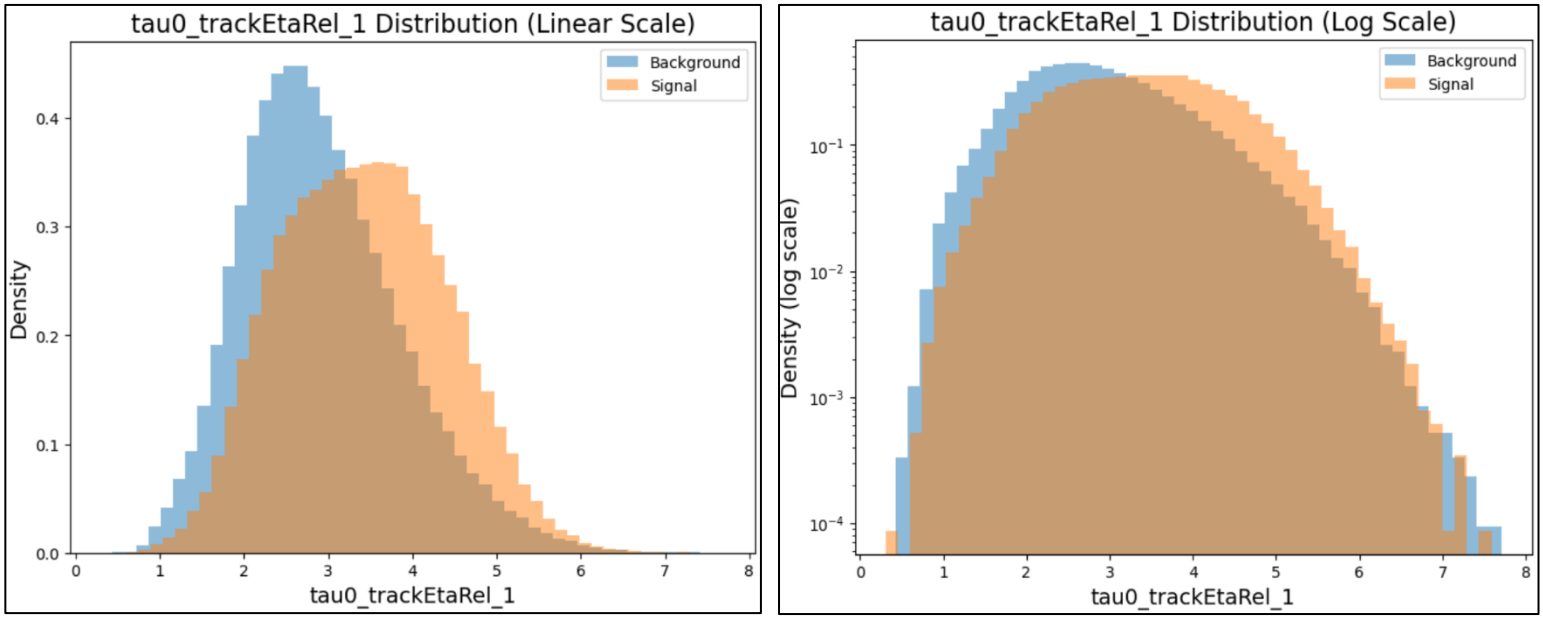
3 Feature Visualisation

Visual analysis included the following which will be talked about in detail:

- Histograms of all features split by event type, in both linear and logarithmic y-scales.
- Correlation heatmap between all numerical variables.

- 2D histograms for the most strongly correlated and anti-correlated variable pairs.
- Mean and standard deviation comparison across all features.

To characterise the data, histograms of all features were plotted for signal and background events. These plots were generated using both linear and logarithmic scales to capture differences across a wide range of values. Figure 1a shows the linear-scale distribution for a *tau0_trackEtaRel_1*, which highlights clear separation between signal and background. The two types of events have a clear distribution where the height and general shape is easy to interpret. However, the data set contained many small numbers and outliers, which were better visualised in the log-scaled histograms (see figure 1b for the logarithmic histogram of the same feature as figure 1a). These visualisations confirm that certain features, particularly those involving impact parameters and vertex kinematics, show strong discriminatory power. The logarithmic distribution shows both types of events being less separated, with a much bulkier distribution. Outliers from both background and signal can be seen, particularly for some background events on the right side.



(a) (b)

Figure 1: (a) Linear scale histogram of *tau0_trackEtaRel_1* for signal and background events; (b) Logarithmic scale histogram of the same feature.

A correlation heatmap (Figure 2) was then used to investigate linear dependencies between variables. Several strong correlations were identified, suggesting redundancy. For example, *trackSip2dSigAboveBottom_0* and *trackSip2dSigAboveBottom _1* were very strongly positively correlated. Figure 3 explores this further through a 2D histogram, which clearly shows the linear joint distribution, which is evidence of high correlation. Features such as tau0_trackEtaRel_1 and tau_vertexDeltaR_0 showed a strong negative correlation (~-0.75).

