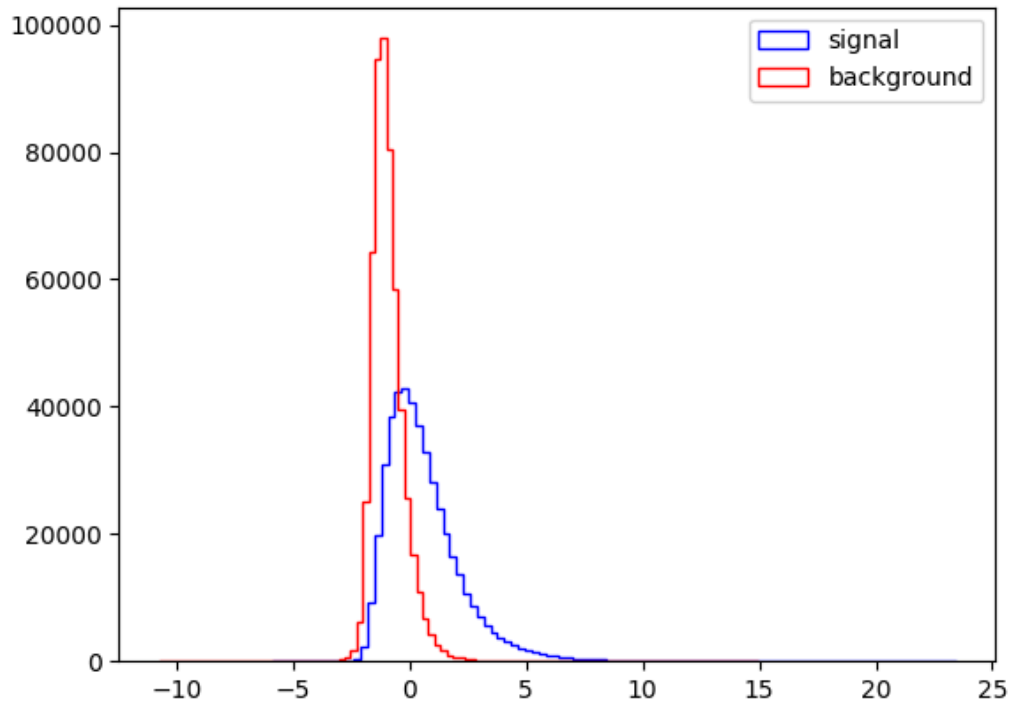


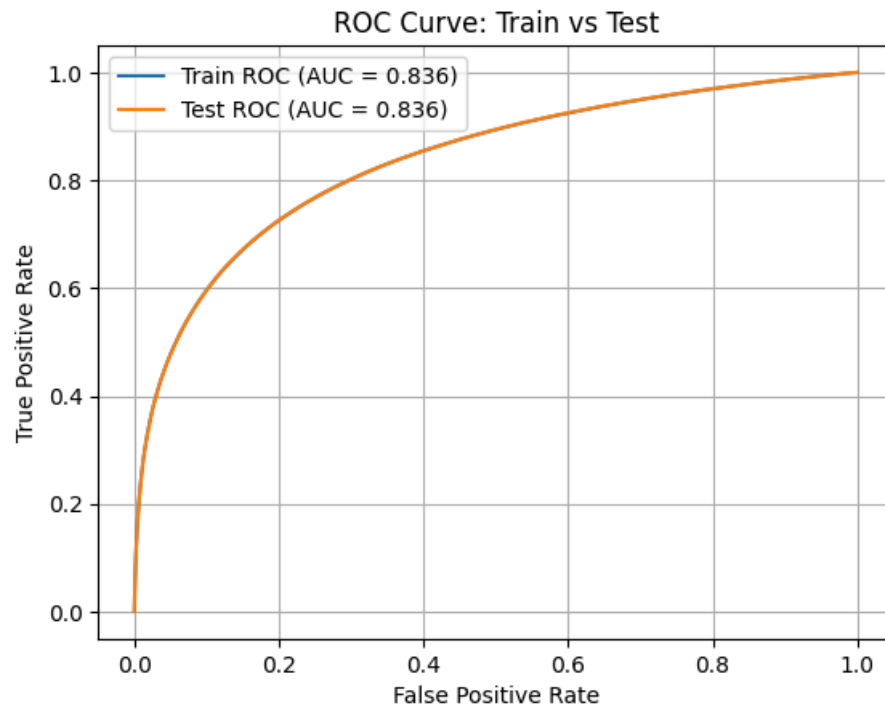
# DATA 3402

Wonho Jeong  
1002242697

## Exercise 3



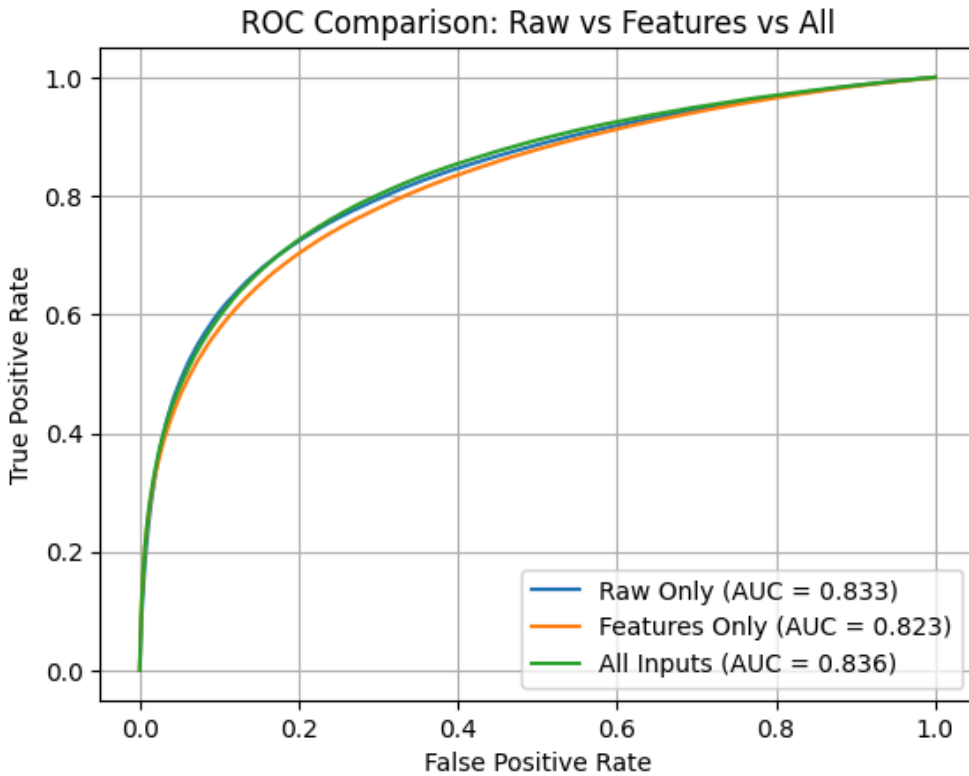
## Exercise 3 - part a



### ### Explanation:

I compare the ROC curves for the training and test datasets. Since both curves overlap almost perfectly and have the same AUC score ( $\sim 0.836$ ), this indicates that the model generalizes well and is not overfitting to the training data.

