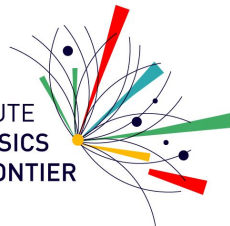




MILLENNIUM INSTITUTE
FOR SUBATOMIC PHYSICS
AT HIGH-ENERGY FRONTIER
SAPHIR



advances in classification of ggF and VBF

J. Tomas Yanez on behalf of Chilean Team

Monday November 20

PREVIOUS CUTS

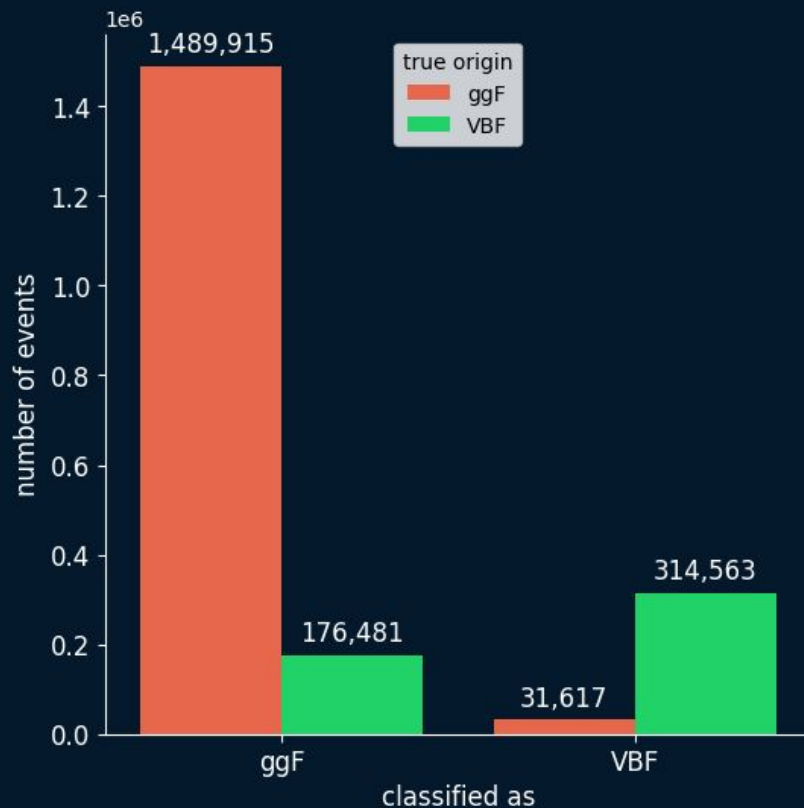
$n_{\text{jet}30} > 1$

$m_{jj} > 1\text{TeV}$

PREVIOUS CUTS

$n_{\text{jet}30} > 1$

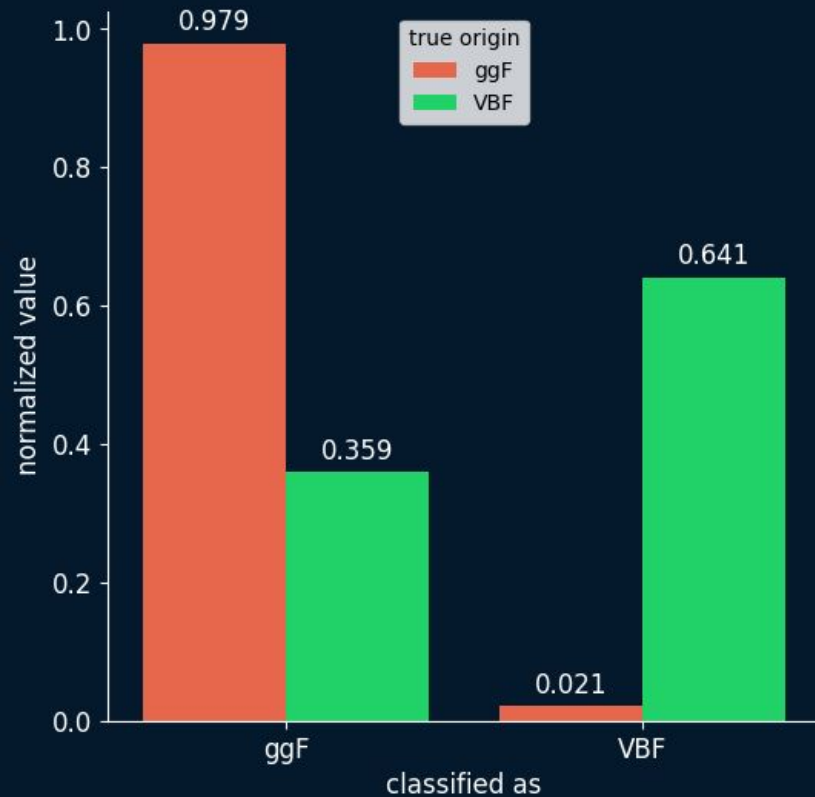
$m_{jj} > 1\text{TeV}$



PREVIOUS CUTS

$n_{\text{jet}30} > 1$

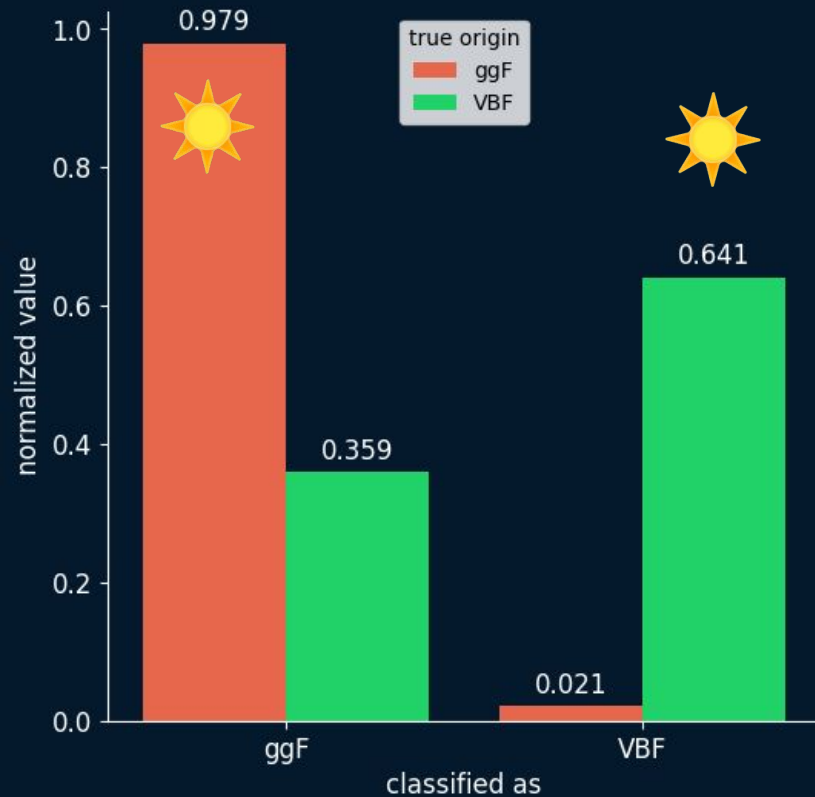