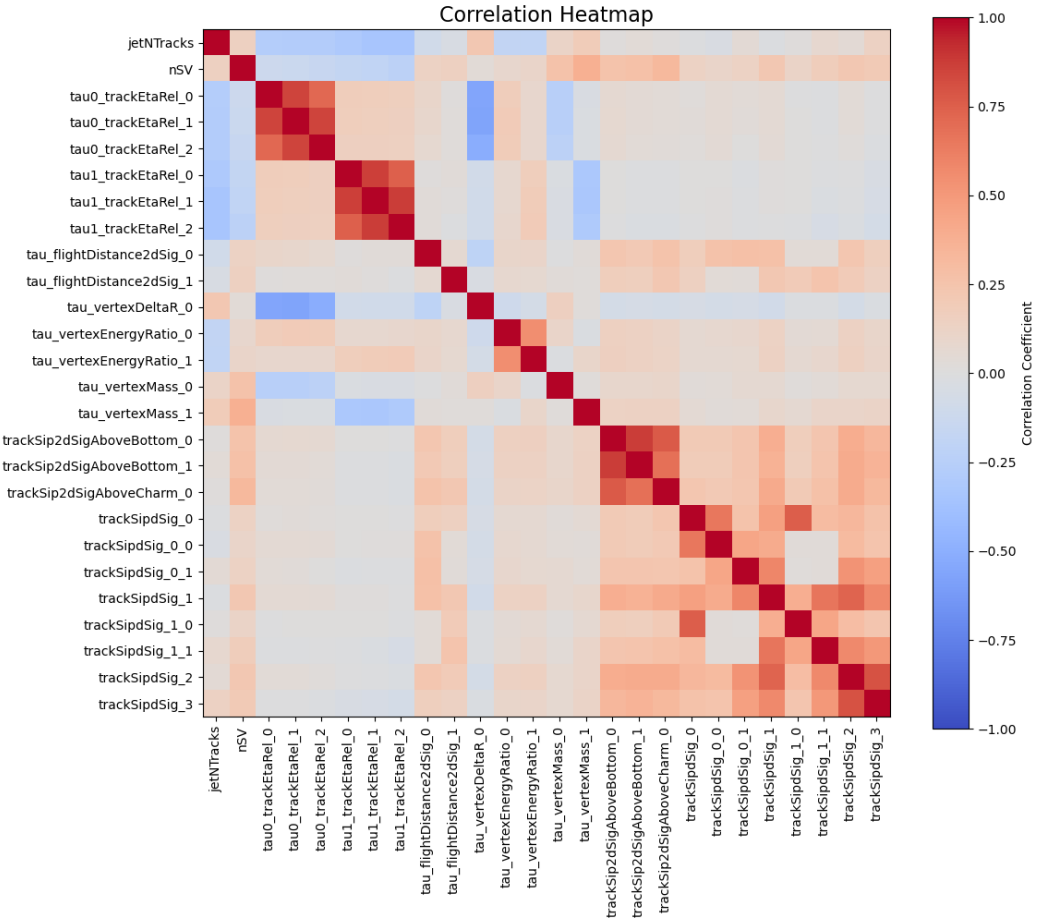


Figure 2: Correlation heat map of variables in the CMS dataset

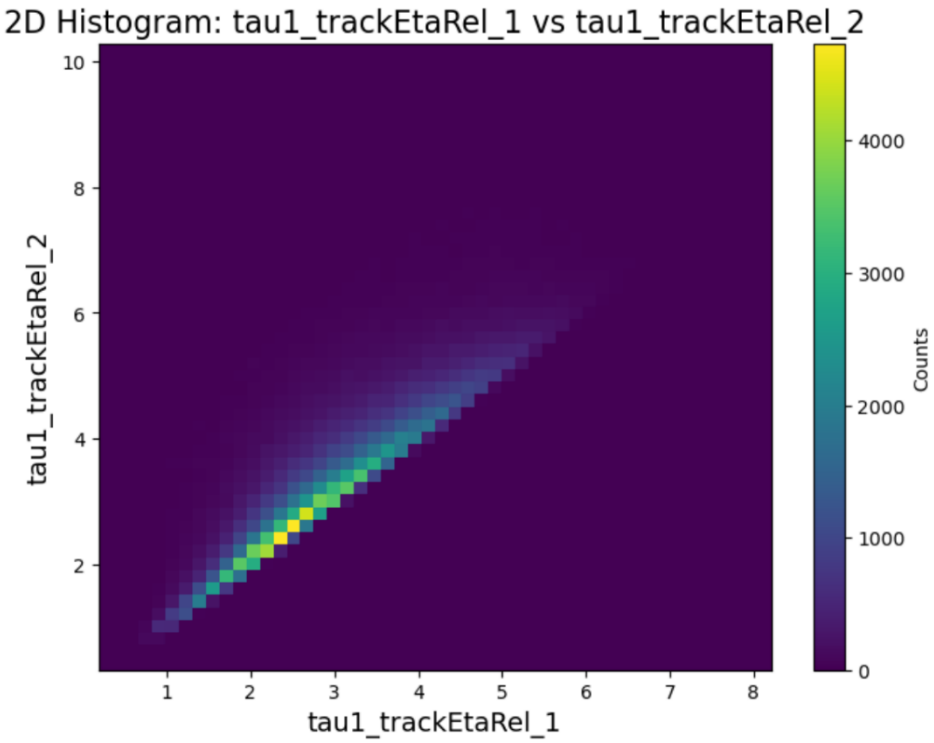


Figure 3: 2D Histogram of the most positively correlated features.

The mean and standard deviation for each numerical feature was calculated and is shown in Figure 4. This grouped bar chart reveals which features have the widest spread or largest magnitudes. Such information is useful for interpreting scale-sensitive methods and supports the decision to standardise features before applying PCA or classification. Notably, some features exhibited large standard deviations relative to their means, suggesting high variability across events. Features with both high mean and high variance may dominate learning algorithms if left unstandardised. Meanwhile, low-variance features may carry less discriminative information.