## Exercise 3 - part b



### ### Explanation:

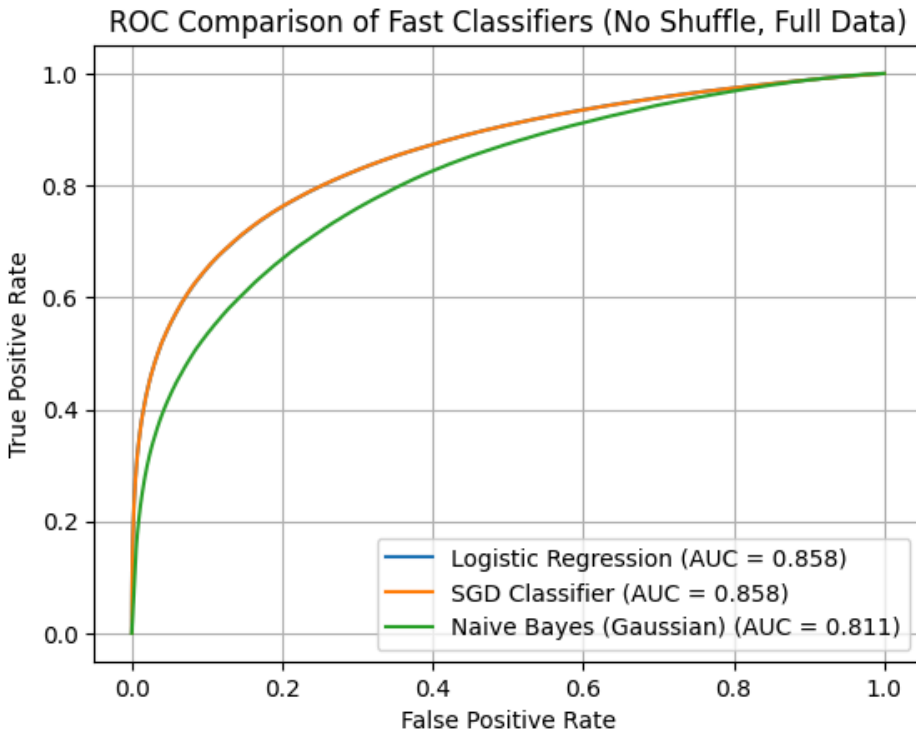
I trained the LDA classifier using three different input sets: raw variables, derived features, and a combination of both. The AUC was highest when using all inputs (0.836), slightly better than using only raw inputs (0.833). The features-only model performed slightly worse (0.823), likely because it lacks some of the low-level information present in the raw inputs.