$m_{jj} > 1\text{TeV}$



PREVIOUS CUTS

$n_{\text{jet}30} > 1$

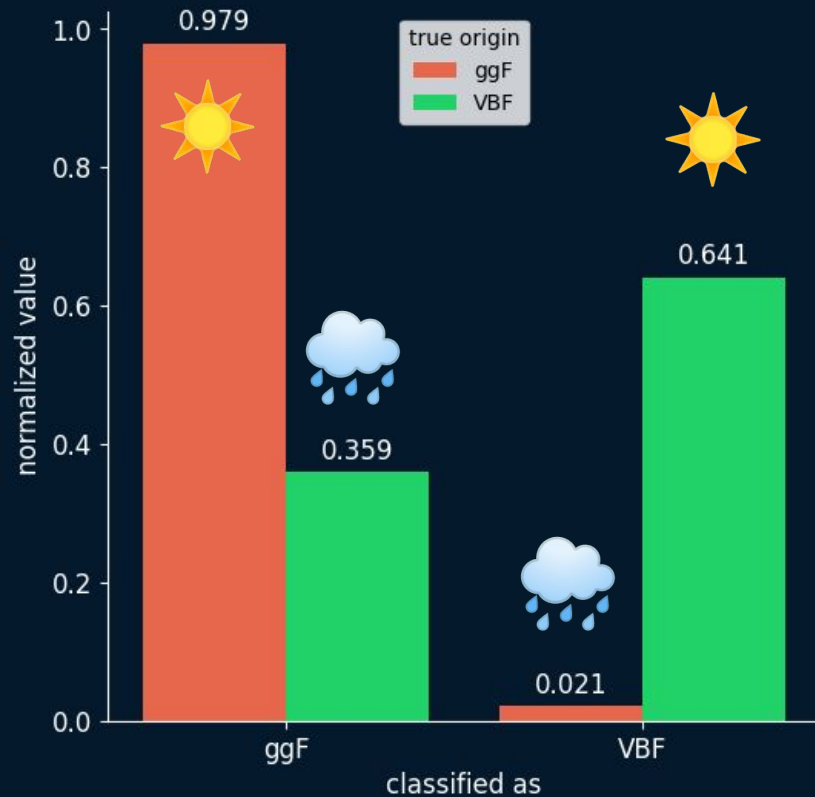
$m_{jj} > 1\text{TeV}$



PREVIOUS CUTS

$n_{\text{jet}30} > 1$

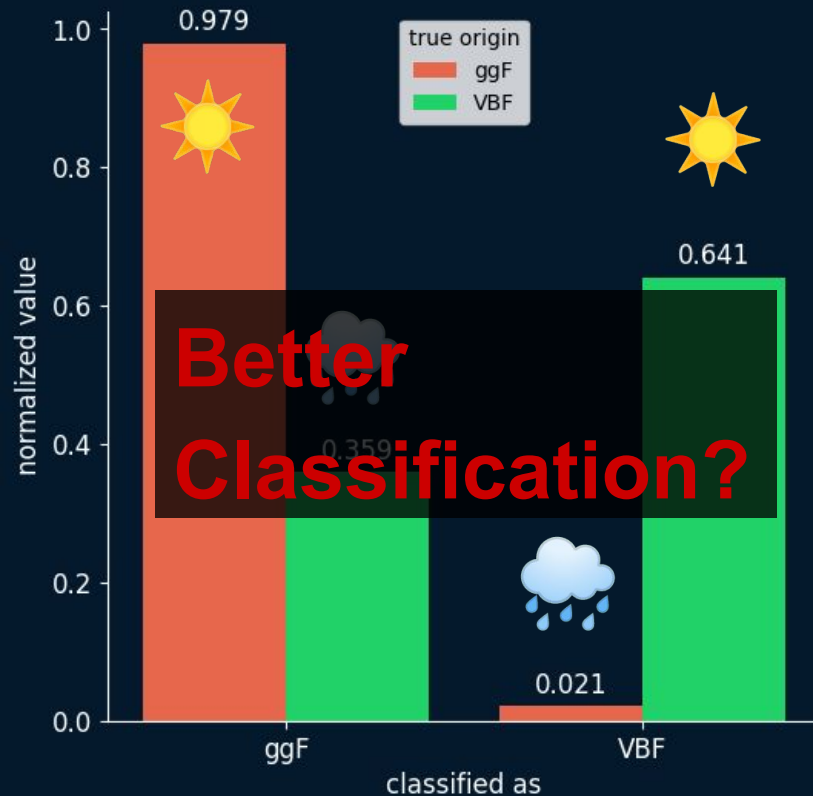
$m_{jj} > 1\text{TeV}$



PREVIOUS CUTS

$n_{\text{jet}30} > 1$

