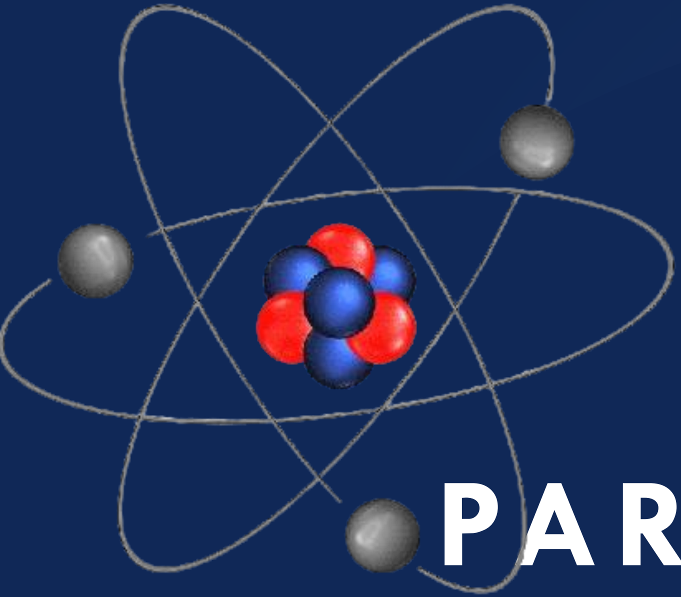
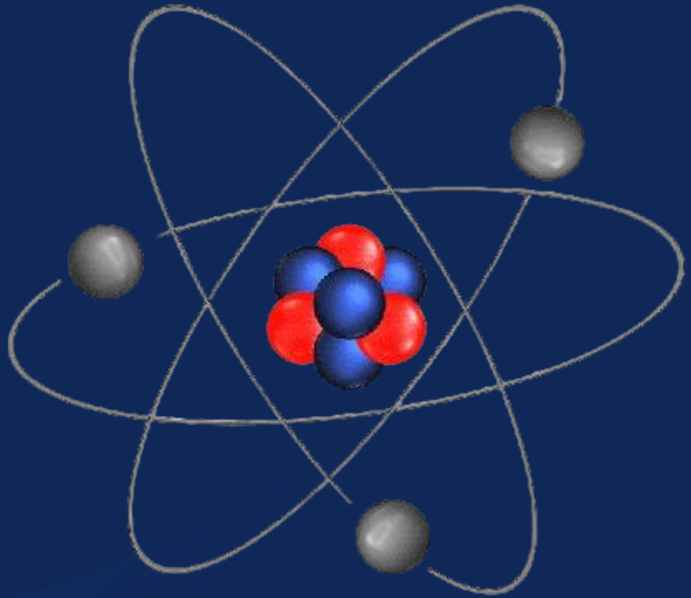


PARTICLE PHYSICS EVENT CLASSIFICATION

By

Chethan



CONTENTS

- Introduction
- Objective
- Data Preprocessing
- Challenges
- Machine Learning Models
- Conclusion

INTRODUCTION

Particle Physics

Particle Physics studies the smallest building blocks of the universe and their interactions