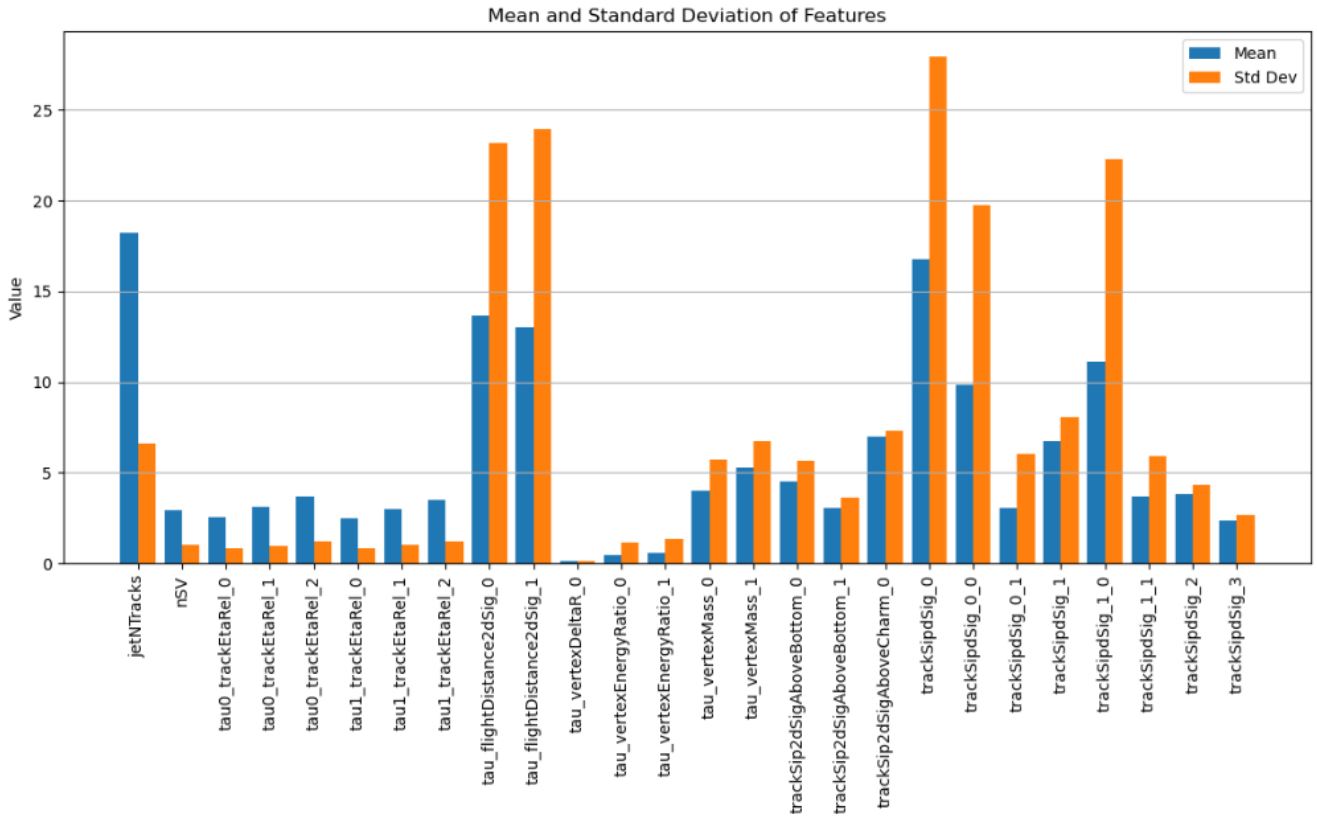


Figure 4: Bar plot of means and standard deviations of all variables.

3 Principal Component Analysis (PCA)

PCA was first applied to the full set of numerical features (excluding labels and index columns), using both signal and background events together. This was done without standardisation, and is intended only for exploratory purposes, not for classification. PCA is a widely used technique for reducing the dimensionality of datasets, retaining the most informative directions of variance. Its foundation is well-documented (Jolliffe, 2002) [2], and it is particularly effective in identifying structure in high-dimensional physics data.

Figure 5 shows the scree plot of explained variance (component weight loadings), this plot shows that the track impact parameters have a strong role in distinguishing between different events in the background sample, while the track pseudorapidity variables associated with the second N-subjettiness axis are less important. Several features contribute less than 0 to the variance in background and signal events. This suggests that some variables may need to be removed for a more accurate pool of predictors for classification.

