Studying supervised and unsupervised machine learning techniques

Dafne Petrelli

Department of Physics, Queen Mary, University of London, Mile End Road, London, E1 4NS, UK

Abstract

This report focuses on the study of applying supervised and unsupervised machine learning techniques, using a data set which describes particle jets, derived from simulated proton-proton collisions like the ones detected by the Compact Muon Detector (CMS) at C.E.R.N., Geneva, Switzerland. The aim of this study is build models which can successfully determine the distinction between these jets which either originate from Higgs bosons decay or from quantumchromodynamic (QCD) multi jet production. This report provides a brief overview on detection at CMS and its contribution when studying the Standard Model in physics. Then it divulges into the data set's exploration. Unsupervised machine learning is explored, through Principal Component Analysis (PCA). Additionally, supervised machine learning techniques such as Naive Bayes, Linear Discriminant Analysis (LDA) and Logistic Regression (LR) were investigated through two cases. One included training a classifier on background vs signal using all variables and the other one only using 85% of the total variance. Finally, an interpretation of the Punzi Significance, which was performed on the best classification model was provided.

1 Introduction

1.1 Motivation of study

1.1.1 CMS detector

The Compact Muon Detector (CMS) is located at one of the four collision points inside the Large Hadron Collider (LHC) at C.E.R.N. in Geneva, Switzerland. The LHC is the largest and most powerful particle accelerator ever built. Its function is to accelerate proton particles at speeds closest to the speed of light, allowing collisions to take place at four locations around its ring. The CMS's purpose is to detect these collisions to gather data about the momenta and energies of these particles, which is then used to roughly characterize quark properties [1].

1.1.2 Higgs bosons

The Higgs boson is a key component of the Standard Model of particle physics, which unifies the fundamental particles and their interactions. The Standard Model describes all matter as consisting of elementary fermions, grouped into quarks and leptons, which interact via fundamental forces, each governed by its own gauge theory [2]. One of these, Quantum Chromodynamics (QCD), describes the strong nuclear force through the exchange of gluons. Nonetheless, most of the visible mass in matter does not come from the Higgs field but rather from QCD dynamics, specifically, the energy stored in the strong interaction between quarks and gluons [3]. Therefore, in studying this distinction it could contribute advancing our understanding of particle interactions and identifying Higgs boson decays within the high background noise from QCD processes. It is crucial for refining experimental strategies in future LHC analyses, where separating signals from background noise will lead to uncovering new physics phenomena and unveiling more information about the Standard Model.

1.2 Data set exploration

The file which was handled throughout this study, comprises of particle jets, derived from a simulated proton-proton collisions like the ones detected by the CMS and it was labeled as "cms_Hbb.csv". The file describes a total of 26 numerical variables which analysis determines the distinction between jets which originate from a Higgs boson decaying to a bottom quark-antiquark pair or jets which originate from quantumchromodynamic (QCD) multijet production as showcased in **Table 1**.

Feature	**Count**	**Non-Null**	**Dtype**	**Mean**	**Std**
Unnamed: 0	225 868.00	non-null	int64	93 832.92	54 160.98
jetNTracks	225 868.00	non-null	float64	18.23	6.61
nSV	225 868.00	non-null	float64	2.95	1.06
tau0_trackEtaRel_0	225 868.00	non-null	float64	2.57	0.85
tau0_trackEtaRel_1	225 868.00	non-null	float64	3.10	0.98
tau0_trackEtaRel_2	225 868.00	non-null	float64	3.67	1.22
tau1_trackEtaRel_0	225 868.00	non-null	float64	2.50	0.87
tau1_trackEtaRel_1	225 868.00	non-null	float64	2.99	2.99
tau1_trackEtaRel_2	225 868.00	non-null	float64	3.51	1.25
tau_flightDistance2dSig_0	225 868.00	non-null	float64	13.64	23.20
tau_flightDistance2dSig_1	225 868.00	non-null	float64	13.01	23.95
tau_vertexDeltaR_0	225 868.00	non-null	float64	0.11	0.12
tau_vertexEnergyRatio_0	225 868.00	non-null	float64	0.48	1.16
tau_vertexEnergyRatio_1	225 868.00	non-null	float64	0.59	1.33
tau_vertexMass_0	225 868.00	non-null	float64	4.03	5.71
tau_vertexMass_1	225 868.00	non-null	float64	5.27	6.76
trackSip2dSigAboveBottom_0	225 868.00	non-null	float64	4.50	5.64
trackSip2dSigAboveBottom_1	225 868.00	non-null	float64	3.04	3.63
trackSip2dSigAboveCharm_0	225 868.00	non-null	float64	6.98	7.29
trackSipdSig_0	225 868.00	non-null	float64	16.74	27.91
trackSipdSig_0_0	225 868.00	non-null	float64	9.87	19.74
trackSipdSig_0_1	225 868.00	non-null	float64	3.03	6.04
trackSipdSig_1	225 868.00	non-null	float64	6.76	8.10
trackSipdSig_1_0	225 868.00	non-null	float64	11.10	22.31
trackSipdSig_1_1	225 868.00	non-null	float64	3.73	5.94
trackSipdSig_2	225 868.00	non-null	float64	3.80	4.32
trackSipdSig_3	225 868.00	non-null	float64	2.35	2.67
isBackground	225 868.00	non-null	float64	0.65	0.48
isSignal	225 868.00	non-null	float64	0.35	0.48

Table 1: Main statistics of "cms.csv".

"cms_Hbb.csv" was handled in Python language. The dataset didn't contain any missing values,so what proceeded was its data analysis.

