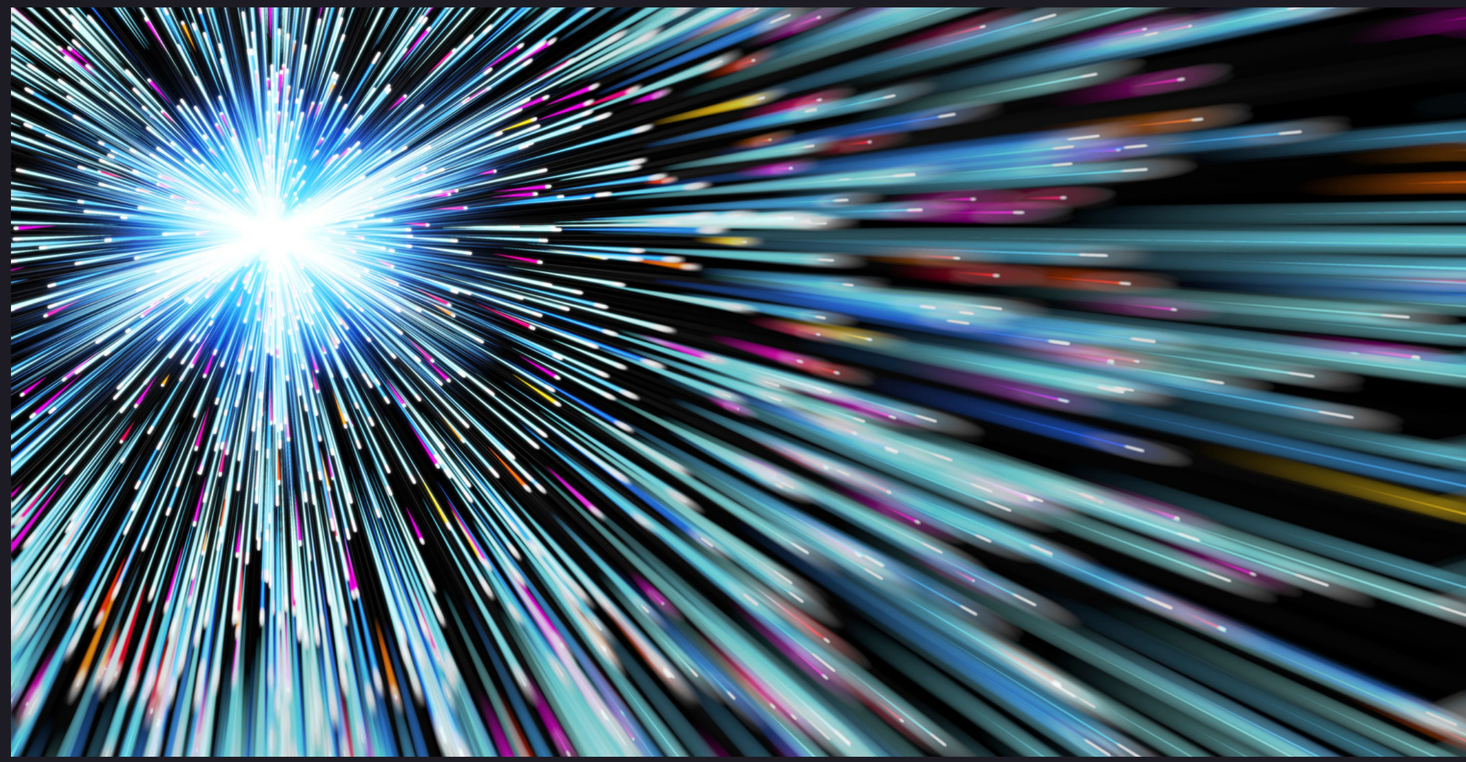
01

In Search of the Higgs Boson

A ML2C1 Project by:
Anshika, Niegil, Sakthisree, Vishnu



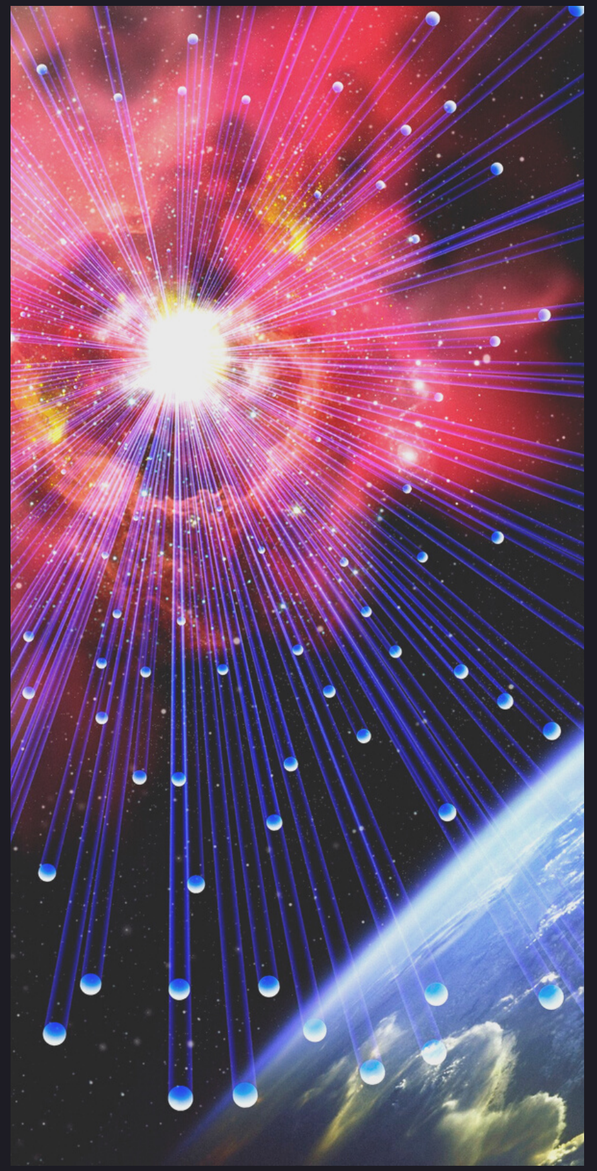
ML2C1 2021

Introduction

The discovery of Higgs particle was announced on 4th July 2012. In 2013, Nobel Prize was conferred upon two scientists, Francois Englert and Peter Higgs for their contribution towards its discovery. A characteristic property of Higgs Boson is its decay into other particles through different processes. At the ATLAS detector at CERN, very high energy protons are accelerated in a circular trajectory in both directions thus colliding with themselves and resulting in hundreds of particles per second.

Goal

By creating a classifier that can improve the procedure that produces the selection region. Given all the features, to predict whether is it Signal or Background.



