HIGGS BOSON PARTICLE ANALYSIS

DHRUVIL PATEL (013729625)
SAUMIL SHAH (013761293)
SAURABH MANE (012548094)

INTRODUCTION

The Higgs Boson particle was first discovered by the ATLAS experiment at the Large Hadron Collider, CERN in 2012.

The Higgs Boson decays into two tau particles giving rise to a small signal buried in background noise.

The goal of the project is to improve the classification of the events into Higgs Boson decay signal and background noise.

DATASET INFORMATION

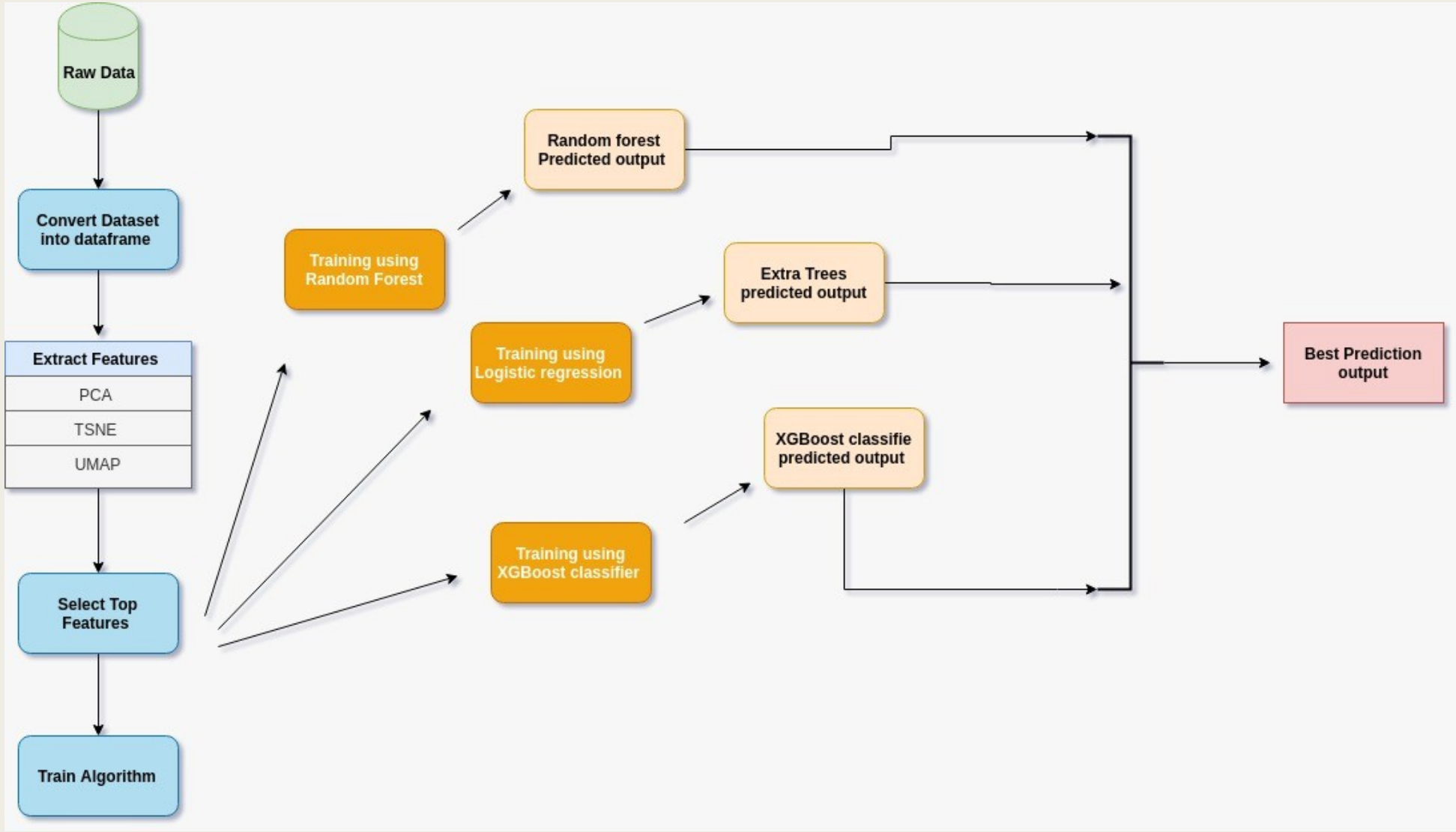
Extracted dataset from ATLAS collaboration (2014). ATLAS Higgs Boson Machine Learning Challenge 2014. CERN Open Data Portal.

Dataset contains 35 features and 818238 unique instances.

The dataset is a combination of categorical and numerical values.

The dataset contains the properties of particles.