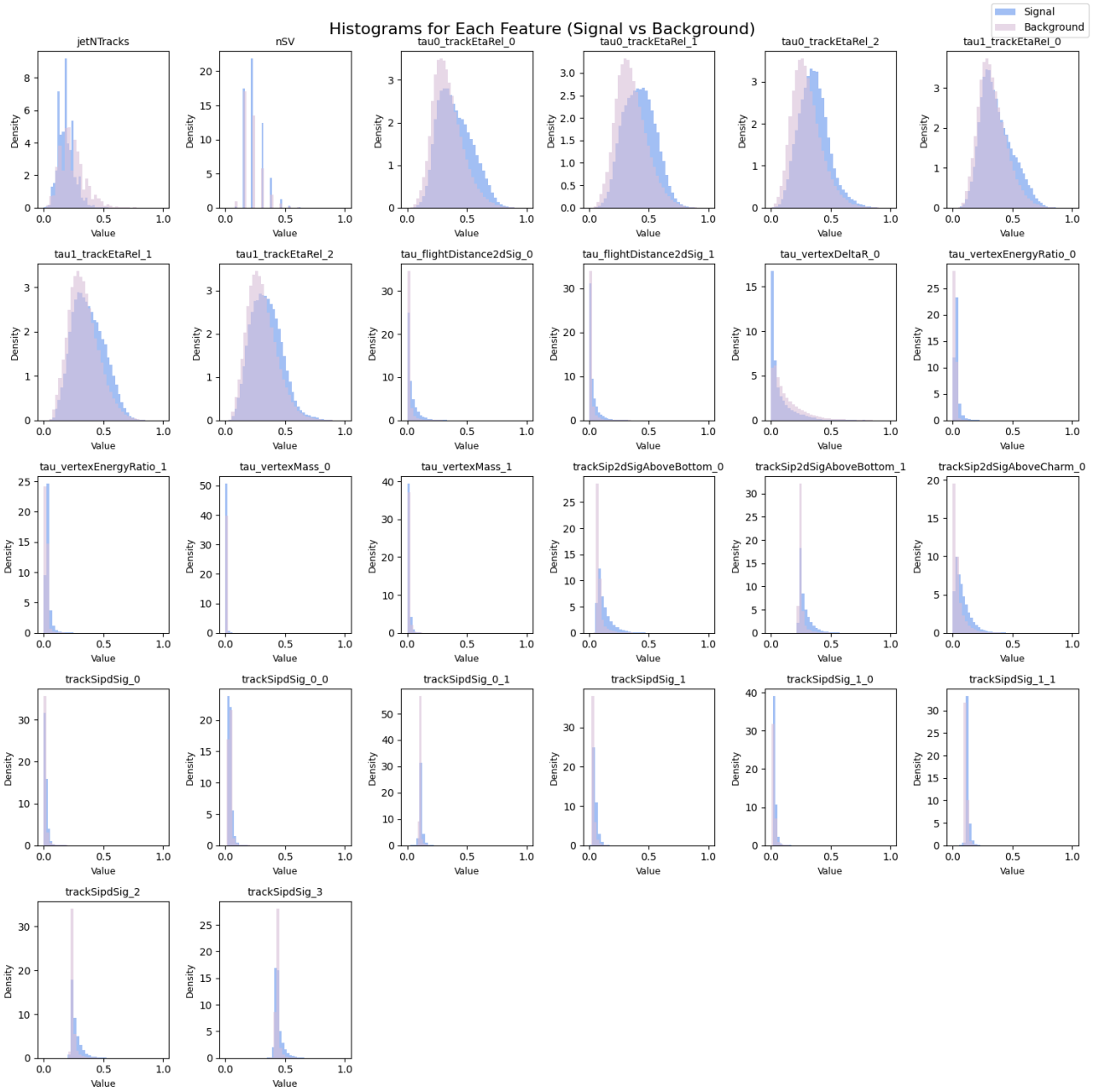

Figure 1: Histogram plots for each feature of the jets data set, in linear scale for the y-axis.

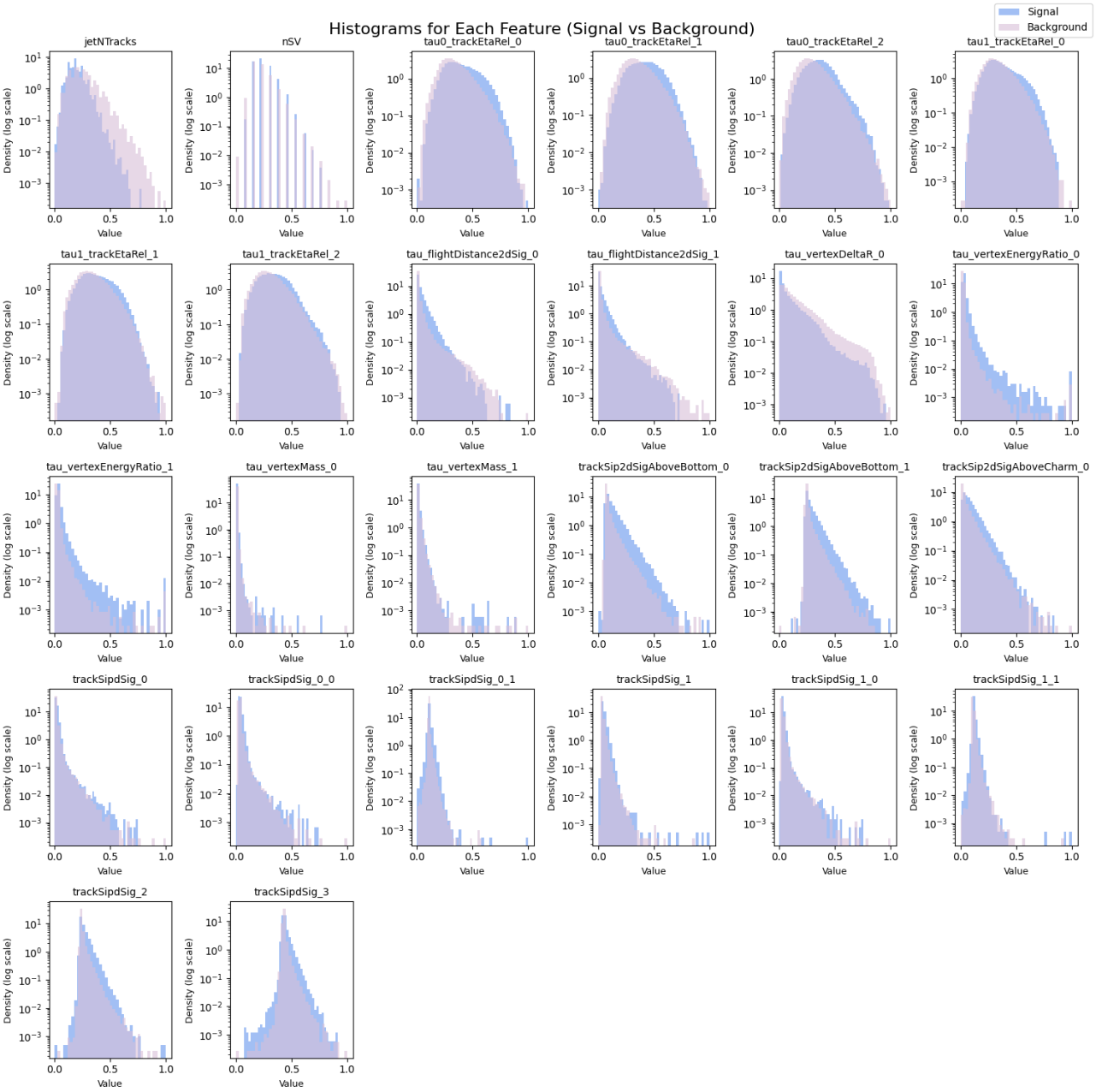

Figure 2: Histogram plots for each feature of the jets data set, in logarithmic scale for the y-axis.

To understand the different features of the jets, the distribution of the histograms was taken into consideration.

Fig. 1 displays that "tau0l_trackEtaRel_0/1/2" and "tau1_trackEtaRel_0/1/2" present a visible central distribution which exhibits the greatest overlap between signal and background compared to the other features. These two sets of features correspond to the first, the second and the third smallest track pseudorapidity $\Delta\eta$, relative to the jet axis, associated to the first N-subjettiness axis and the second N-subjettiness axis respectively. Pseudorapidity is a quantity used to describe the rapidity of massless or nearly massless particles. It expresses the angle of a particle's trajectory relative to the beam axis, especially for particles moving close to the speed of light. N-subjettiness is a jet shape variable used to identify boosted heavy particles that decay

into N partons. It characterizes the internal structure of jets by measuring how closely the jet constituents align with N subjet axes. The tau (τ) values approach zero when the jet is consistent with having N or fewer subjets. Due to Pseudorapidity being directly correlated to the detector components, features such as "tau0l_trackEtaRel_0/1/2" and "tau1_trackEtaRel_0/1/2" show clear signal-background separation on a linear scale which is of a highly importance when studying jet substructure. However, the rest of the features showcased low visibility when plotting them in linear scale, therefore a logarithmic scale was applied such as in **Fig. 2**. This was done to analyze rare events displayed in the extreme tails which the histogram produces when comparing signal-background distribution. From this analysis it can be observed that "trackSipdSig_0/1/2/3" "trackSipdSig_0/1_0/1" exhibit the greatest signal-background separation. These two features correspond to the first/second/third/fourth largest track 3D signed impact parameter significance and to the first/second (first index) largest track 3D signed impact parameter significance associated to the first/second (second index) N-subjettiness axis respectively. The impact parameter is defined as the shortest distance between a reconstructed particle track and the primary collision vertex. When Higgs bosons decay to a bottom quark-antiquark pair, these pair produces B-hadrons which show large impact parameters, since they originate from a secondary vertex, which are rare events but appear very distinct in the histograms due to their long tail distribution. On the other hand, background tracks align usually with the primary vertex, which leads to the impact parameter significance to approximately being 0. Hence, the plotting in the logarithmic scale facilitates the signal-to-background distinction as the signal tails stand out more evidently from the background [4].