INTRODUCTION

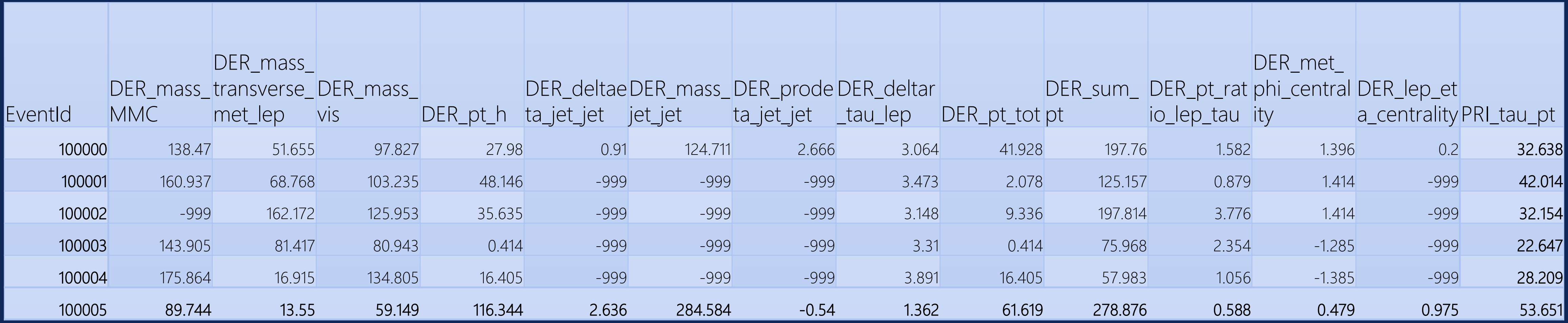
Background

Accurate classification crucial in particle physics for understanding fundamental particles and interactions.

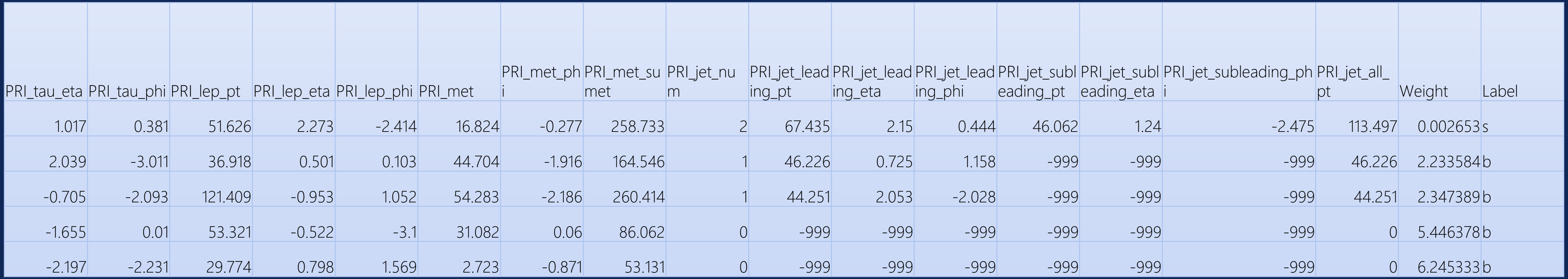
OBJECTIVE

Build a machine-learning model to classify events into Signal (s) and Background (b) categories.

DATASET



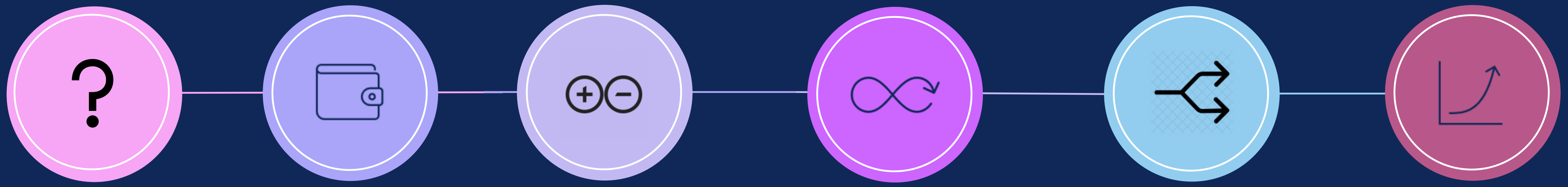
EventId	DER_mass_MMC	DER_mass_transverse_met_lep	DER_mass_vis	DER_pt_h	DER_deltaeta_jet_jet	DER_mass_jet_jet	DER_prodeta_jet_jet	DER_deltar_tau_lep	DER_pt_tot	DER_sum_pt	DER_pt_ratio_lep_tau	DER_met_phi_centrality	DER_lep_eta_centrality	PRI_tau_pt
100000	138.47	51.655	97.827	27.98	0.91	124.711	2.666	3.064	41.928	197.76	1.582	1.396	0.2	32.638
100001	160.937	68.768	103.235	48.146	-999	-999	-999	3.473	2.078	125.157	0.879	1.414	-999	42.014
100002	-999	162.172	125.953	35.635	-999	-999	-999	3.148	9.336	197.814	3.776	1.414	-999	32.154
100003	143.905	81.417	80.943	0.414	-999	-999	-999	3.31	0.414	75.968	2.354	-1.285	-999	22.647
100004	175.864	16.915	134.805	16.405	-999	-999	-999	3.891	16.405	57.983	1.056	-1.385	-999	28.209
100005	89.744	13.55	59.149	116.344	2.636	284.584	-0.54	1.362	61.619	278.876	0.588	0.479	0.975	53.651