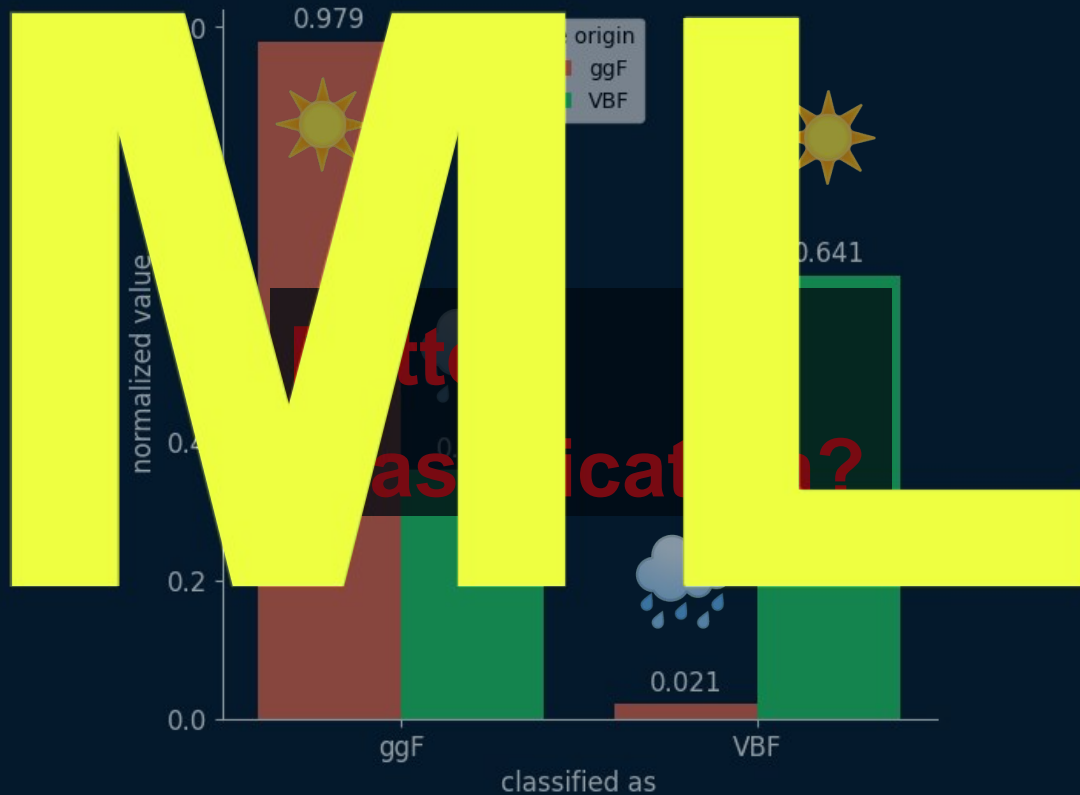
$m_{jj} > 1\text{TeV}$



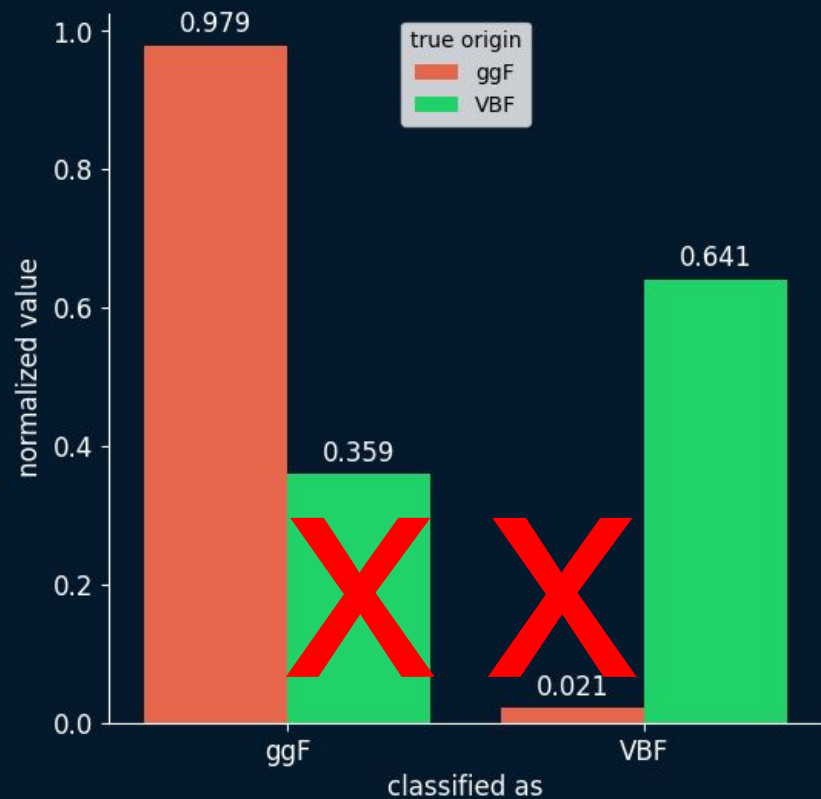
PREVIOUS CUTS

$n_{\text{jet}30} > 1$

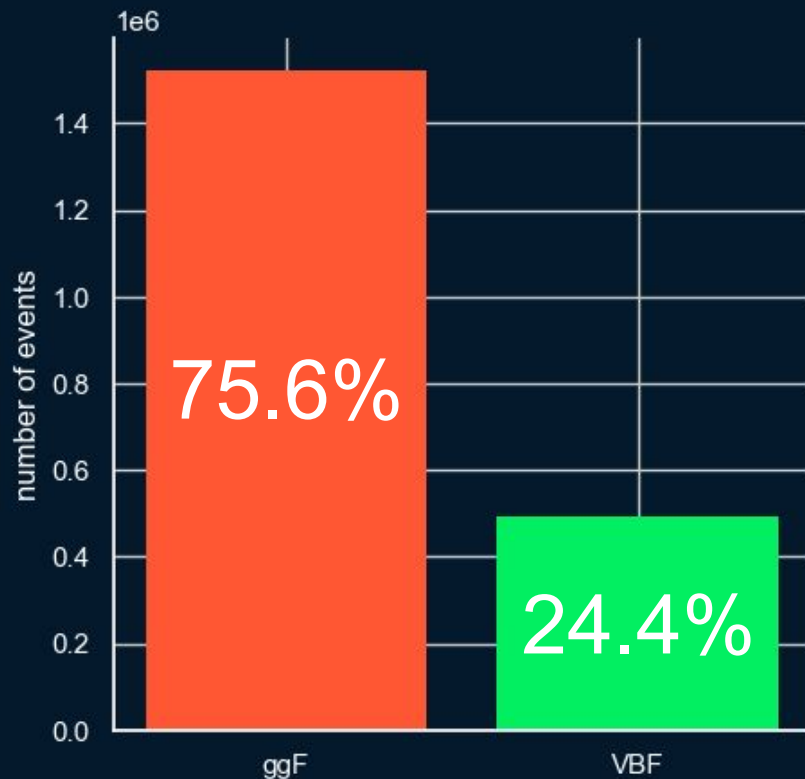
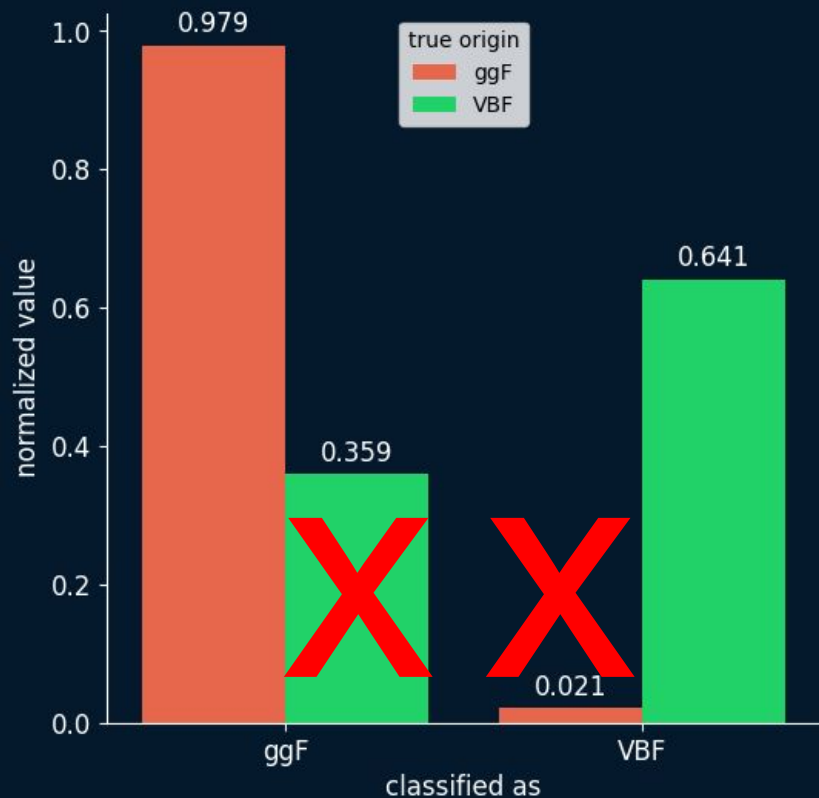
$m_{jj} > 1\text{TeV}$