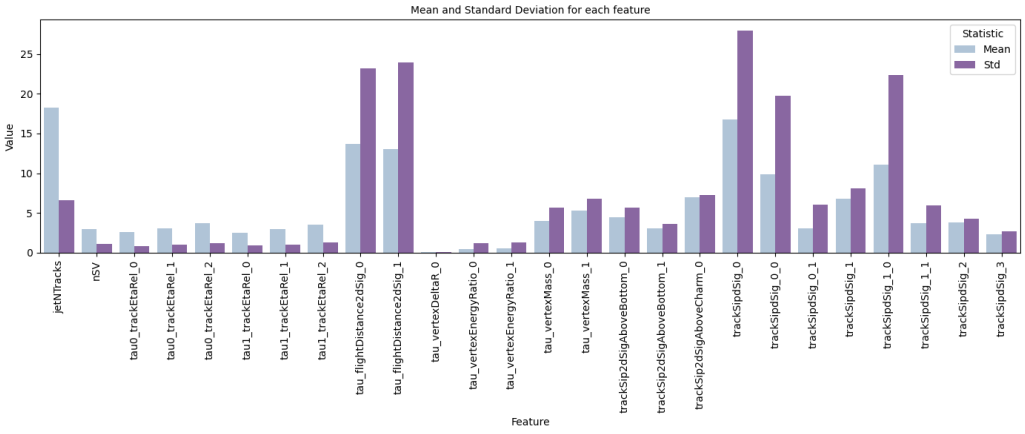

Figure 3: Bar plot comparing the mean and standard deviation for each of the particle jets features.

To visualize the main statistics from *Table 1*, a comparison bar plot for each feature was computed as in **Fig. 3**. The feature which showcases the highest mean corresponds to "jetNTracks" which refers to the number of tracks associated with the jet. Since QCD jets are high-energy and can fragment into multiple charged particles, each producing a track, the mean if the "jetNTracks" being the highest (mean = 18.23) could indicate that the data set mainly consists of background events.
On the other hand, the feature with the highest standard deviation is yielded by the "trackSipdSig_0" feature (std = 27.91). Compared to its standard deviation, its mean is very low. This may entail

that most of the tracks are consistent and their significance is near 0, but there a few which are heavily displaced, which is what is causing the standard deviation to be inflated. Finally, features such as "tau_flightDistance2dSig_0" and "tau_flightDistance2dSig_1", present a standard deviation (std = 23.20 and std = 23.95 respectively) almost doubled their mean (mean = 13.64 and mean = 13.01 respectively). These features correspond to how significant the transverse flight's distance is from its primary vertex to its secondary vertex, based on its N-subjettiness axis [4]. The low mean could be due to most decays showing minimal displacement, leading to very small flight distance being of approximately 0. However, the high variance must be due to a small fraction of events, which exhibit a significant vertex displacement, leading to very large values.

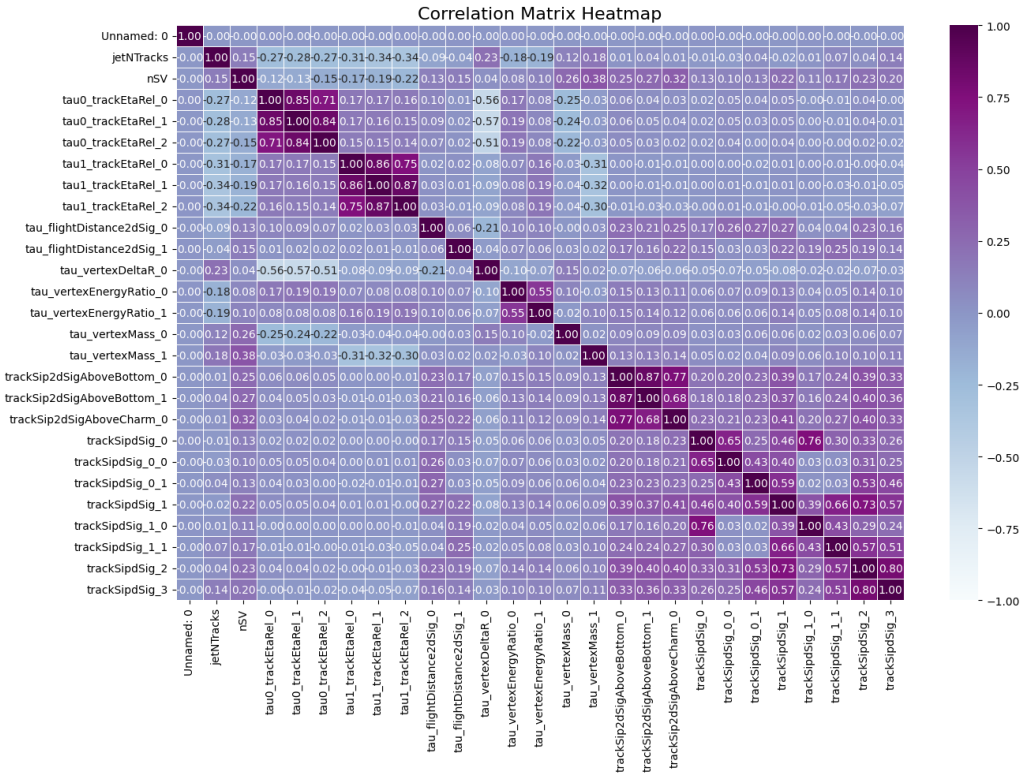

Figure 4: Heat map for the particle jets features.

A heat map was computed, as in **Fig. 4**, to visualize the correlations between different features, with darker colors indicating stronger correlations. The two most prominent positive correlations were observed between the "trackSip2dSigAboveBottom_1" and "trackSip2dSigAboveBottom_0" (r = 0.87), and between the "tau1_trackEtaRel_1" and "tau1_trackEtaRel_0" (r = 0.86). These strong relationships may be attributed to their co-linearity in the context of signal detection. The first strongest correlation is due to both features measuring the 2D impact parameter significance of the first or second track, which when added up, they result in the total mass of the tracks to overcome the B-hadron threshold [4]. Therefore, these tracks yield multiple large and similar displacements from the collision points. The second strongest correlation is due to both features stemming from the same kinematic conditions in the symmetric decaying of Higgs bosons into bottom quark-anti quark jets.
On the other hand, the two strongest negative correlations were observed between the "tau_vertexDeltaR_0" and "tau0_trackEtaRel_1" (r = -0.56) and between the "tau_vertexDeltaR_0 and

"tau0_trackEtaRel_0" (r = -0.55). The "tau_vertexDeltaR_0" corresponds to the angular distance ΔR between the first N-subjettiness axis and number of secondary vertices (SV) direction. This inverse relationship suggests that as the tau decay vertex moves further from the beam axis, the track pseudorapidities decrease, which leads to the jet track showcasing as narrower and more tightly-wrapped, this leads to the jet structure being more focused once the decay is taking place. This strong negative correlation quantifies the decaying of Higgs bosons into bottom quark-anti quark signal jets.

The 4 previous correlations were then plotted as 2D histograms to analyse their relation further.

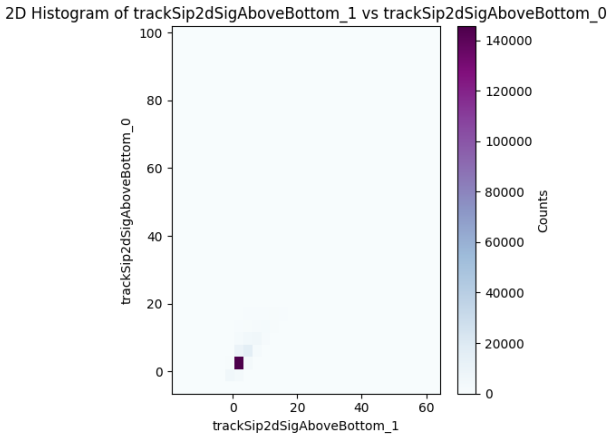

Figure 5: 2D Histogram of tS2dSAB_0 as a function of tS2dSAB_1.