PROJECT WORKFLOW



SOLUTIONS

ALGORITHM 1 – LOGISTIC REGRESSION

- WHY?
  - *Gives unbiased results, with low variance.*
- PARAMETERS –
  - max_iter = 100

ALGORITHM 2 – RANDOM FOREST CLASSIFIER

- WHY?
  - *Handles intensive calculations.*
  - *Ensures all features contributes to the model.*
- PARAMETERS –
  - n_estimators = 300

SOLUTIONS

ALGORITHM 3 – XGBOOST CLASSIFIER

- WHY?
  - *Faster than other implementations of gradient boosting.*
- PARAMETERS –
  - *learning_rate = 0.1*

DATA PREPROCESSING

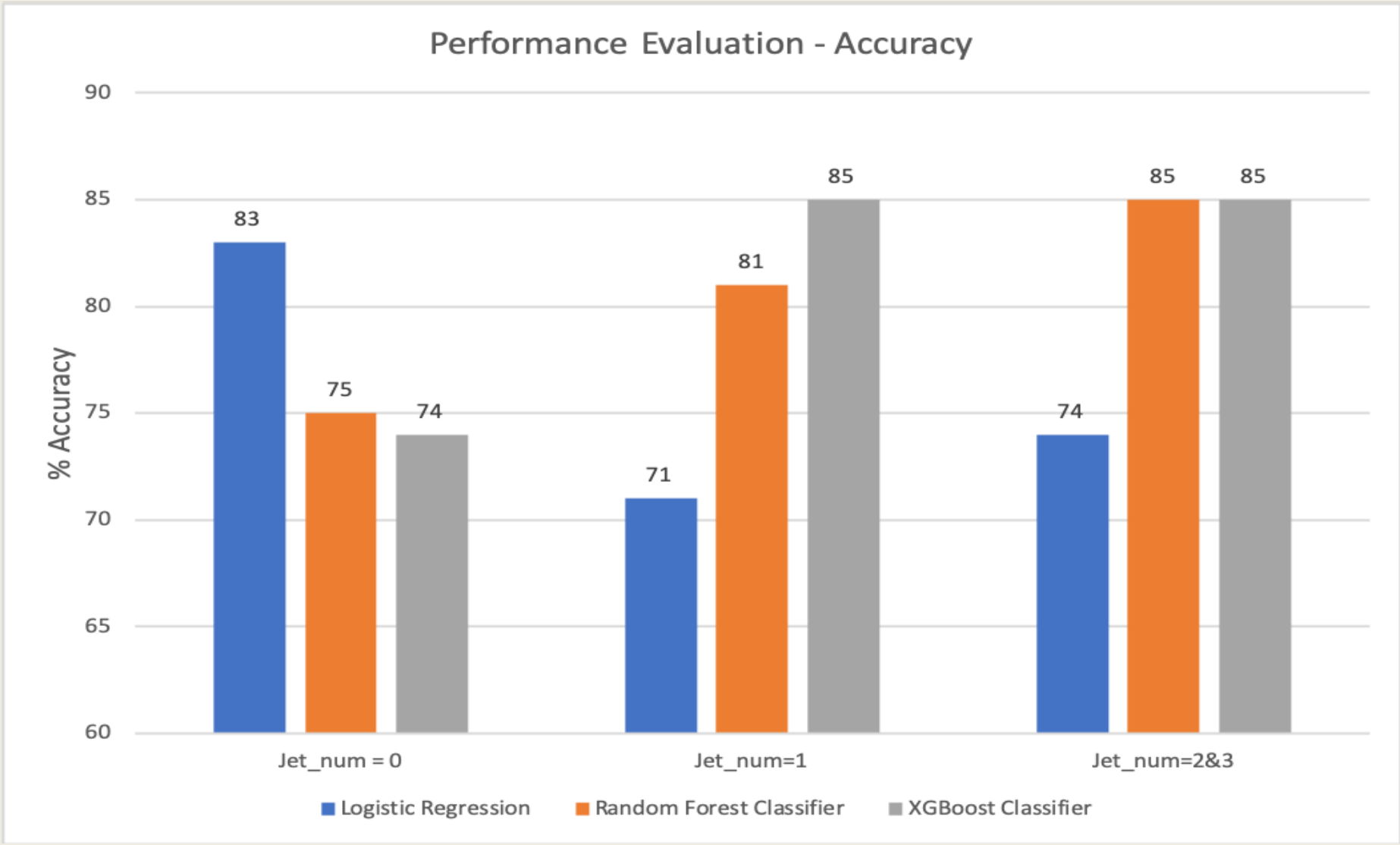
Grouping the data according to the PRI_jet_num to remove null value columns.

Applied PCA to extract new features and reduce the dimensionality.

Applied Robust Scaler to handle outliers in the data.

Applied Standard Scaler as a preprocessing step for PCA.

RESULT


Performance Evaluation - Accuracy
% Accuracy
90
85
80
75
70
65
60
83
75
74
71
81
85
74
85
85
Jet_num = 0
Jet_num=1
Jet_num=2&3
Logistic Regression
Random Forest Classifier
XGBoost Classifier

CHALLENGES

- THINGS THAT WORKED:
  - We divided the dataset into 3 parts, depending on the number of ‘PRI_jet_num’. This gave us good results when we implemented our algorithms.
  - Applying PCA as a preprocessing step for t-SNE saved a lot of time to visualize the data in a lower dimensional space as it reduced the number of features.
  - Applying ensemble methods gave decent results after some preprocessing was applied.
- THINGS THAT DIDN’T WORK:
  - Applying XGBoost Classifier without optimally preprocessing the dataset led to overfitting.
  - UMAP and t-SNE did not give satisfactory results in clustering.
  - Applying SVM on this dataset was very difficult. Training the model on 10% of the data took a very long time and we decided not to implement it on the entire dataset.

CONCLUSION

We understood the importance of data preprocessing and how it is the most important step when applying any algorithm.

Random Forest CLassifier tends to overfit with an increase in estimators.

The XGBoost Classifier gave us the best prediction result.

THANK YOU