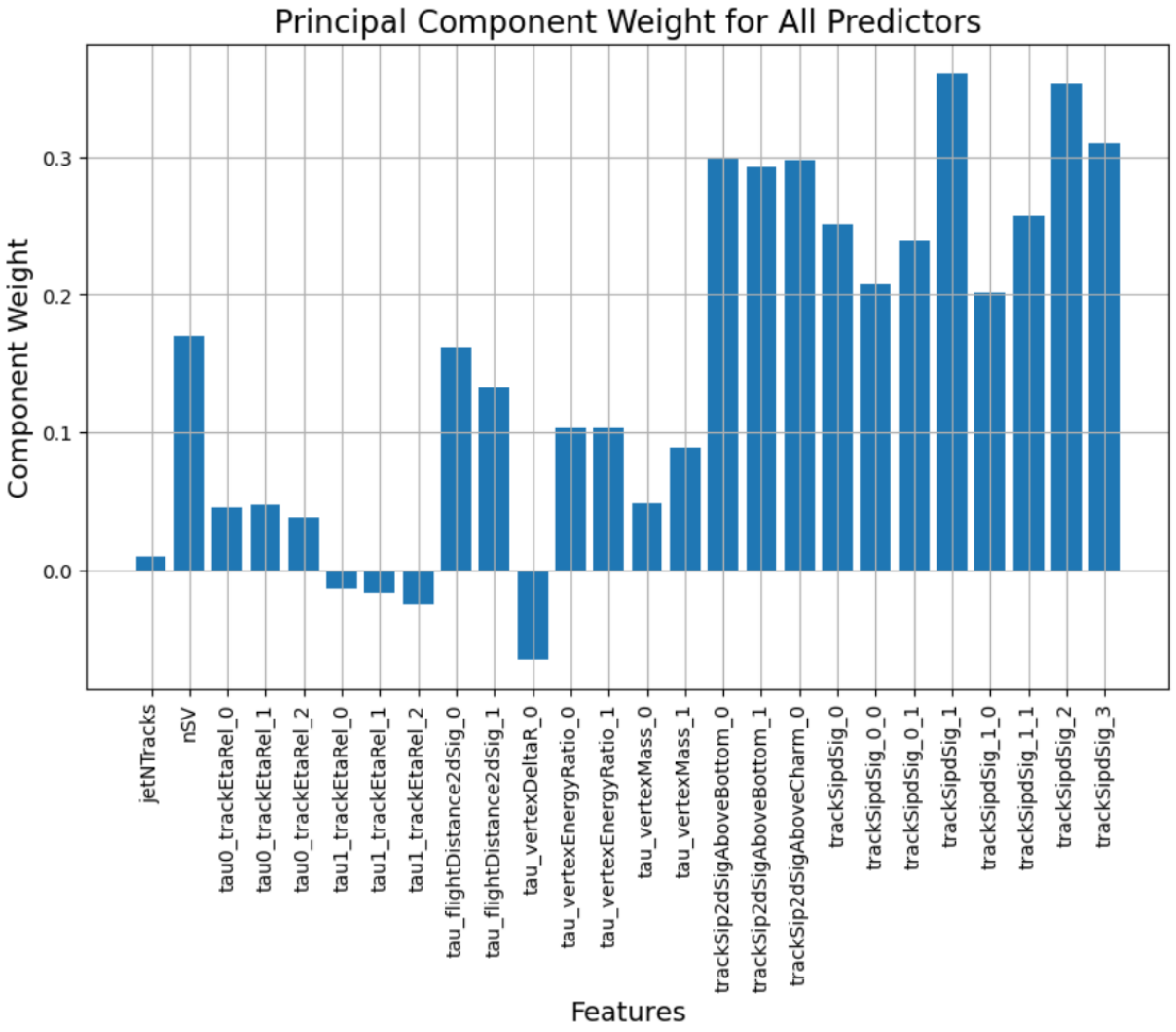


Figure 5: Principal Component Weights for all variables (predictors), showing how much each variable contributes to the variance in the two types of events (background and

Based on this, we considered two cases for classification:

- Case 1 uses all original numerical features without applying PCA. This serves as the baseline to evaluate performance without dimensionality reduction.
- Case 2 uses only the top 9 principal components, which were chosen based on the 85% cumulative variance threshold observed in the scree plot.

These two cases were evaluated to assess whether reduction in variables, those contributing around 15% in total, could improve or simplify classification without compromising performance.

4 Classification Models and Case Performance

Three classification models were trained and tested on both full feature data (Case 1) and PCA-reduced data (Case 2). These were implemented using the scikit-learn Python library (Pedregosa et al., 2011) [3], which provides robust and efficient tools for statistical learning. Figure 7 presents the ROC curves for all models in Case 1. ROC analysis is a standard approach for evaluating binary classifiers (Fawcett, 2006) [4], and here, Logistic Regression achieved the highest AUC (0.86), followed closely by LDA (0.84), and then Naive Bayes (0.82). The ROC curve for Logistic Regression is notably steeper, especially at low false positive rates, indicating superior discriminative performance.