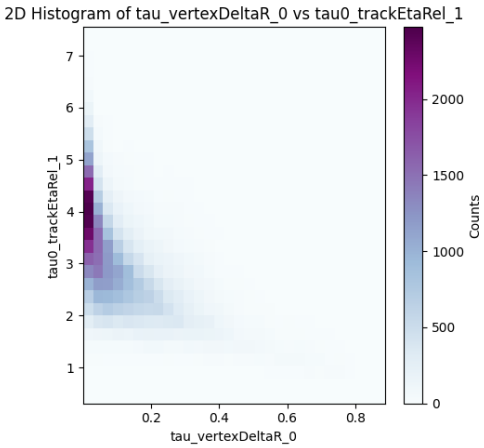

Figure 7: 2D Histogram of tau0_tER_1 as a function of tau_vDR_0.

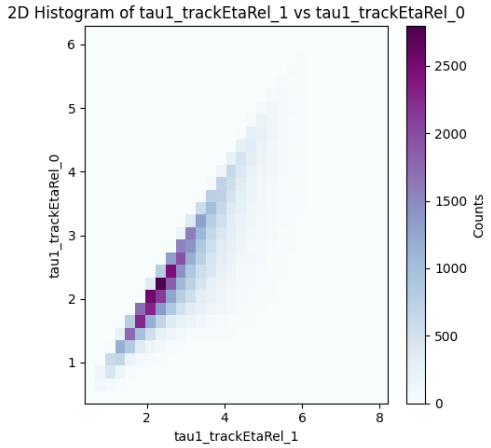

Figure 6: 2D Histogram of tau1_tER_1 as a function of tau1_tER_0 .

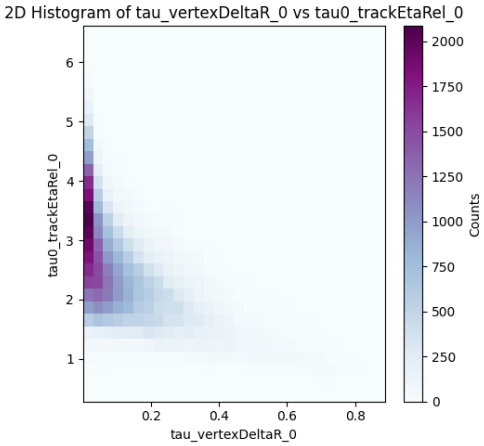

Figure 8: 2D Histogram of tau0_tER_0 as a function of tau_vDR_0.

Fig. 5 displays a single, dense purple region that fades outward, indicating that most data points are concentrated in a narrow area. This suggests low variance between the two features plotted and possibly limited discriminative power. On the other hand, **Fig. 6** exhibits a tightly clustered diagonal pattern, reflecting a strong correlation and coherent track directionality between the features. Both **Fig. 7** and **Fig. 8** show visible clustering of points in the range of

0.0 to 0.2 for tau_vertexDeltaR_0, suggesting that the tracks are well-aligned with the vertex and exhibit moderate relative pseudorapidity. The strong density in this region may point to reconstructed jets, which are informative for classification and could improve discrimination against background.

---

2 Modeling of the CMS data

2.1 Principal Component Analysis

2.1.1 Theory

To quantify the way in which the numerical variables covary, Principal Component Analysis (PCA) was performed, separately on background and signal events. This model's purpose is to reduce dimensionality of large data sets, by combining multiple predictor variables into a smaller set of variables, defined as the "principal components". The principal components correspond to weighted linear combinations of the originals set [5]. The weights assigned to each original variable indicate how much that variable contributes to each principal component. The principal components can be defined in **Eq. 1** as,

$$Z_i = \sum_{j=1}^{p} w_{ij} X_j \tag{1}$$

where Z_i corresponds to the principal component, X_j refers to the original variable, w_{ij} is the weight of the variable j^{th} in the component i^{th} and p is the total number of variables on which PCA is performed on.

2.1.2 Performance of PCA

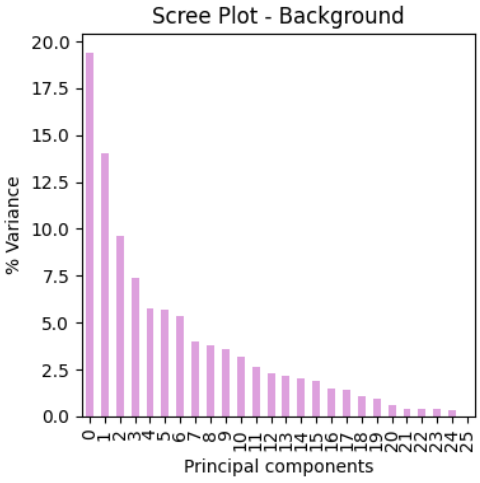

Figure 9: Scree plot performed for Background.

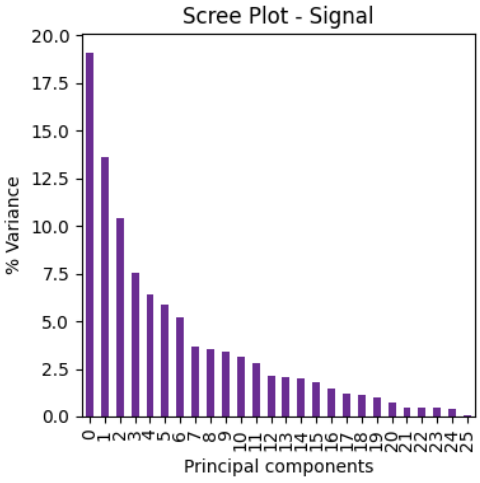

Figure 10: Scree plot performed for Signal.

Fig. 9 and **Fig. 10** display the scree plots which were plotted for background and signal respectively. Both of the graphs' PC1 (which is represented by index 0) exhibit the highest

variance which is about 20%, then PC2, PC3, PC4 (which are represented by index 1,2 and 3 respectively) with a variance ranging from 13% to 7% and finally beyond PC7, the variance drops gradually. This suggests that the first 5 principal components retain the most information about the variance. In **Fig. 9** from PC20 and in **Fig. 10** from PC19, all principal components contribute a variance of less than 1%, which leads them to statistically redundant, making these predictors suitable to be dropped, in further analysis.