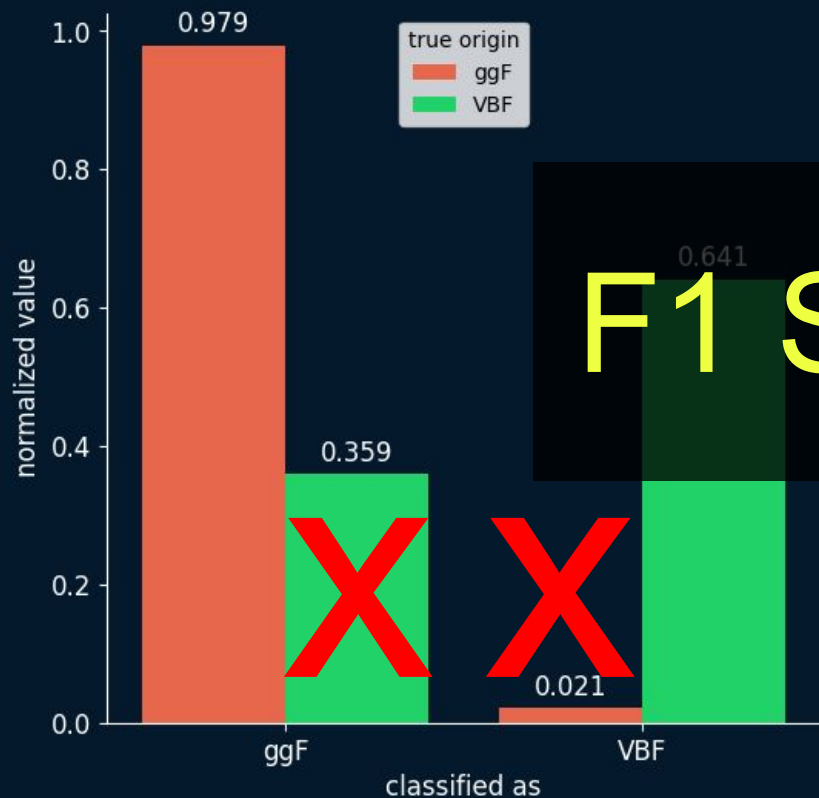
OUR OBJECTIVE



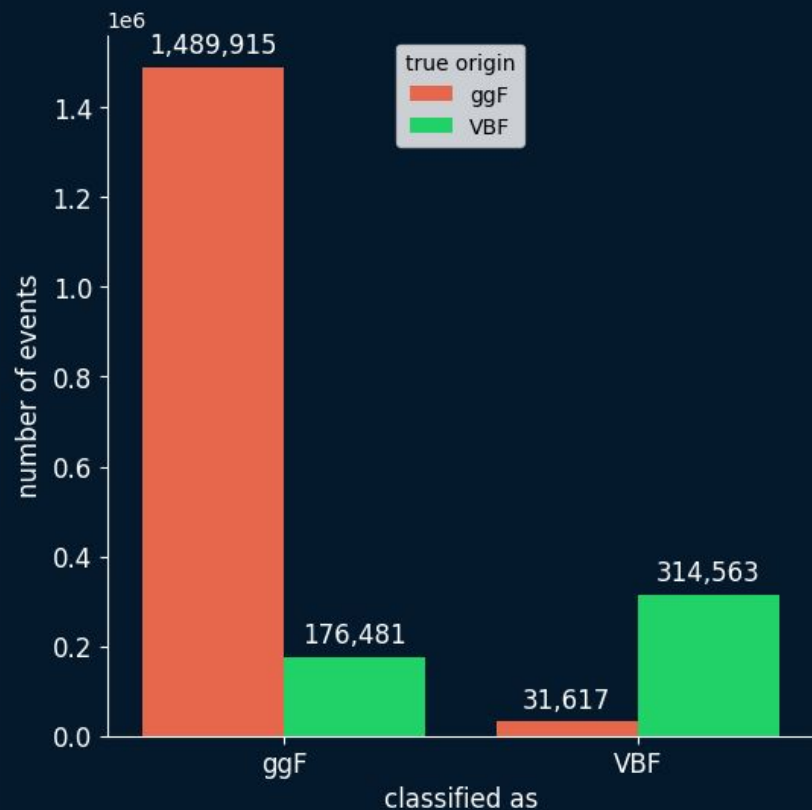
OUR OBJECTIVE



OUR OBJECTIVE

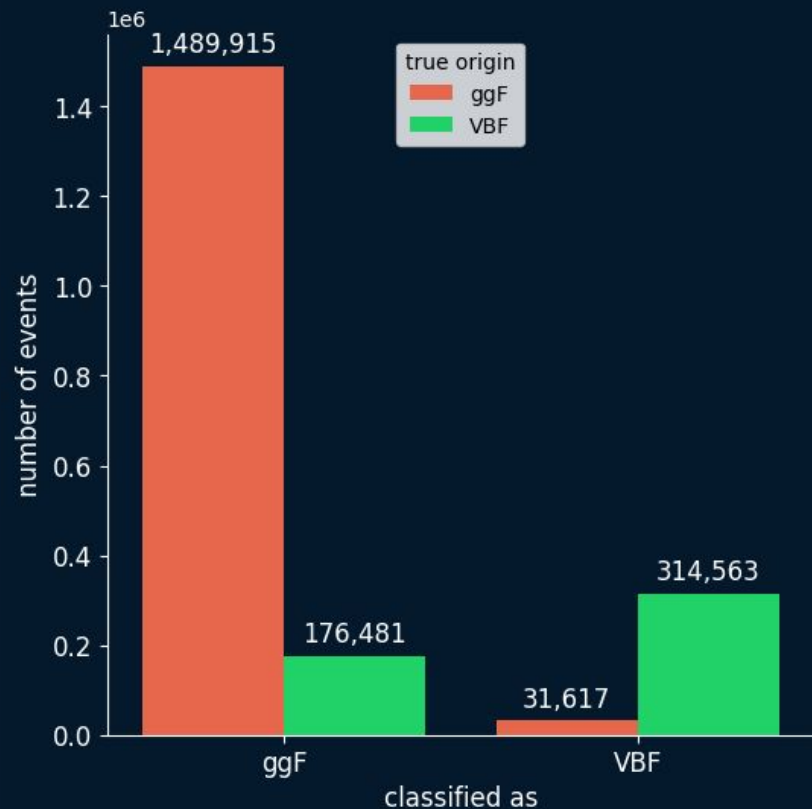


F1 SCORE



$$\frac{VBF_{good} + ggF_{good}}{all} \approx 0.9$$