PRI_tau_eta	PRI_tau_phi	PRI_lep_pt	PRI_lep_eta	PRI_lep_phi	PRI_met	PRI_met_phi	PRI_met_sumet	PRI_jet_num	PRI_jet_leading_pt	PRI_jet_leading_eta	PRI_jet_leading_phi	PRI_jet_subleading_pt	PRI_jet_subleading_eta	PRI_jet_subleading_phi	PRI_jet_all_pt	Weight	Label
1.017	0.381	51.626	2.273	-2.414	16.824	-0.277	258.733	2	67.435	2.15	0.444	46.062	1.24	-2.475	113.497	0.002653	s
2.039	-3.011	36.918	0.501	0.103	44.704	-1.916	164.546	1	46.226	0.725	1.158	-999	-999	-999	46.226	2.233584	b
-0.705	-2.093	121.409	-0.953	1.052	54.283	-2.186	260.414	1	44.251	2.053	-2.028	-999	-999	-999	44.251	2.347389	b
-1.655	0.01	53.321	-0.522	-3.1	31.082	0.06	86.062	0	-999	-999	-999	-999	-999	-999	0	5.446378	b
-2.197	-2.231	29.774	0.798	1.569	2.723	-0.871	53.131	0	-999	-999	-999	-999	-999	-999	0	6.245333	b

No. of columns – 33
No. of rows - 250000

DATA PREPROCESSING



Step 1 ●

Handling Missing Data

Step 2 ●

Data Cleaning

Step 3 ●

Feature Engineering

Step 4 ●

Dealing with Categorical Data

Step 5 ●

Data Splitting

Step 6 ●

Standardization

CHALLENGES & DIFFICULTIES

Imbalanced data

- Target column Data was imbalanced with 70-30 ratio
- Under-sampling method is used to balance the data

Outliers/Error

- -999 value was found in the data set which was considered as missing value.
- All features with more than 30% missing value was removed
- Other Few of the outlier value was removed