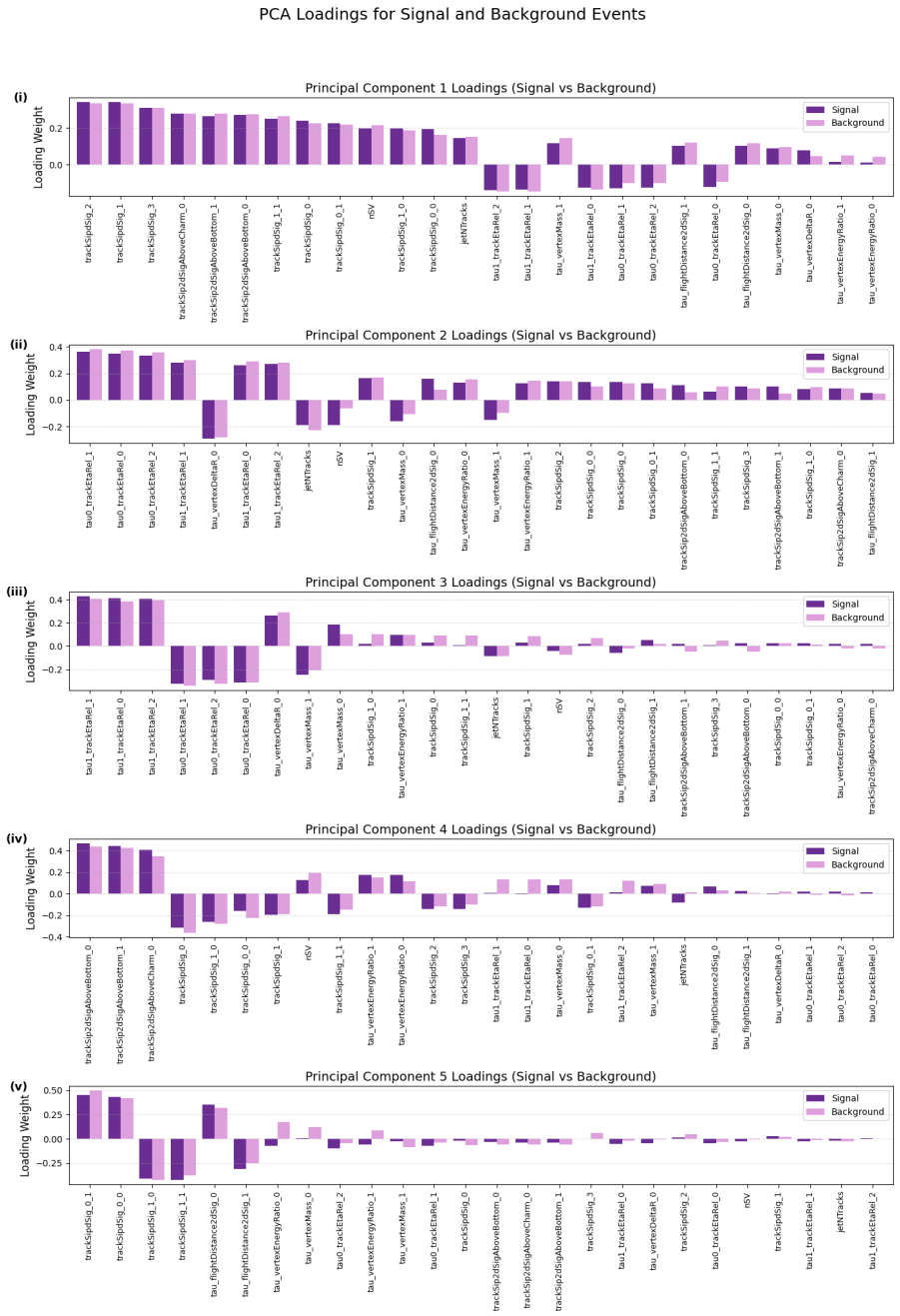

Figure 11: Comparison plot for signal and background plots for PC1, PC2, PC3, PC4, PC5.

In **Fig. 11**, the weights for PC1, PC2, PC3, PC4 and PC5 were computed, showing a comparison between the signal and background events. The reason why these first 5 components were the only ones which were plotted is because they held the most statistically significant variance. PC5, PC6, PC7 share a similar variance, therefore only PC5 was selected and beyond

that an "elbow" shape exhibits in the graph, signifying a linear decline.
(i) The features that contributed most significantly to PC1 were identified as "trackSipdSig_1/2/3" (positively) and "tau1_trackEtaRel_0/1/2" (negatively), indicating that these variables exhibit the most dominant variance along this principal component. PC1's role may involve separating linear components from angular components. This leads to a strong discriminative power in distinguishing signal (Hbb decay) events from QCD background, likely due to notable differences in impact parameter significance between these two jet types.
(ii) The features that contributed most significantly to PC2 were identified as "tau0_trackEtaRel_0/1/2" (positively) and "tau_vertexDeltaR_0" (negatively). These geometric properties may be describing the particles decay topology. This implies that variations in the angular distribution and alignment of tracks relative to the jet axis represent key secondary factors in distinguishing signal from background.
(iii) The features that contributed most significantly to PC3 were identified as "tau0_trackEtaRel_0/1/2" (positively) and "tau1_trackEtaRel_0/1/2" (negatively). PC3's role might be differentiating between the 2 most-prominent N-subjetiness axis.
(iv) The features that contributed most significantly to PC4 were found to be "trackSip2dSigAboveBottom_0_1" and "trackSip2dSigAboveCharm_0" (positively) and "trackSipdSig_0. Their prominence in PC4 suggests that its purpose is based on differentiating between 2D and 3D tracks in relation to the primary vertex, properties which were not explored from the first three principal components.
(v) In contrast to PC1, which captures broader variance in the dataset, PC5 narrows into more localized, jet details. This is evidenced by the dominant features "trackSipdSig_0_1—0_0" (positively and negatively), which might be differentiating between the directionality of the jets. These measurements reveal subtle distinctions in jet behavior that contribute to more detailed signal-background separation [6].

2.2 Classification Models

Classification is a form of predictive modeling, in this project the aim is to predict if the jet corresponds to a background or to a signal. To achieve this, three types of classification techniques were studied including Naive Bayes, Linear Discriminant Analysis (LDA) and Logistic Regression (LR). These were performed in two different cases. One involved training a classifier using all variables, while the second case trained a classifier using the PCA while explaining 85% of the variance.
To compare the classification models, their respective classification reports were computed, including metrics such as the precision, recall, F1-score and accuracy.
The precision measures the accuracy of a predicted positive outcome as

$$precision = \frac{\Sigma TP}{\Sigma TP + \Sigma FP} \tag{2}$$

where TP refers to True positive predictions and FP refers to True negative predictions [7].
The recall measures the strength of the model to predict a positive outcome as

$$recall = \frac{\Sigma TP}{\Sigma TP + \Sigma FN} \tag{3}$$

where FN refers to False negative predictions [7].
The F1 score measures the harmonic mean between the precision and the recall as

$$F1 - score = 2\frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

The accuracy measures the total correct assignments as

$$accuracy = \frac{\Sigma TP + \Sigma TN}{Sample_t} \tag{5}$$

where TN refers to True negative predictions and $Sample_t$ refers to the total sample size [7]. Finally, confusion matrices were computed for the techniques for each case, as they provide significant statistical analysis. In a confusion matrix, the rows correspond to the true class labels, while the columns represent the predicted labels. The entries along the diagonal show the number of correct classifications, whereas the off-diagonal entries indicate instances that were misclassified.

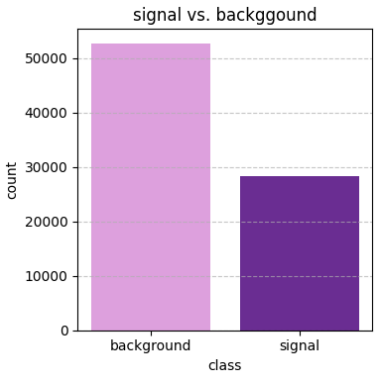

Figure 12: Class distribution categorizing signal and background.

To visualize the class distribution that characterizes the data set a bar plot was computed as in **Fig. 12**. As discussed earlier, QCD jets can fragment into multiple particles, each producing a track, which leads them to form the majority of the data set. Specifically 64.89% of the data corresponds to background events and 35.11% to signal events. This disparity once again highlights the challenges in identifying signal data.
The class imbalance was dealt with, by down sampling the background data to match the signal data. The data was then standardized, so that all the features would have the same variance. This was done to ensure a fair comparison between all the features and improve the classification models performance.

2.2.1 Theory

The Naive Bayes (NB) model observes the probability of the outcome by using a set of predictor values. The probability of an exact matching taking place decreases as the number of combinations increases, since the entire data set is handled. The probability of observing the outcome $Y{=}i$, given by a set of predictors follows **Eq. 6** as

$$P(Y = i|X_1, X_2, ..., X_p) \tag{6}$$

The Linear Discriminant Analysis (LDA) model assumes that the predictor variables are continuous and follow a normal distribution. It emphasizes the "between" group sum of squares, which describes the variation between different classes, and the "within" group sum of squares, which reflects the variation among observations within the same class .

$$\frac{SS_b}{SS_w} \tag{7}$$

where SS_b is the sum in of the squares in "between" and SS_w is the sum of the squares "within". The Logistic Regression (LR) model employs the logistic response function combined with a logit transformation, allowing probabilities,limited to the [0, 1] interval, to be mapped onto an unrestricted continuous scale. This is showcased by **Eq. 8** as,

$$p = \frac{1}{1 + e^{-}(\beta_0 + \beta_1 x_2 + \beta_2 x_2 + ... + \beta_q x_q)} \tag{8}$$

where p is modeled and this transformation allows it to stay between the values of 0 and 1 [7].