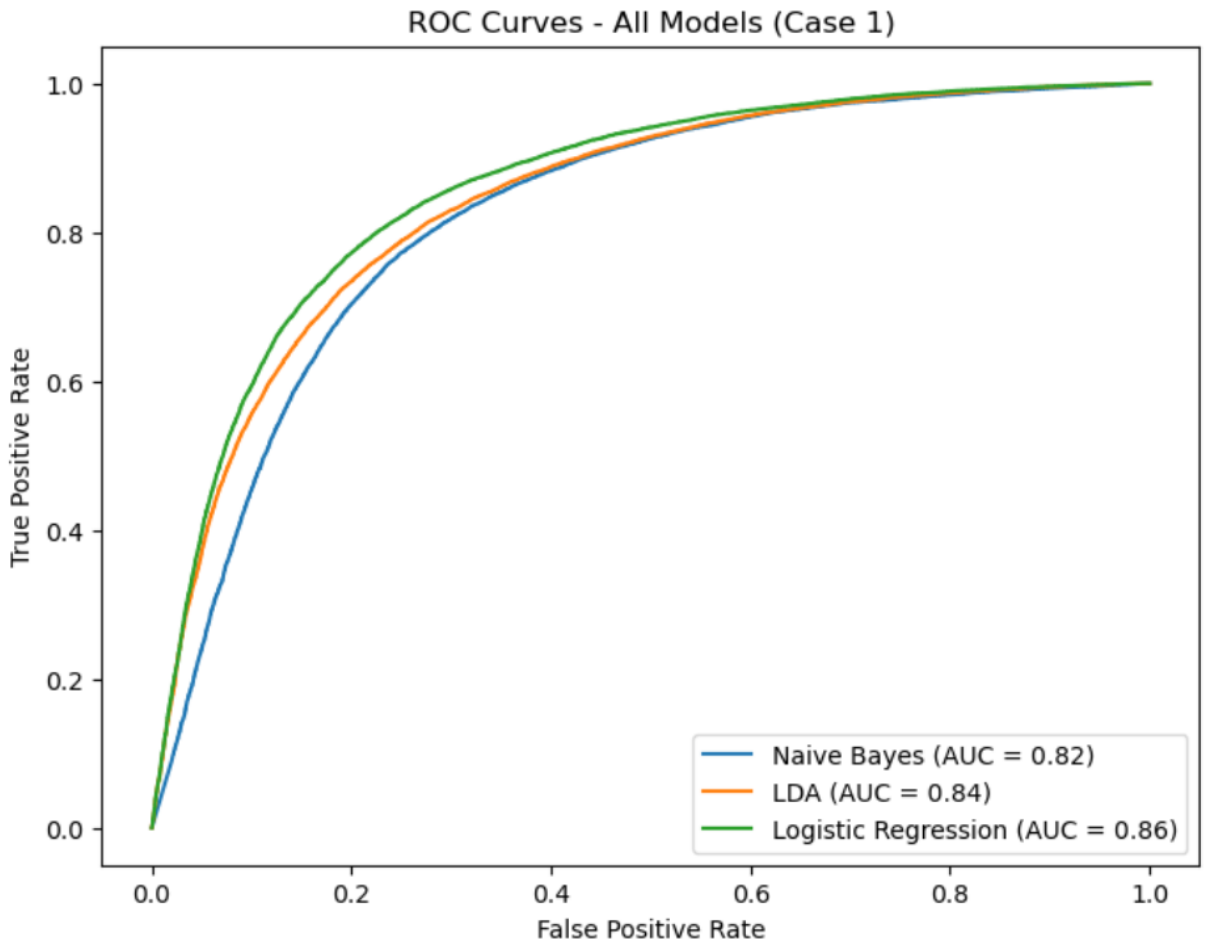


Figure 7: ROC Curves of all three classification models for case 1 (all variables).

To further examine how each model differentiated signal from background, histograms of predicted probabilities were plotted. Figure 8 shows the results for Logistic Regression (Case 1), where signal and background distributions are well-separated. The background peaks near zero while the signal distribution peaks closer to one, suggesting this model assigns high confidence scores to correct classifications.