Data

```
1  train.shape, test.shape
```

```
((250000, 32), (550000, 30))
```

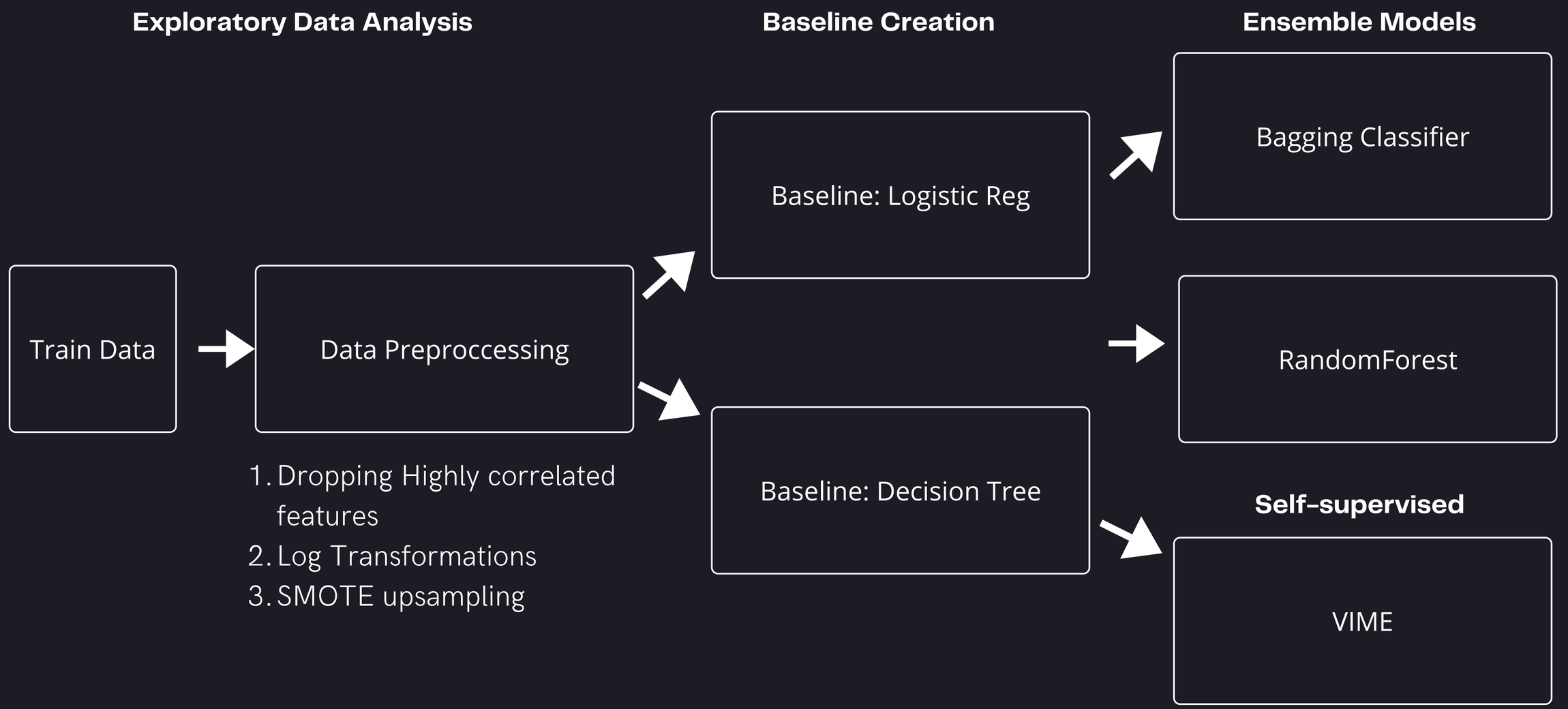
	dtype	nulls	num_uniques	value_counts
DER_mass_MMC	float64	0	108338	HC
DER_mass_transverse_met_lep	float64	0	101637	HC
DER_mass_vis	float64	0	100558	HC
DER_pt_h	float64	0	115563	HC
DER_deltaeta_jet_jet	float64	0	7087	HC
DER_mass_jet_jet	float64	0	68366	HC
DER_prodeta_jet_jet	float64	0	16593	HC
DER_deltar_tau_lep	float64	0	4692	HC
DER_pt_tot	float64	0	59042	HC
DER_sum_pt	float64	0	156098	HC
DER_pt_ratio_lep_tau	float64	0	5931	HC
DER_met_phi_centrality	float64	0	2829	HC
DER_lep_eta_centrality	float64	0	1002	HC
PRI_tau_pt	float64	0	59639	HC
PRI_tau_eta	float64	0	4971	HC
PRI_tau_phi	float64	0	6285	HC
PRI_lep_pt	float64	0	61929	HC
PRI_lep_eta	float64	0	4987	HC
PRI_lep_phi	float64	0	6285	HC
PRI_met	float64	0	87836	HC
PRI_met_phi	float64	0	6285	HC
PRI_met_sumet	float64	0	179740	HC
PRI_jet_num	int64	0	4	0:99913 1:77544 2:50379 3:22164
PRI_jet_leading_pt	float64	0	86590	HC
PRI_jet_leading_eta	float64	0	8558	HC
PRI_jet_leading_phi	float64	0	6285	HC
PRI_jet_subleading_pt	float64	0	42464	HC
PRI_jet_subleading_eta	float64	0	8628	HC
PRI_jet_subleading_phi	float64	0	6286	HC