2.2.2 Training a classifier on background vs signal using all the variables vs 85% of the variance

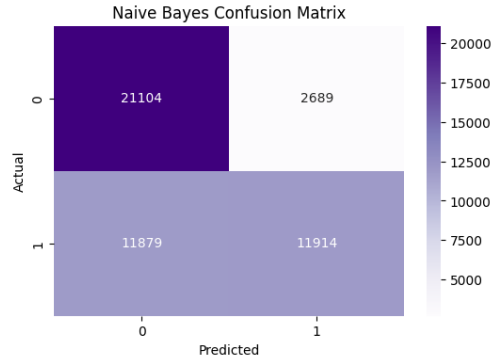

Figure 13: Confusion Matrix for Naive Bayes for all variables.

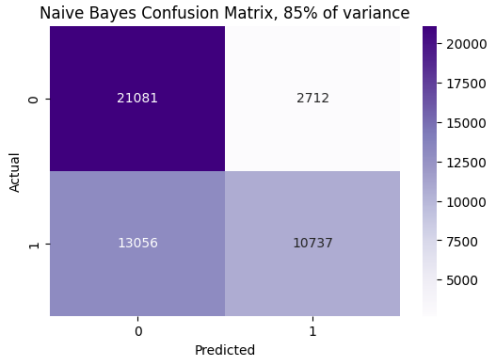

Figure 14: Confusion Matrix for Naive Bayes for 85% of the variance.

When all variables are used, Naive Bayes performed the best when predicting True Negatives (TN = 21104) and False Positives (FP = 2689), which showcased strong precision. However, this model predicts too many False Negatives (FN = 11879), which reduces its efficiency when predicting True Positives. When only 85% of the variance is used, the model's prediction for False negatives and False Positives increases (FN = 13056 and FP = 2712) and for True Positives it decreases (TP = 10737). This indicates that Naive Bayes is quite sensitive when there is a reduction in features.

Class	Precision	Recall	F1- score	Accuracy
0	0.64	0.89	0.74	0.69
1	0.82	0.50	0.62	0.69

Table 2: Classification report for Naive Bayes for all variables.

Class	Precision	Recall	F1- score	Accuracy
0	0.62	0.89	0.73	0.67
1	0.80	0.45	0.58	0.67

Table 3: Classification report for Naive Bayes for 85% of the variance.

From *Table 2* and *Table 3*, it can be gathered that the recall for background events remains the same (= 0.89) and for the precision and the F1-score, there is a slight drop in performance (precision = 0.64 to precision = 0.62 and F1-score = 0.74 to F1-score = 0.73). This showcases how Naive Bayes is still very successful at identifying negatives. However, when it comes a predicting positive values, such as the recall (recall = 0.50 to recall = 0.45) and F1-score (F1-score = 0.62 to F1-score = 0.58) there is a less balanced performance.

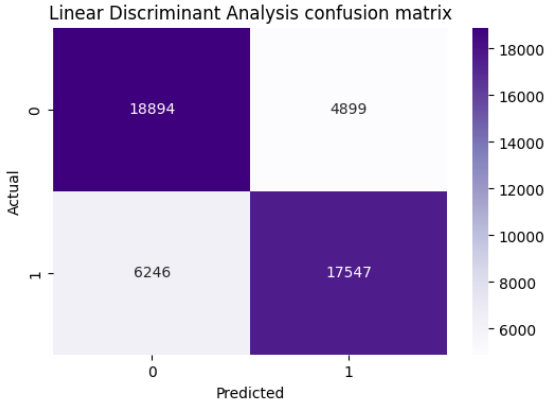

Figure 15: Confusion Matrix for LDA for all variables.

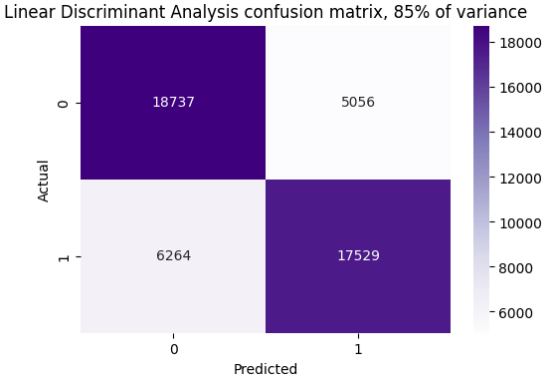

Figure 16: Confusion Matrix for LDA for 85% of the variance.

When all variables are used, LDA had the best performance when predicting False Negatives (FN = 6246). When only 85% of the variance is used, there is a slight increase in False Positives (FP = 5056) and False Negatives (FN = 6264), which showcases the model's sensitivity when reducing variables.

Class	Precision	Recall	F1- score	Accuracy
0	0.75	0.79	0.77	0.77
1	0.78	0.74	0.76	0.77

Table 4: Classification report for LDA for all variables.

Class	Precision	Recall	F1- score	Accuracy
0	0.75	0.79	0.77	0.76
1	0.78	0.74	0.76	0.76

Table 5: Classification report for LDA for 85% of the variance.

From *Table 4* and *Table 5*, it can be observed that LDA performs exceptionally well when retaining all of the metrics when features reduction takes place. The only metric which showcases a reduction is accuracy (accuracy = 0.77 to accuracy = 0.76). Overall, it was best model when retaining dimensionality balance.

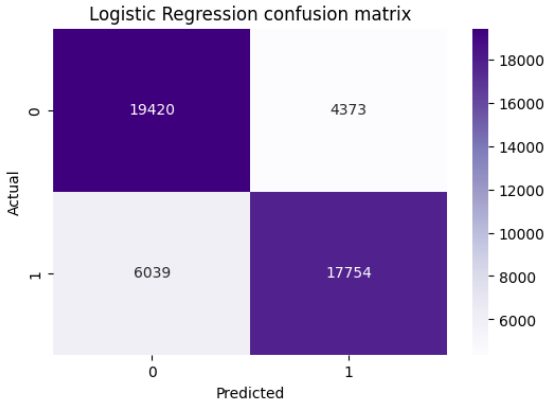

Figure 17: Confusion Matrix for LR for all variables.

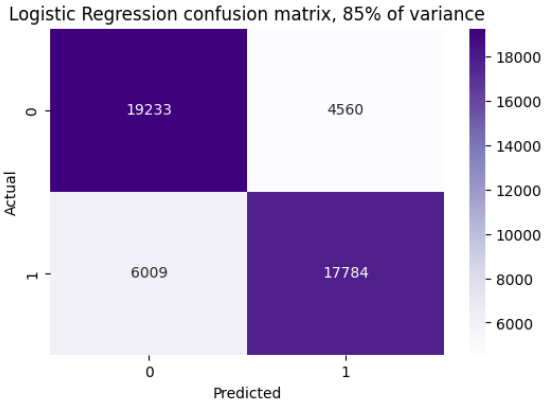

Figure 18: Confusion Matrix for LR for 85% of the variance.

When all variables are used, LR performed the best in retaining stability in both cases. A slight increase in False Positives (from FP = 4373 for all variables to FP = 4560) and a slight decrease

in False Negatives (FN = 6039 to FN = 6009) suggest that LR performs well in dimensionality reduction, showcasing great sensitivity and specificity.