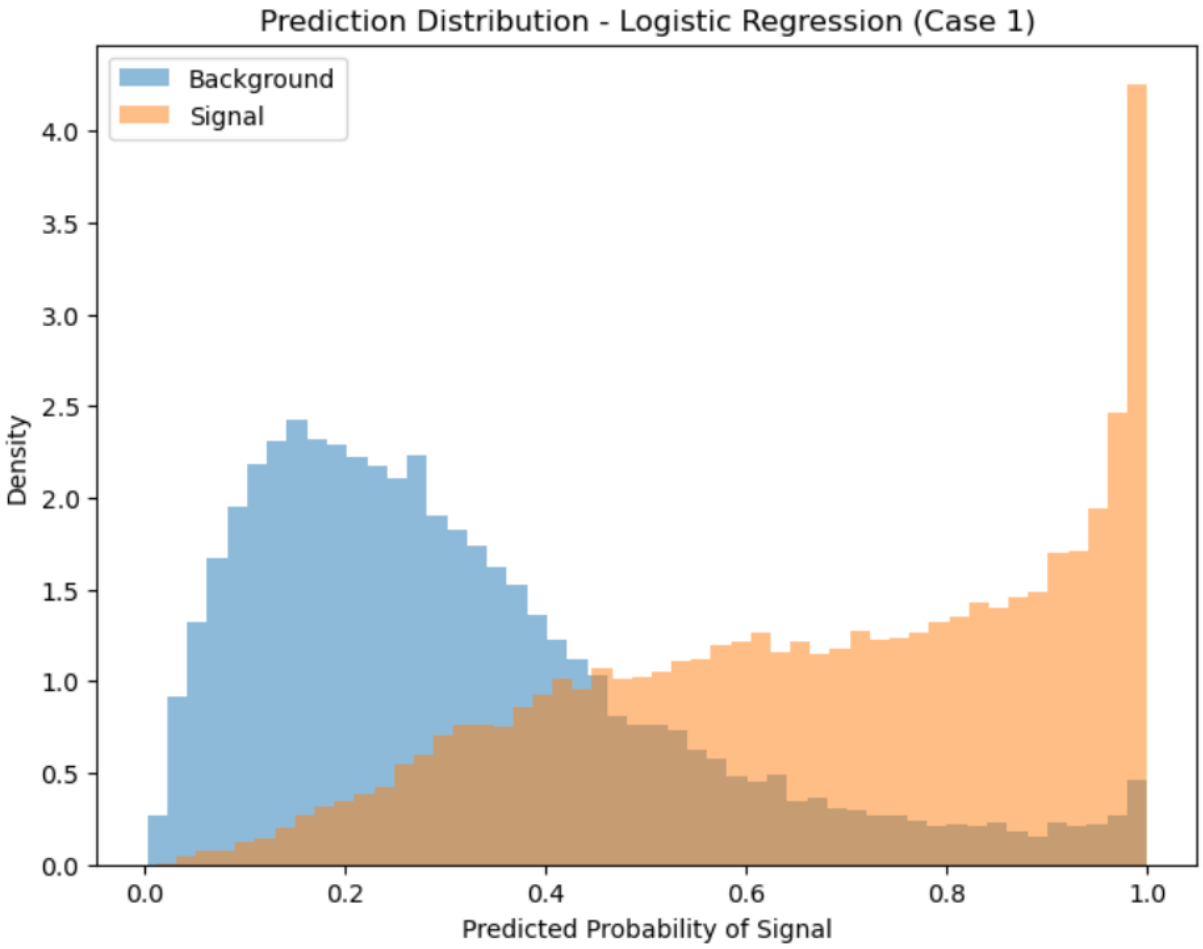


Figure 8: Histogram of predicted probabilities for logistic regression (case 1).

In addition to ROC curves, the confusion matrices for each classifier offer insight into the types of classification errors made. Below are the confusion matrices for Case 1:

- Naive Bayes:
    - True Negatives: 21,172
    - False Positives: 2,575

    - False Negatives: 12,281
    - True Positives: 11,516
- LDA:
    - True Negatives: 19,101
    - False Positives: 4,646
    - False Negatives: 6,439
    - True Positives: 17,358

- Logistic Regression:
    - True Negatives: 19,560
    - False Positives: 4,187
    - False Negatives: 6,161
    - True Positives: 17,636

These matrices confirm that Logistic Regression achieves the best balance between true positives and false positives, making fewer misclassifications overall compared to the other models. Naive Bayes shows the highest false negative rate, while LDA shows a good balance but is slightly outperformed by Logistic Regression.

For Case 2, where PCA was used to reduce the input space to 9 principal components, the models performed similarly but with subtle shifts in their trade-offs. The confusion matrices for Case 2 are:

- Naive Bayes:
    - True Negatives: 21,038
    - False Positives: 2,709
    - False Negatives: 13,913
    - True Positives: 9,884

- LDA:
    - True Negatives: 18,903
    - False Positives: 4,844

- False Negatives: 6,391
- True Positives: 17,406

- Logistic Regression:
  - True Negatives: 19,406
  - False Positives: 4,341
  - False Negatives: 6,211
  - True Positives: 17,586

While Logistic Regression again provides the strongest overall performance, it is worth noting that LDA is nearly as effective in this case, with slightly fewer false negatives but slightly more false positives. Naive Bayes underperforms significantly on this reduced feature space, showing the highest number of misclassified signals. The ROC curve for case 2 was quite similar to case 1, it can be seen in figure 9 below. One difference is the curve for Naïve Bayes, which is noticeably lower and outperformed compared to its performance in case 1.