For the Challenge, we have been provide simulated events using the official ATLAS full detector simulator. The simulator has two parts. In the first, random proton-proton collisions are simulated based on all the knowledge that we have accumulated on particle physics. It reproduces the random microscopic explosions resulting from the proton-proton collisions. In the second part, the resulting particles are tracked through a virtual model of the detector. The process yields simulated events with properties that mimic the statistical properties of the real events with additional information on what has happened during the collision, before particles are measured in the detector

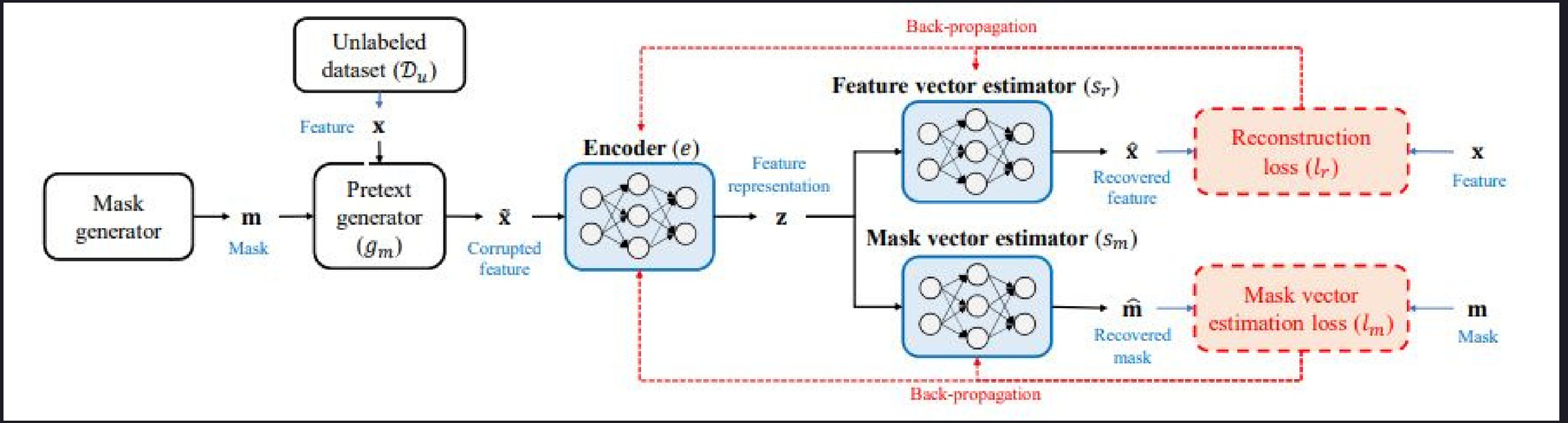
Variable	Description
EventId	An unique integer identifier of the event.
DER_mass_MMC	The estimated mass m_H of the Higgs boson candidate, obtained through a probabilistic phase space integration.
DER_mass_transverse_met_lep	The transverse mass between the missing transverse energy and the lepton.
DER_mass_vis	The invariant mass of the hadronic tau and the lepton.
DER_pt_h	The modulus of the vector sum of the transverse momentum of the hadronic tau, the lepton and the missing transverse energy vector.
DER_deltaeta_jet_jet	The absolute value of the pseudorapidity separation between the two jets (undefined if PRI_jet_num $\leq$ 1).
DER_mass_jet_jet	The invariant mass of the two jets (undefined if PRI_jet_num $\leq$ 1).
DER_prodeta_jet_jet	The product of the pseudorapidities of the two jets (undefined if PRI_jet_num $\leq$ 1).
DER_deltar_tau_lep	The R separation between the hadronic tau and the lepton.
DER_pt_tot	The modulus of the vector sum of the missing transverse momenta and the transverse momenta of the hadronic tau, the lepton, the leading jet (if PRI_jet_num $\geq$) and the subleading jet (if PRI jet num = 2) (but not of any additional jets).
DER_sum_pt	The sum of the moduli of the transverse momenta of the hadronic tau, the lepton, the leading jet (if PRI jet num $\geq$ 1) and the subleading jet (if PRI jet num = 2) and the other jets (if PRI jet num = 3).
DER_pt_ratio_lep_tau	The ratio of the transverse momenta of the lepton and the hadronic tau.
DER_met_phi_centrality	The centrality of the azimuthal angle of the missing transverse energy vector w.r.t. the hadronic tau and the lepton.