## Exercise 4

Logistic Regression: Significance  $\sigma = 487.195$ , Ns = 268129, Nb = 34759

SGD Classifier: Significance  $\sigma = 483.994$ , Ns = 262813, Nb = 32045

Naive Bayes (Gaussian): Significance  $\sigma = 452.609$ , Ns = 253626, Nb = 60382



### ### Explanation- part a

In this part, I import and define three fast and commonly used classification algorithms to compare their performance on the SUSY dataset:

- **Logistic Regression** - A linear model widely used as a strong baseline for classification tasks.
- **SGD Classifier** - A linear model trained using Stochastic Gradient Descent. When used with 'log\_loss', it behaves similarly to logistic regression but is highly scalable to large datasets.
- **Gaussian Naive Bayes** - A probabilistic classifier based on applying Bayes' theorem with the assumption of feature independence and Gaussian distribution. It's extremely fast to train and predict.

These models were selected due to their high speed, suitability for large datasets, and their ability to produce probability scores for ROC and AUC evaluation.

### ### Explanation- part b

This section defines the `evaluate_classifier` function, which:

- Trains each classifier using the provided training data.
- Computes prediction scores using either `decision_function` or `predict_proba`, depending on the model.
- Calculates the Receiver Operating Characteristic (ROC) curve, showing the trade-off between true positive rate and false positive rate.
- Computes the Area Under the Curve (AUC) as a single scalar value summarizing classifier performance – the higher the AUC, the better the model is at ranking positive instances higher than negatives.

This function is key to understanding the discriminatory power of each classifier.

### ### Explanation- part c

In this part, I assessed how well each classifier distinguishes signal events from background using a physics-inspired metric called statistical significance.

The goal was not just to evaluate the classifier's accuracy, but to measure how confidently it can identify rare signal events while minimizing false positives from background data.

To do this:

- I converted the predicted scores into binary predictions using a threshold (typically 0.5).

Then, I counted:

- The number of correctly predicted signal events (true positives)
- The number of background events that were incorrectly classified as signal (false positives)
- Based on these values, I computed a significance score, which increases when the model finds more true signals with fewer false detections.

This metric is especially useful in scientific applications, such as particle physics, where the challenge is to detect a small number of important signal events hidden within a much larger background.

For each classifier, I printed:

- The significance score
- The number of signal events it correctly identified
- The number of background events it misclassified

This approach complements the ROC and AUC evaluation by adding a more domain-specific perspective on model performance.




### ### Final Output

The plot generated compares all three classifiers visually through ROC curves, and prints:

AUC values for quantitative performance comparison.

Significance ( $\sigma$ ) values to assess the classifiers' utility in signal detection contexts.

## Exercise 5 - Part b

	Metric	Value	
0	Accuracy	0.775137	 
1	Precision	0.885241	
2	Recall (TPR)	0.585137	
3	F1 Score	0.704563	
4	FPR	0.064159	
5	AUC	0.858159	
6	Significance $\sigma$	487.194924	
7	Ns	268129.000000	
8	Nb	34759.000000	

### ### Explanation - part b

- Using the Logistic Regression model from Exercise 4, I evaluated several performance metrics on the test dataset.
  - The classifier achieved an AUC of 0.858, indicating strong discriminatory power.
- It also reached a high precision of 0.885, meaning that when it predicts a signal, it is usually correct.
- However, the recall (0.585) was moderate, suggesting that some true signal events were missed.
  - The significance score of 487.2 confirms that the classifier is highly effective at distinguishing signal from background in a high-energy physics context.
  - Overall, the model performs well in terms of both statistical and classification metrics, though improving recall could further enhance its sensitivity to rare signal events.