Overfitting

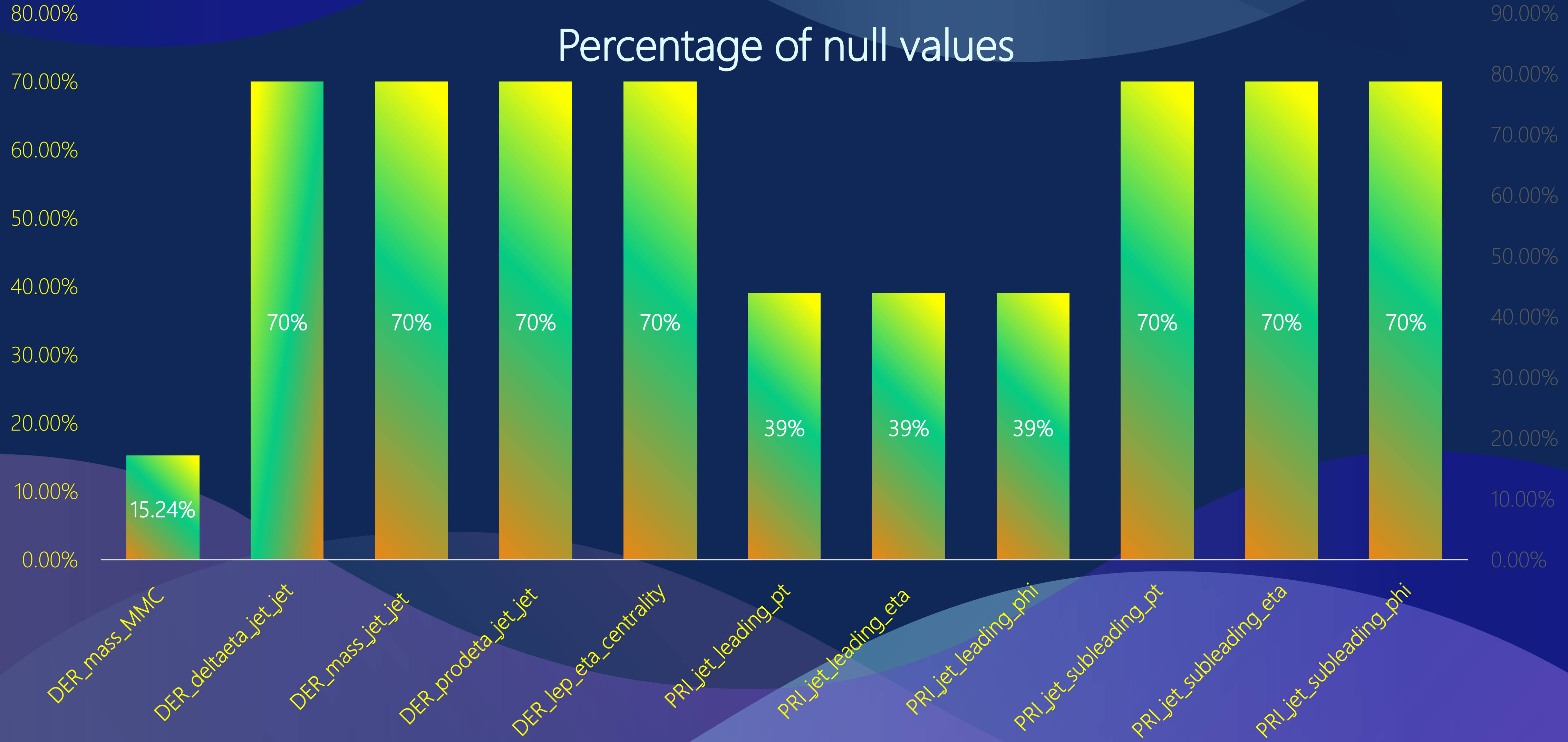
- Almost all the models were overfitting
- L2 regularization and estimators were used to avoid the overfitting

Handling Missing Data

- Identify and handle missing values in the dataset.

Data Cleaning

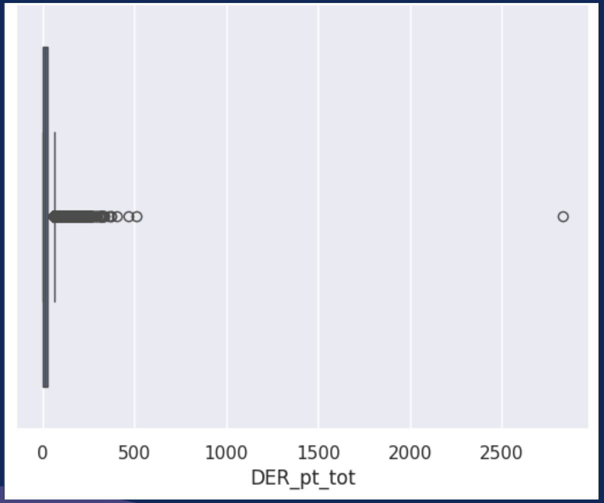
- Address any inconsistencies or errors in the dataset, -999 error found
- Correct or remove outliers that may negatively impact model performance.