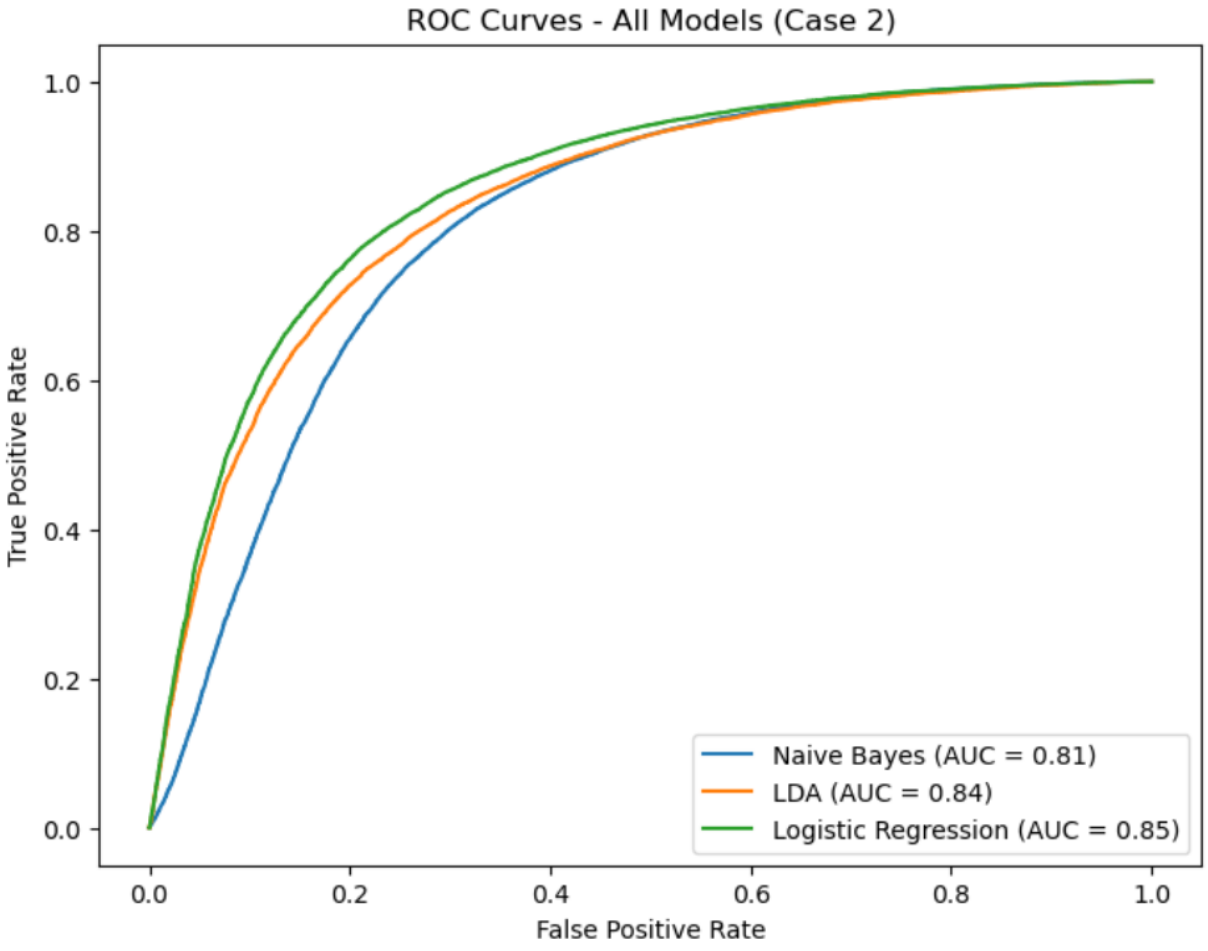


Figure 9: ROC curve of classification models for case 2 (only the variables explaining 85% of the variance).

Figure 10 shows the prediction distribution for case 2. It is again very similar to the same graph plotted for case 1, with a clear separation between signal and background probabilities, although the distinction is slightly less pronounced, indicating a small drop in classification power when using the PCA-reduced feature set.

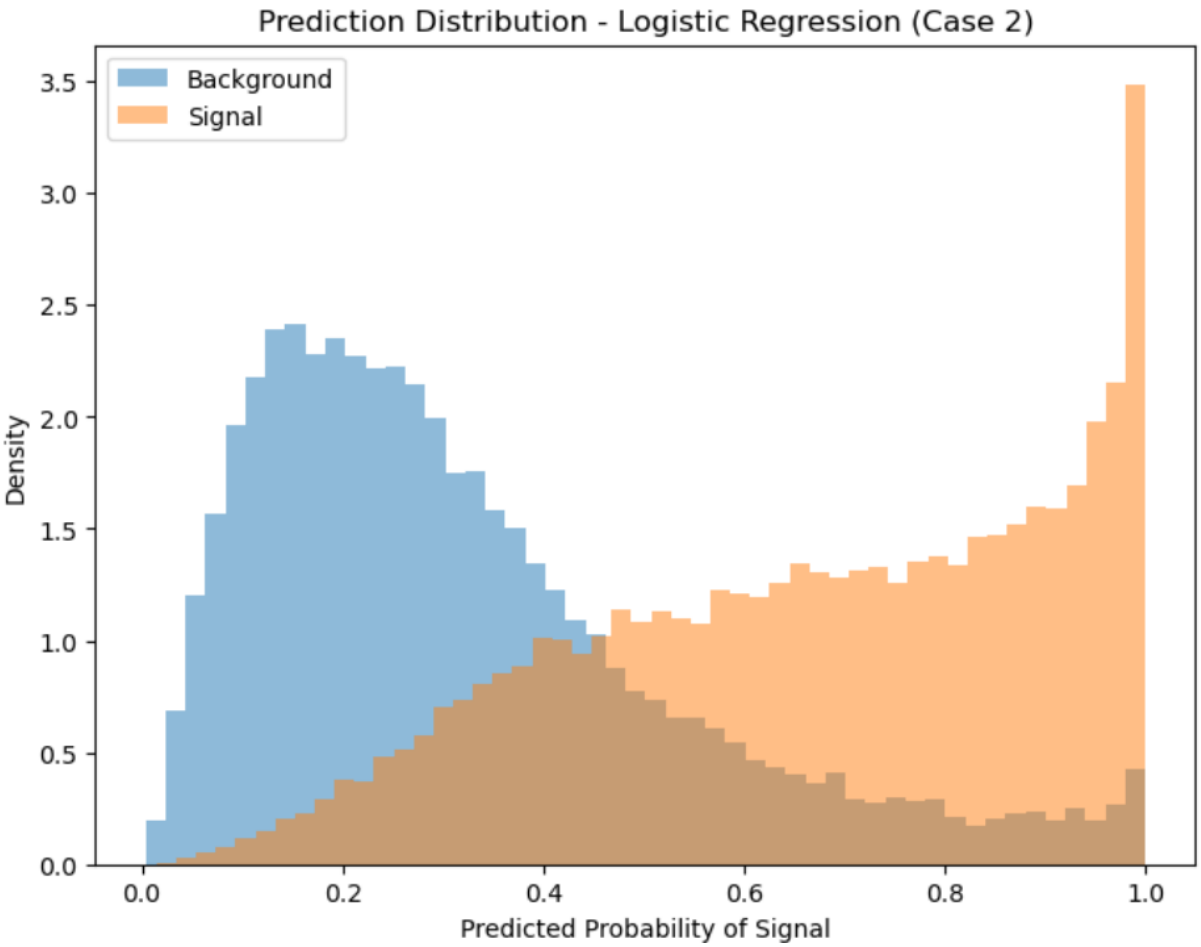


Figure 10: Histogram of predicted probabilities for logistic regression (case 1).

5 Threshold Optimisation Using Punzi Significance

To optimise the classification threshold beyond the default 0.5, the Punzi significance was calculated across a range of thresholds. The Punzi figure of merit is specifically designed for searches involving rare signal events, such as those in particle physics (Punzi, 2003) [1]. It is defined as:

$SP(t) = \varepsilon(t) / (1 + \sqrt{B(t)})$

where ε(t) is the signal efficiency and B(t) is the number of background events that survive the threshold t. This metric balances the competing needs of retaining signal while minimising background.