PRI_tau_pt	The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the hadronic tau.
PRI_tau_eta	The pseudorapidity η of the hadronic tau.
PRI_tau_phi	The azimuth angle ϕ of the hadronic tau.
PRI_lep_pt	The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the lepton (electron or muon).
PRI_lep_eta	The pseudorapidity η of the lepton.
PRI_lep_phi	The azimuth angle ϕ of the lepton.
PRI_met	The missing transverse energy E_T^{miss}
PRI_met_phi	The azimuth angle ϕ of the mssing transverse energy
PRI_met_sumet	The total transverse energy in the detector.
PRI_jet_num	The number of jets (integer with value of 0, 1, 2 or 3; possible larger values have been capped at 3).
PRI_jet_leading_pt	The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the leading jet, that is the jet with largest transverse momentum (undefined if PRI_jet_num = 0).
PRI_jet_leading_eta	The pseudorapidity η of the leading jet (undefined if PRI jet num = 0).
PRI_jet_leading_phi	The azimuth angle ϕ of the leading jet (undefined if PRI jet num = 0).
PRI_jet_subleading_pt	The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the leading jet, that is, the jet with second largest transverse momentum (undefined if PRI_jet_num $\leq$ 1).
PRI_jet_subleading_eta	The pseudorapidity η of the subleading jet (undefined if PRI_jet_num $\leq$ 1).
PRI_jet_subleading_phi	The azimuth angle ϕ of the subleading jet (undefined if PRI_jet_num $\leq$ 1).
PRI_jet_all_pt	The scalar sum of the transverse momentum of all the jets of the events.
Weight	The event weight w_i
Label	The event label (string) $y_i \in \{s, b\}$ (s for signal, b for background).

APPROACH


Exploratory Data Analysis
Baseline Creation
Ensemble Models
Train Data
Data Preproccessing
1. Dropping Highly correlated features
2. Log Transformations
3. SMOTE upsampling
Baseline: Logistic Reg
Baseline: Decision Tree
Bagging Classifier
RandomForest
Self-supervised
VIME

EXPLORED A NEW CONCEPT - VIME



We introduce two pretext tasks: feature vector estimation and mask vector estimation. Our goal is to optimize a pretext model to recover an input sample (a feature vector) from its corrupted variant, at the same time as estimating the mask vector that has been applied to the sample

- **Mask vector estimator**, $s_m : \mathcal{Z} \to [0,1]^d$, takes $\mathbf{z}$ as input and outputs a vector $\hat{\mathbf{m}}$ to predict which features of $\tilde{\mathbf{x}}$ have been replaced by a noisy counterpart (i.e., $\mathbf{m}$);
- **Feature vector estimator**, $s_r : \mathcal{Z} \to \mathcal{X}$, takes $\mathbf{z}$ as input and returns $\hat{\mathbf{x}}$, an estimate of the original sample $\mathbf{x}$.

The encoder e and the pretext predictive models (in our case, the two estimators s_m and s_r) are trained jointly in the following optimization problem,

$$\min_{e,s_m,s_r} \mathbb{E}_{\mathbf{x}\sim p_X,\mathbf{m}\sim p_\mathbf{m},\tilde{\mathbf{x}}\sim g_m(\mathbf{x},\mathbf{m})}\Big[l_m(\mathbf{m},\hat{\mathbf{m}}) + \alpha \cdot l_r(\mathbf{x},\hat{\mathbf{x}})\Big] \tag{4}$$