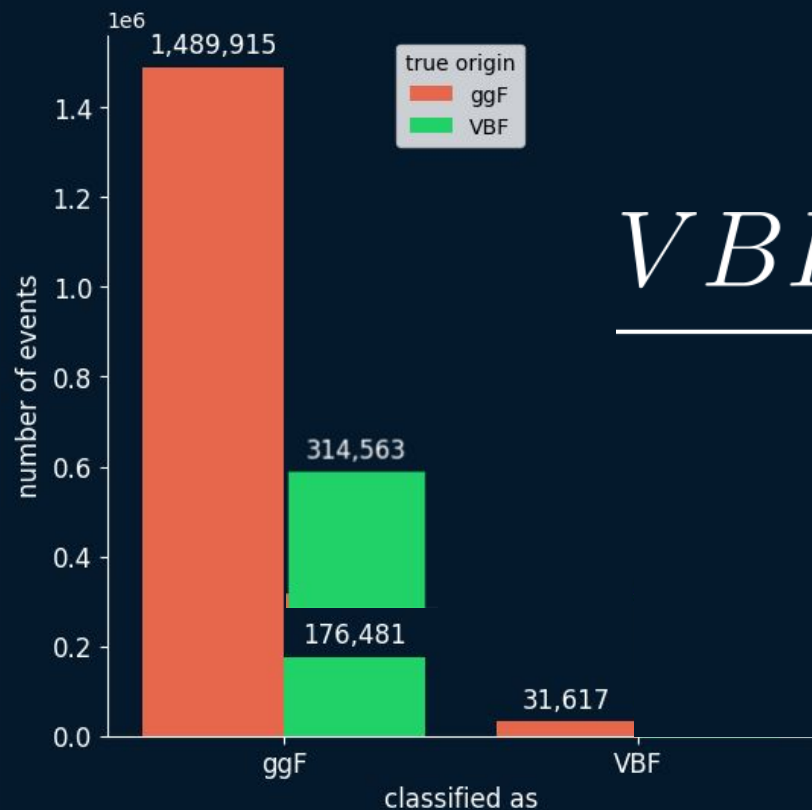
F1 SCORE



$$\frac{VBF_{good} + ggF_{good}}{all} \approx 0.9$$

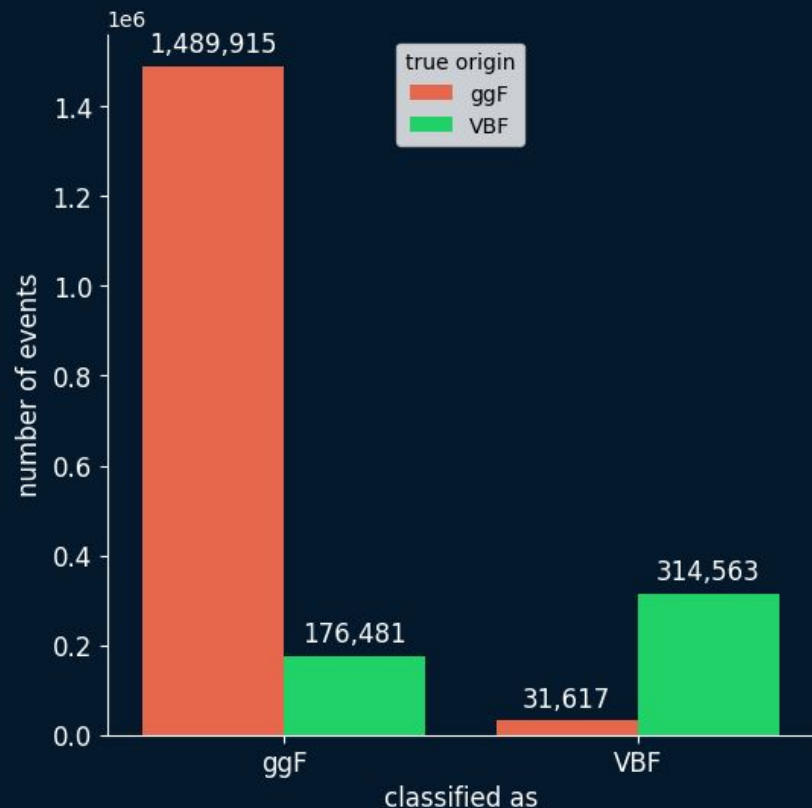
but VBF got 36% of error

F1 SCORE



$$\frac{VBF_{good} + ggF_{good}}{all} \approx 0.74$$