Class	Precision	Recall	F1- score	Accuracy
0	0.76	0.82	0.79	0.78
1	0.80	0.75	0.77	0.78

Table 6: Classification report for LR for all variables.

Class	Precision	Recall	F1- score	Accuracy
0	0.76	0.76	0.81	0.78
1	0.80	0.80	0.75	0.78

Table 7: Classification report for LR for 85% of the variance.

From *Table 6* and *Table 7*, LR shows the best performance across all of the classification models, when retaining dimensionality balance after the features drop. The precision for both events remains unchanged (precision for background = 0.76, precision for signal = 0.80), showing the robustness of the model. Recall dropped slightly for background events (recall = 0.82 to recall = 0.76) but it increased for signal events (recall = 0.75 to recall = 0.80), meaning that less False negatives were predicted. On the other hand, the opposite took place with the F1-score, the signal's F1-score decreased (F1-score = 0.77 to F1-score = 0.75) and the background's F1-score increased (F1-score = 0.79 to F1-score = 0.81), indicating there was a small reduction in balance.

2.2.3 Analysis

For each classification model, plots comparing their probabilities, by overlapping the two distributions from the two type of events were computed.

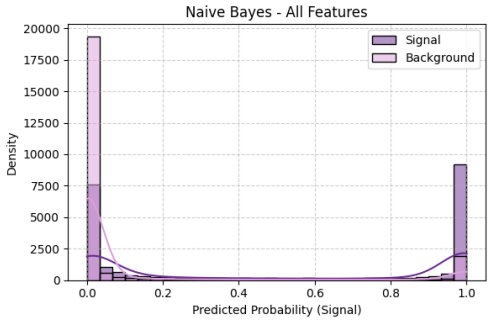

Figure 19: Histogram plot showcasing Naive Bayes's probability distribution for all features.

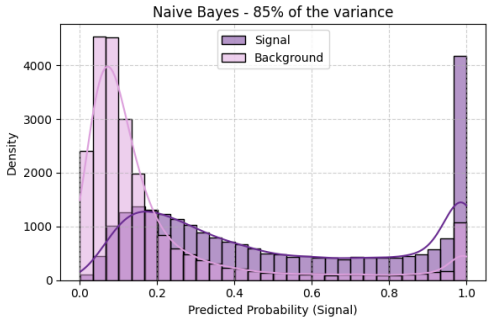

Figure 20: Histogram plot showcasing Naive Bayes's probability distribution for 85% of the variance.

In **Fig. 19** there is a very high separation between the two event's distributions, which suggests that the model is very good at distinguishing between signal and background events. However there is about 45% of the predicted probabilities for signal events lie in the range between 0.0 and 0.2, which coincides with the background peak. This significant overlap in the low-probability region reveals a weakness in the model's precision, particularly in classifying true signal events. Such behavior may also suggest potential overfitting, where the model performs well on the training data but may struggle to generalize to unseen events. **Fig. 20**, introduces an overlap between the events, when reducing the dimensionality of the training data. This displays the performance drop of the model in its precision, recall and F1-score, which was

already discussed when the classification report was computed. By zooming in the 0.0 to 0.2 region, that is where majority of the overlapping takes place, indicating the model's accuracy drop when predicting True Positives.

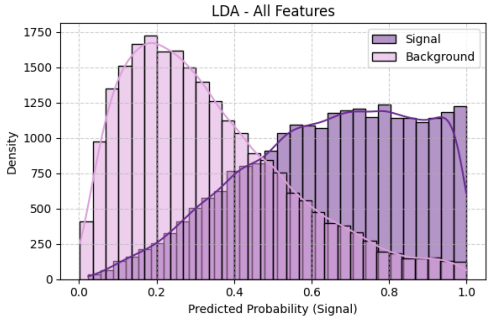

Figure 21: Histogram plot showcasing LDA's probability distribution for all features.

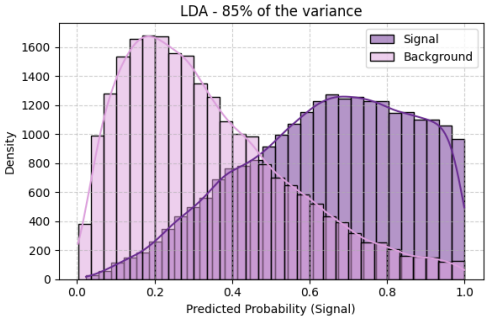

Figure 22: Histogram plot showcasing LDA's probability distribution for 85% of the variance.

Since Linear Discriminant Analysis (LDA) assumes that the predictors follow a normal distribution, it tends to yield more stable and interpretable predictions, which enhance class separability, as demonstrated **Fig. 21**. Both **Fig. 21** and **Fig. 22** exhibit similar levels of separation and overlap between signal and background events. However, fewer peaks are observed in each event distribution when the classifier is trained using 85% of the variance, which may be associated with a slight drop in accuracy due to the reduced feature set. Despite this, the overall similarity in separation between the two figures highlights LDA's robustness and its ability to maintain performance even after dimensionality reduction.

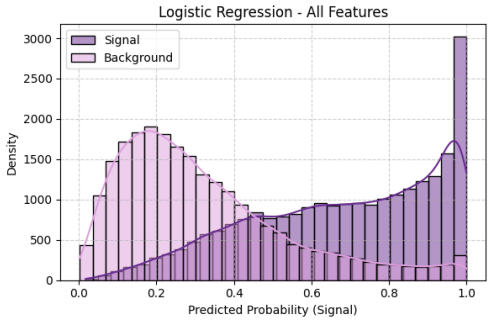

Figure 23: Histogram plot showcasing LR's probability distribution for all features.

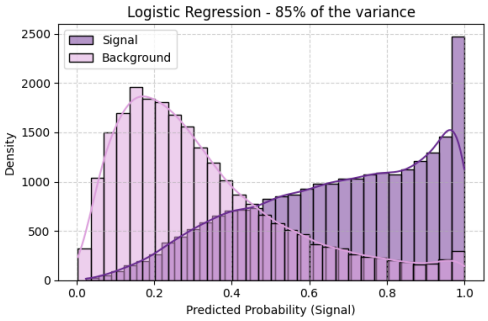

Figure 24: Histogram plot showcasing LR's probability distribution for 85% of the variance.

From **Fig. 23** and **Fig. 24** it can be gathered that LR can moderately distinguish signal events from background events. The overlap between the two increases in **Fig. 24**, meaning more uncertainty is introduced when predicting False Negatives, due to the reduce in dimensionality.

Overall, LR was identified as the best-performing classification model, based on its consistent performance across both cases. It achieved the highest accuracy in both full feature and reduced dimensionality scenarios (accuracy = 0.78), outperforming the other models.
While dimensionality reduction typically leads to a performance decline, especially when many

features contribute valuable information, LR remained robust. In this dataset, it was shown that most features carry nearly equal importance. This was first demonstrated through the correlation heatmap (**Fig. 4**), which indicated low multicollinearity — suggesting that each feature offers unique information. Additionally, scree plots for background and signal events (**Fig. 9**, **Fig. 10**) confirmed that approximately 90% of the total variance is significant when studying jet properties. As expected, when reducing the dataset to 85% of the total variance, a drop in accuracy was observed for both Naive Bayes and LDA. However, Logistic Regression maintained its performance, demonstrating its stability under dimensionality reduction. Further support for LR's effectiveness is provided by its probability distribution plots (**Fig. 23** ,**Fig. 24**), which show minimal overlap between signal and background distributions. This indicates lower prediction uncertainty, particularly in reducing false negatives, and reflects LR's strong discriminative ability.

2.3 Punzi significance

The best performing model turned out to be LR, therefore a threshold optimization using Punzi significance was performed.