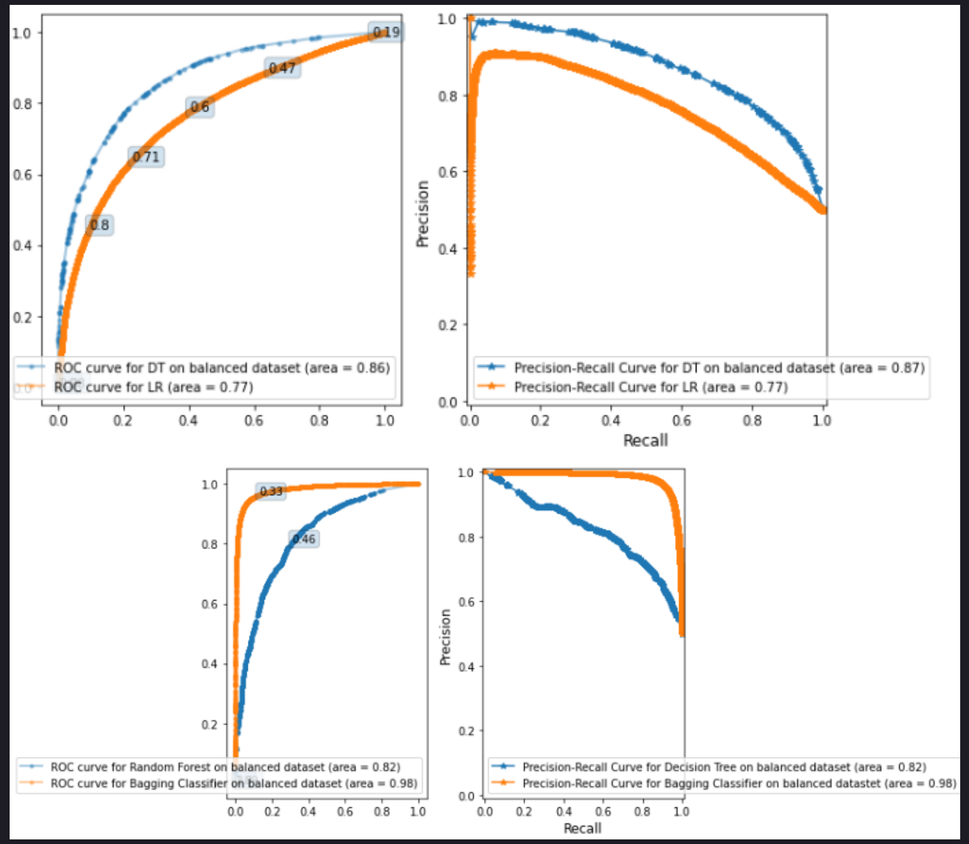
where $\hat{\mathbf{m}} = (s_m \circ e)(\tilde{\mathbf{x}})$ and $\hat{\mathbf{x}} = (s_r \circ e)(\tilde{\mathbf{x}})$. The first loss function l_m is the sum of the binary cross-entropy losses for each dimension of the mask vector[2]:

$$l_m(\mathbf{m},\hat{\mathbf{m}}) = -\frac{1}{d}\Big[\sum_{j=1}^{d} m_j \log\left[(s_m \circ e)_j(\tilde{\mathbf{x}})\right] + (1-m_j)\log\left[1-(s_m \circ e)_j(\tilde{\mathbf{x}})\right]\Big], \tag{5}$$

and the second loss function l_r is the reconstruction loss,

$$l_r(\mathbf{x},\hat{\mathbf{x}}) = \frac{1}{d}\Big[\sum_{j=1}^{d}(x_j - (s_r \circ e)_j(\tilde{\mathbf{x}}))^2\Big]. \tag{6}$$

RESULTS



These are the following accuracies and F1 Score for the various classifiers used:

1. Logistic Regression - 71% Accuracy and 0.70 F1 Score
2. Decision Tree Classifier - 78% Accuracy and 0.78 F1 Score
3. Bagging Classifier - 84% Accuracy and 0.84 F1 Score
4. Random Tree Classifier - 84% Accuracy and 0.84 F1 Score

Clearly, the ensemble models performed bettter. We would like to further our work by spending more time in pre-processing the data and also fine tuning our ensemble methods more.

We would like to further our exploration of the VIME models and see if self-supervision proves to be fruitful in the area of physics.