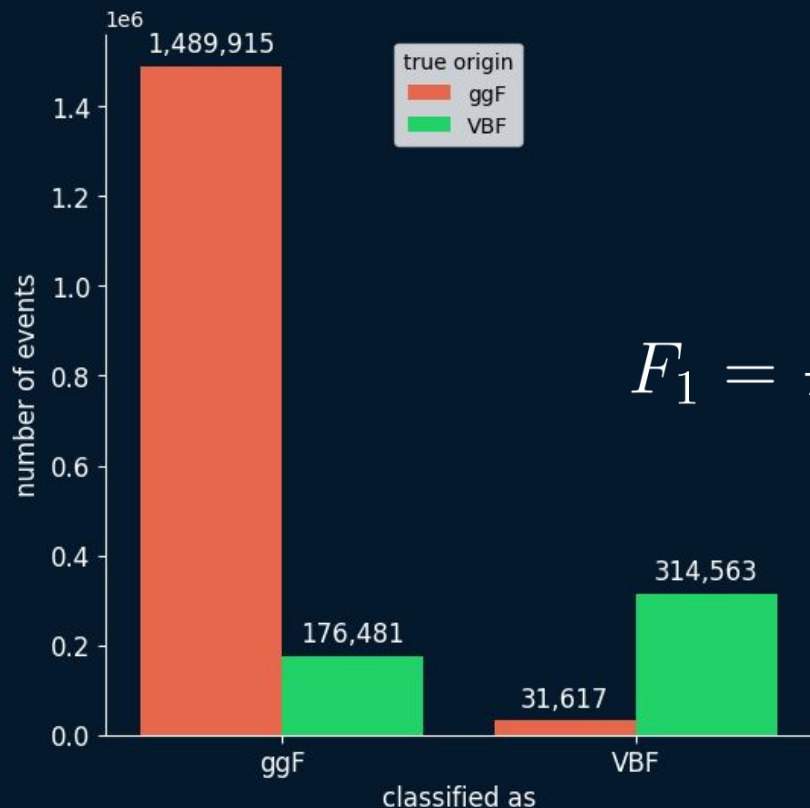
F1 SCORE



$$\frac{VBF_{good} + ggF_{good}}{all} \approx 0.9$$

but VBF got 36% of error

F1 SCORE



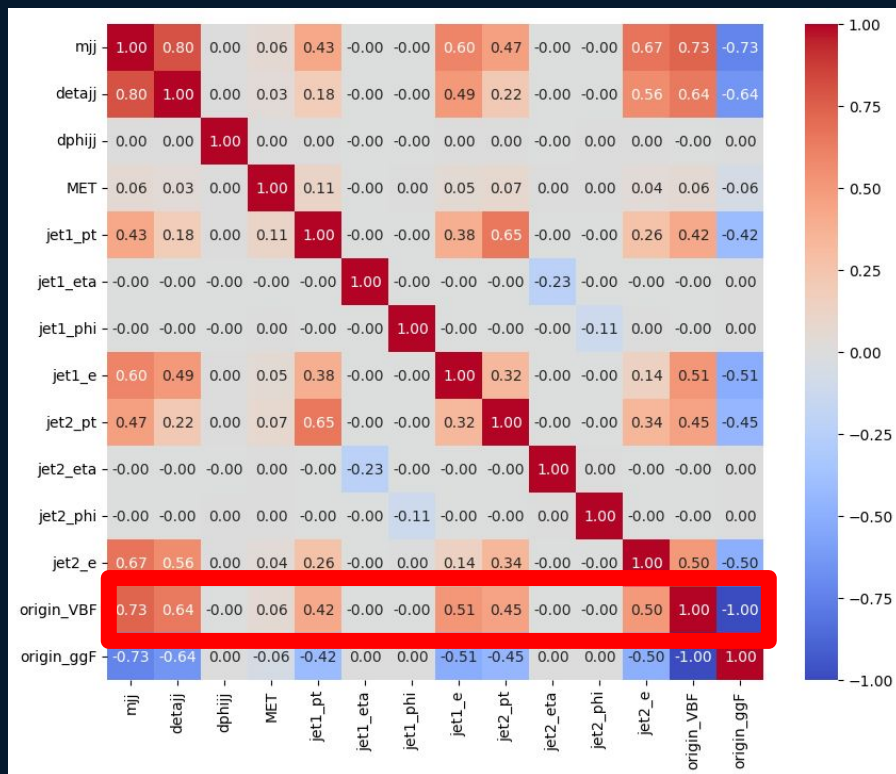
$$\frac{VBF_{good} + ggF_{good}}{all} \approx 0.9$$