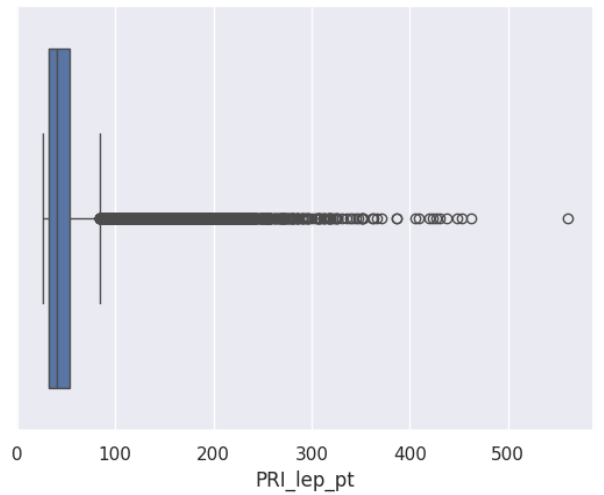
DATA PREPROCESSING

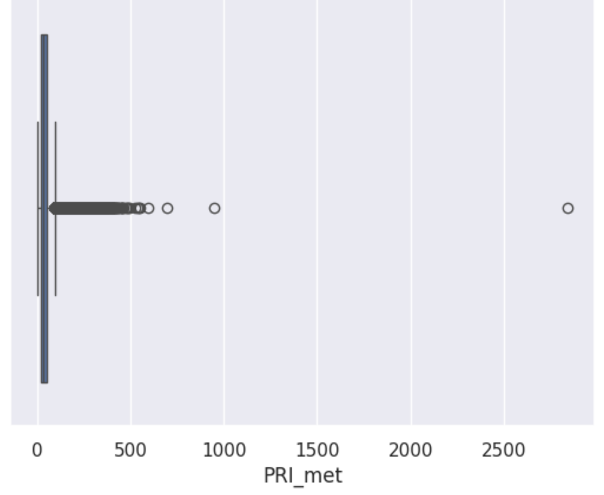
Feature Engineering

- Removing irrelevant or redundant features that do not contribute meaningful information
- Features with more than 30% missing values are dropped and event id feature is dropped.



1. DER_deltaeta_jet_jet
2. DER_mass_jet_jet
3. DER_prodeta_jet_jet
4. DER_lep_eta_centrality
5. PRI_jet_leading_pt
6. PRI_jet_leading_eta
7. PRI_jet_leading_phi
8. PRI_jet_subleading_pt
9. PRI_jet_subleading_eta
10. PRI_jet_subleading_phi
11. EventId





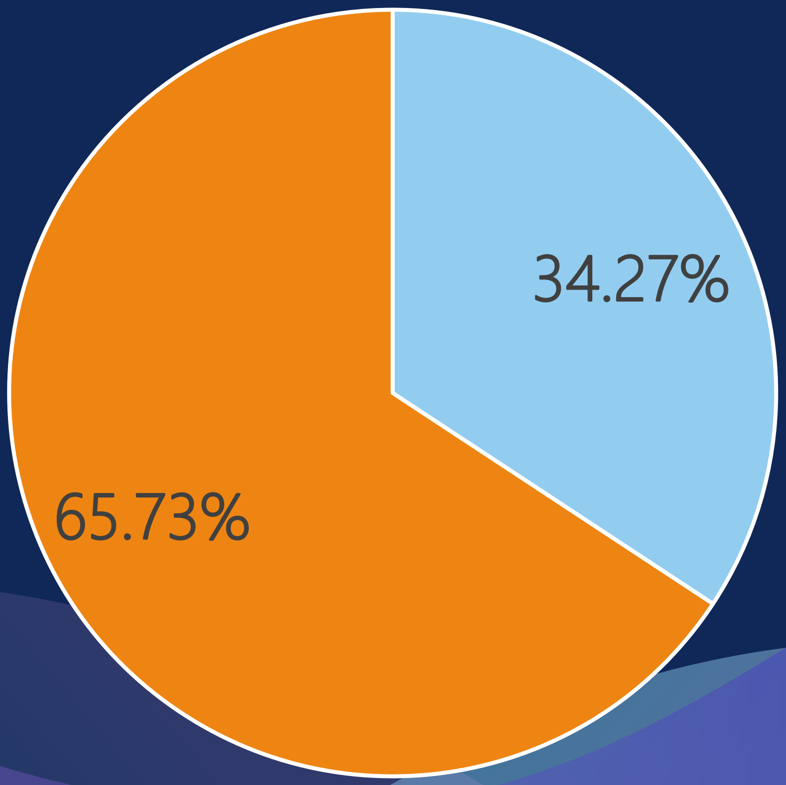
DATA PREPROCESSING

Dealing with Categorical Data

- Converting categorical variables into a format suitable for machine learning models. (Target feature)
- Since data is imbalanced, data is balanced first, using under-sampling method