Figure 11 shows the Punzi significance curve as a function of threshold. A maximum significance of 0.0061 was observed at a threshold of 0.645. This value represents the optimal trade-off between signal efficiency and background contamination, which helps to pick up on rare events, which is a key concern in high-energy physics analysis such as those performed by the CMS Collaboration (CMS, 2014) [5].

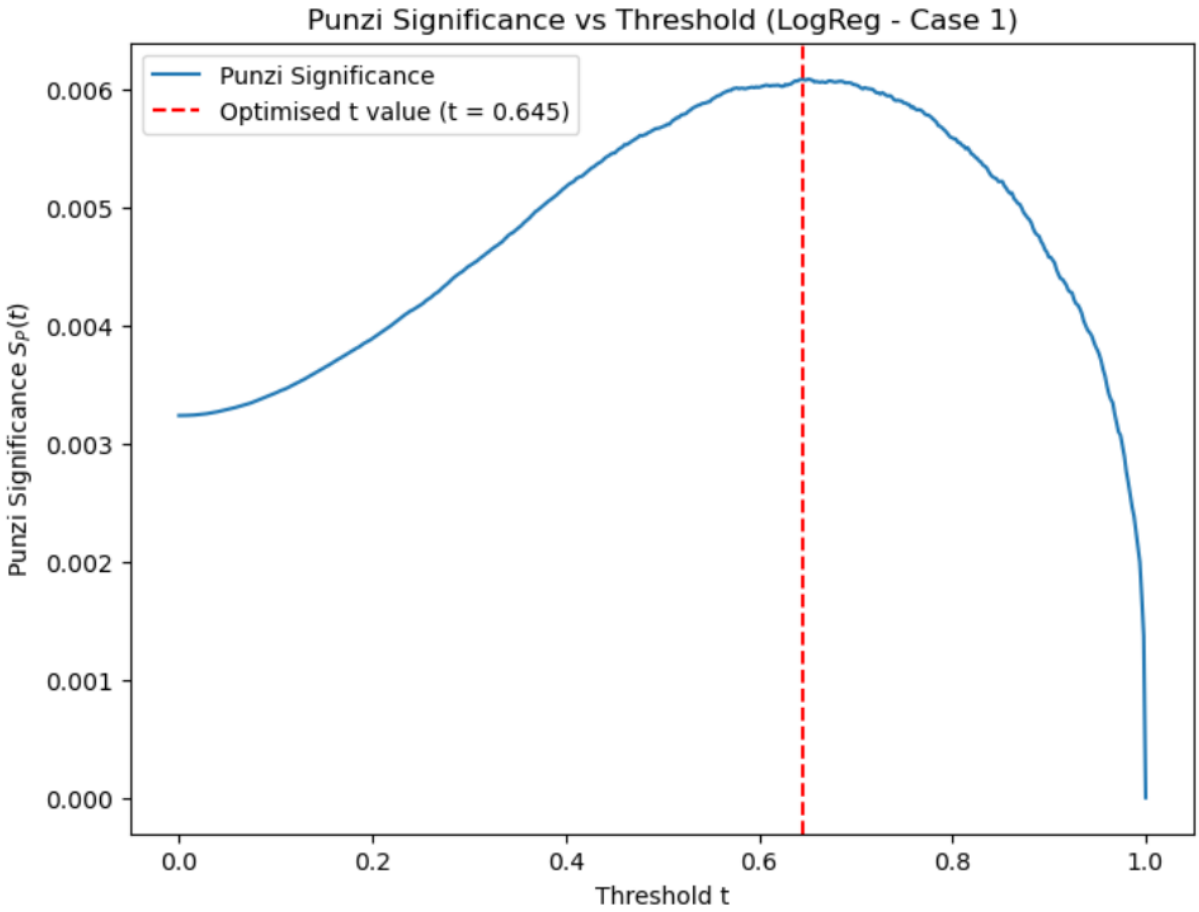


Figure 11: Punzi Significance Curve of Logistic Regression (case 1).

6 Conclusion

Logistic Regression using the original features (Case 1) was the most effective classifier for this dataset, achieving a high AUC and clean signal-background separation. PCA revealed variable redundancy and confirmed the importance of track and vertex-based features, though using the full feature set yielded the best performance overall.

The Punzi significance analysis enabled optimal threshold selection, helping to make the final classification decision. Future work could explore additional event-level features (e.g. MET, jet multiplicity) or test advanced classifiers like decision trees or boosted ensembles.

7 References

[1] CMS Collaboration (2014) 'Identification of b-quark jets with the CMS experiment', *Journal of Instrumentation*, 8, P04013.
Available at: https://doi.org/10.1088/1748-0221/8/04/P04013 (Accessed: 28 April 2025).

[2] Fawcett, T. (2006) 'An introduction to ROC analysis', *Pattern Recognition Letters*, 27(8), pp. 861–874.
Available at: https://doi.org/10.1016/j.patrec.2005.10.010 (Accessed: 28 April 2025).

[3] Jolliffe, I.T. (2002) *Principal Component Analysis*. 2nd edn. New York: Springer.

[4] Pedregosa, F. et al. (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830.
Available at: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (Accessed: 28 April 2025).

[5] Punzi, G. (2003) 'Sensitivity of searches for new signals and its optimization', *Proceedings of the Statistical Problems in Particle Physics, Astrophysics and Cosmology Workshop*.
Available at: https://arxiv.org/abs/physics/0308063 (Accessed: 28 April 2025).