2.3.1 Theory

The Punzi significance is a metric which handles the balance between signal efficiency and background rejection [8], which optimizes the search for rare events such as the one studied in this report, where there is a significant higher presence for background compared to signal. This value is computed in **Eq. 9** as,

$$S_p(t) = \frac{\varepsilon(t)}{1 + B\sqrt{(t)}} \tag{9}$$

where $\varepsilon(t)$ corresponds to the efficiency of the signal dependent on threshold t on the classification output and $B(t)$ refers to the number of background events dependent on the same threshold t [4].

2.3.2 Analysis

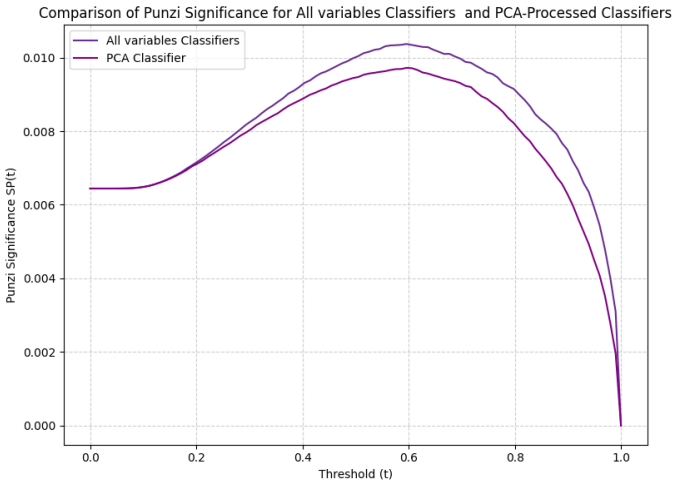

Figure 25: $S_p(t)$ as function of t.

A plot comparing the Punzi Significance (S_p) between the two classifiers was performed as in **Fig. 25**. The optimal threshold for all variance was observed to be 0.69, yielding a Punzi significance of 0.0117, while the optimal threshold for 85% of the variance corresponded to 0.63, yielding a Punzi significance of 0.0114.
This means that the all variables classifier slightly outperforms the PCA classifier, confirming the disadvantage discussed previously when reducing features dimensionality. The optimal threshold corresponds to the maximum signal significance, which allows to misclassify background events. The thresholds of 0.69 and 0.63, should be used when distinguishing between signal and background events for each respective testing.
Moreover, the values for the S_p's for both classifiers are both very low, which is expected when dealing with such a large data set in particle physics, when signal detection is extremely rare.

The difference between the two classifiers performance (Δ $S_p = 0.0003$) being very small, once again confirms all the evaluation that has been performed so far throughout this study. By dropping features, the signal resolution will decrease as a consequence.

---

3 Conclusion

In this report, a dataset consisting of simulated particle jets, analogous to those detected in the Compact Muon Solenoid (CMS) experiment at C.E.R.N, was analyzed. The primary goal was to develop both unsupervised and supervised machine learning models to classify jets as originating either from the decay of Higgs bosons into a pair of bottom and anti-bottom quarks (Hbb decays), or from gluon exchanges as predicted by Quantum Chromodynamics (QCD).
The initial exploration of the dataset involved the use of histograms, heatmaps, and bar plots. These visualizations indicated that the features within the dataset exhibited low pairwise correlations, suggesting a limited presence of multicollinearity. To further investigate the structure and relevance of the features, Principal Component Analysis (PCA), an unsupervised dimensionality reduction technique, was applied separately to signal events (Hbb decays) and background events (QCD). The resulting scree plots demonstrated that the first five principal components (PCA's) accounted for approximately 65% of the total variance in both cases. Further examination of the component weightings enabled insights into which original features contributed most significantly to the variance, offering potential interpretations related to the physical properties of the jets. The study then focused on evaluating the performance of different supervised machine learning models in classifying signal versus background events. Classifiers were trained using both the full set of principal components (representing 100% of the variance) and a reduced set capturing 85% of the variance. The models tested included Naive Bayes, Linear Discriminant Analysis (LDA), and Logistic Regression (LR). Performance was assessed using classification metrics, including precision, recall, accuracy, and F1-score, alongside their resulting confusion matrices.
Logistic Regression was found to exhibit the best overall performance. Although LDA achieved the same values in all metrics expect accuracy, Logistic Regression, not only it maintained a strong balance across all metrics, but it also preserved accuracy after dimensionality reduction. Importantly, further analysis using probability distribution histograms revealed that Logistic Regression exhibited the least overlap between signal and background event predictions, indicating the lowest uncertainty among the models tested.
An important insight from the PCA was that nearly 90% of the original features played a meaningful role in identifying signal events. Consequently, reducing the variance to 85% led to a measurable drop in classification performance, particularly in detecting signal events. This highlighted the importance of retaining a high-dimensional feature space for accurate and robust event classification. Finally, to study how to optimize signal detection and background rejection, Punzi Significance was implemented to LR. The all-variables classifier achieved an optimal Punzi significance of 0.0117 at a threshold of 0.69, while the PCA-reduced classifier reached 0.0114 at a threshold of 0.63. Although the difference between the two values is small (Δ $S_p = 0.0003$), it supports the overall conclusion that dimensionality reduction slightly impairs the classifier's ability to distinguish signal from background. The low values obtained for the S_p's reflect the intrinsic difficulty of detecting extremely rare signals within large background dominated datasets such as the one in this study.
There were three challenges which were identified in this study. The first one involved overfitting. When dealing with such a large data set with this many features holding mostly-equal

importance, there might be a chance in which the classification models overestimate the magnitude of the numerical variables [9]. This exactly what took place when training a classifier using all variance in Naive Bayes, hence why it was ruled out as the weakest performing model. Overfitting can inflate the performance of all metrics, which can lead to a drop in discriminative power, especially when dealing with rare events such as signals. The second challenge was the high class imbalance in this data set, which was highlighted by the S_p values being very low, emphasizing the statistical difficulty that raises when observing signals. At last, the third challenge was the dimensionality reduction. PCA demonstrated to be very good at avoiding overfitting, but it led to a drop in accuracy for both Naive Bayes and LDA. This was because the variance that was chosen corresponded to 85%, when instead it should have been approximately 5% for this specific data set, due to all features retaining significant information about the jets properties and characterization. Therefore a way to improve this study would be to next time perform PCA on 90% of the variance instead. Additionally a model which could be implemented in this study could be Random Forest, which is a supervised machine learning technique which handles high-dimensional data well and is quite robust to overfitting [10].

References

[1] CERN. *Detector — CMS Experiment*. [Online]. Available: https://cms.cern/detector

[2] CERN. (2023). *The Higgs boson*. [Online]. Available: https://home.cern/science/physics/higgs-boson

[3] CMS Collaboration. (2018). Evidence for the Higgs boson decay to a bottom quark–antiquark pair. *Physical Review Letters*, **121**(12), 121801.

[4] Bona, M. (n.d.). *SPA5131 Practical Techniques for Data Science – Final Project*. Queen Mary University of London.

[5] Bona, M. (n.d.). *Lecture Notes 6: Practical Techniques in Data Science*. Queen Mary University of London.

[6] Bass, S.D., De Roeck, A., & Kado, M. (2021). The Higgs boson: Its implications and prospects for future discoveries. *Nature Reviews Physics*, **3**(7), 397–410.

[7] Bona, M. (n.d.). *Lecture Notes 5: Practical Techniques in Data Science*. Queen Mary University of London.

[8] Punzi, G. (2003). Sensitivity of searches for new signals and its optimization. In *Proceedings of the Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology* (pp. 79–83).

[9] Subramanian, J., & Simon, R. (2013). Overfitting in prediction models – Is it a problem only in high dimensions? *Contemporary Clinical Trials*, **36**(2), 636–641.

[10] Baladram, S. (2024). *Random Forest, Explained: A Visual Guide with Code Examples*. Towards Data Science. [Online]. Available: https://towardsdatascience.com/random-forest-explained-a-visual-guide-with-code-examples-9f736a6e1b3c/