Target Column - Label

Data is imbalanced



🟦 S =1

🟧 B =0

DATA PREPROCESSING

Data Splitting

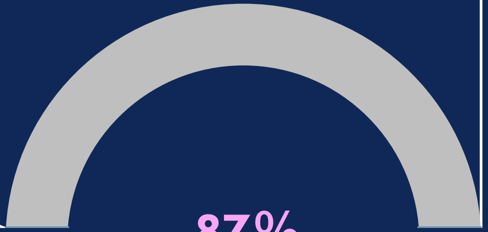
- Divide the dataset into training, and testing sets to evaluate the model's performance on unseen data.
- 80%-20% splitting is done

Normalization/Standardization

- Scale numerical features to a standard range to prevent certain features from dominating others
- Standard scaler is used in this dataset

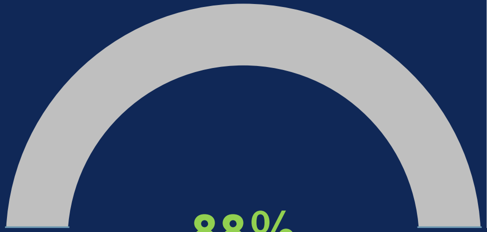
MACHINE LEARNING MODELS

Logistic Regression



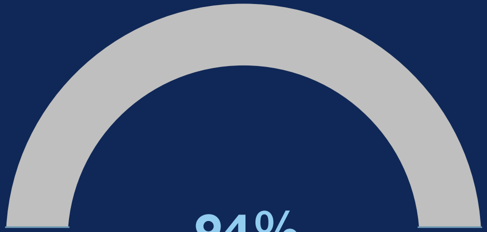
87% Accuracy

Kneighbors Classifier



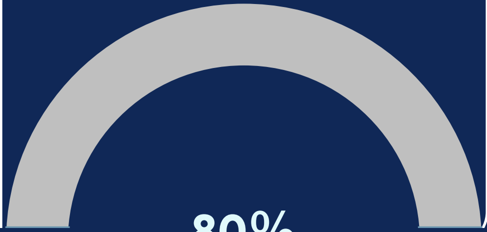
88% Accuracy

XGBoost



94% Accuracy

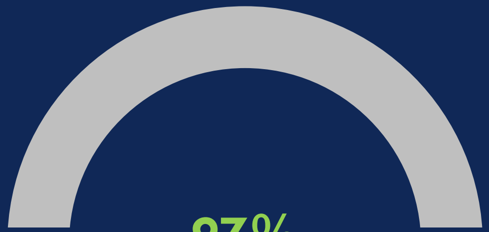
Naive Bayes Theorem



80% Accuracy

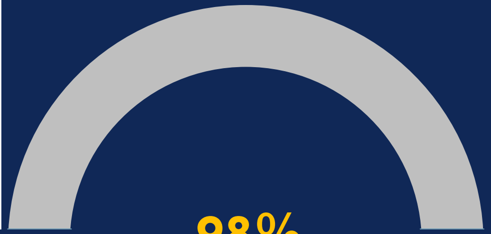
MACHINE LEARNING MODELS

Random Forest Classifier

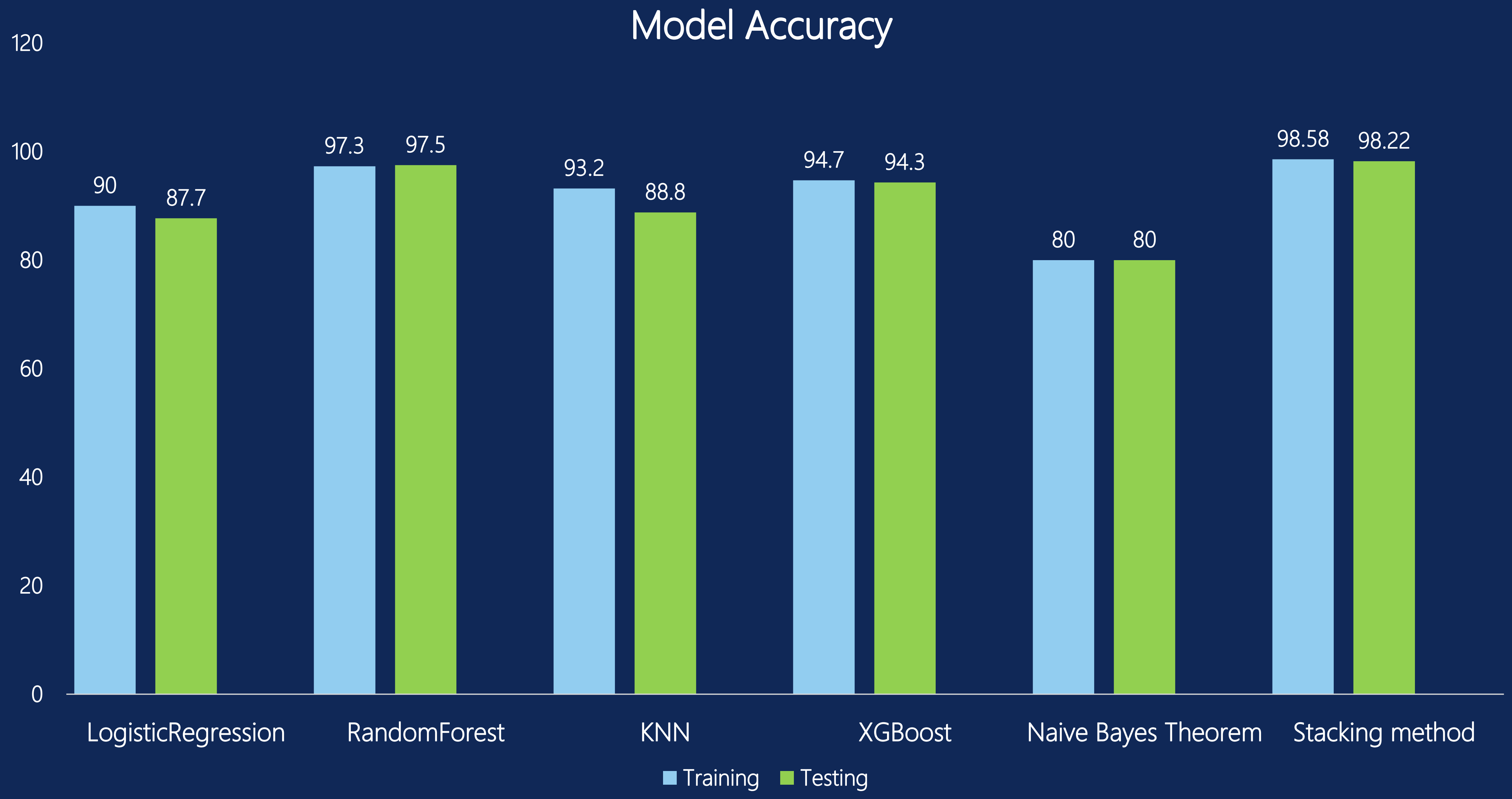


97%
Accuracy

Stacking Method



98%
Accuracy


Model Accuracy
120
100
80
60
40
20
0
90
87.7
97.3
97.5
93.2
88.8
94.7
94.3
80
80
98.58
98.22
LogisticRegression
RandomForest
KNN
XGBoost
Naive Bayes Theorem
Stacking method
Training
Testing

FINAL STACKING MODEL

Cross-Validation Scores (Accuracy): [0.981349 0.98116651 0.98222498 0.97992481 0.98160383]

Mean Cross-Validation Accuracy: 98.13%

Standard Deviation of Cross-Validation Accuracy: 0.08%

CONCLUSION

In this physics particle event classification project utilizing machine learning algorithms, we applied a range of models to predict and classify events into signal and background categories. The accuracy results obtained from different models provide valuable insights into the effectiveness of each algorithm.

ACKNOWLEDGEMENT

I would like to thank Learnbay for providing this dataset

THANK YOU

Chethan

chethan.kmr07@gmail.com