but VBF got 36% of error

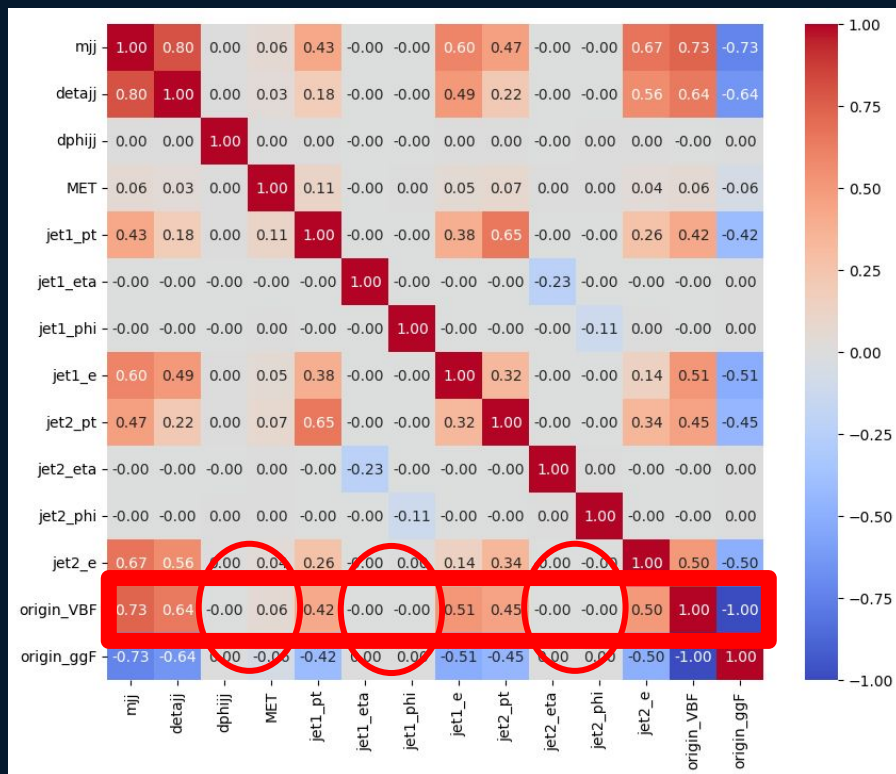
$$F_1 = \frac{VBF_{good}}{VBF_{good} + 0.5(VBF_{bad} + ggF_{bad})}$$

$$F_1 = 0.75$$

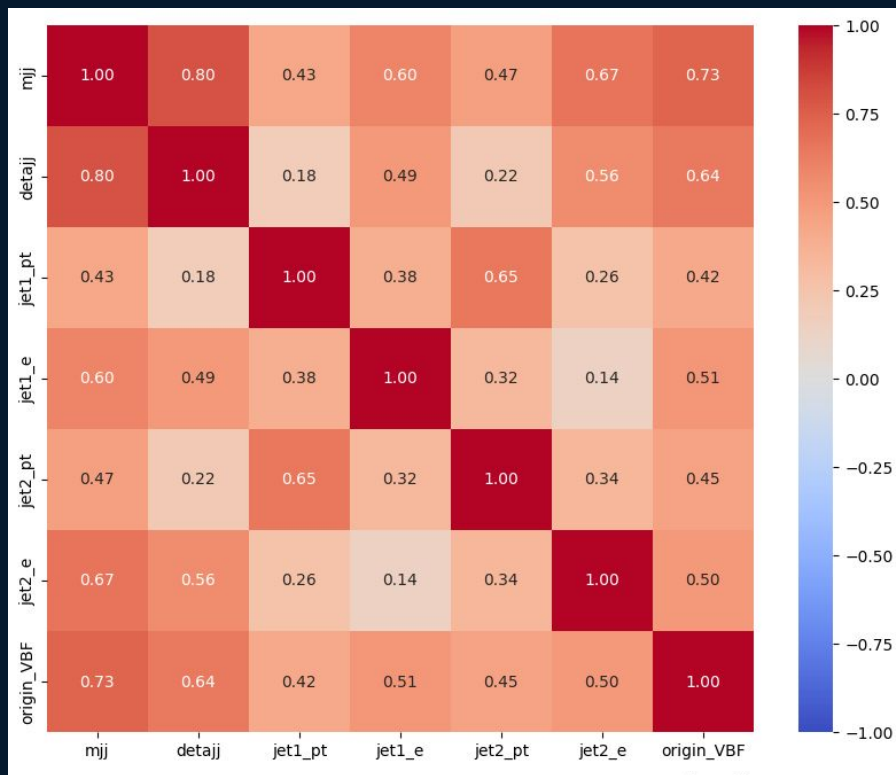
CORRELATION



CORRELATION

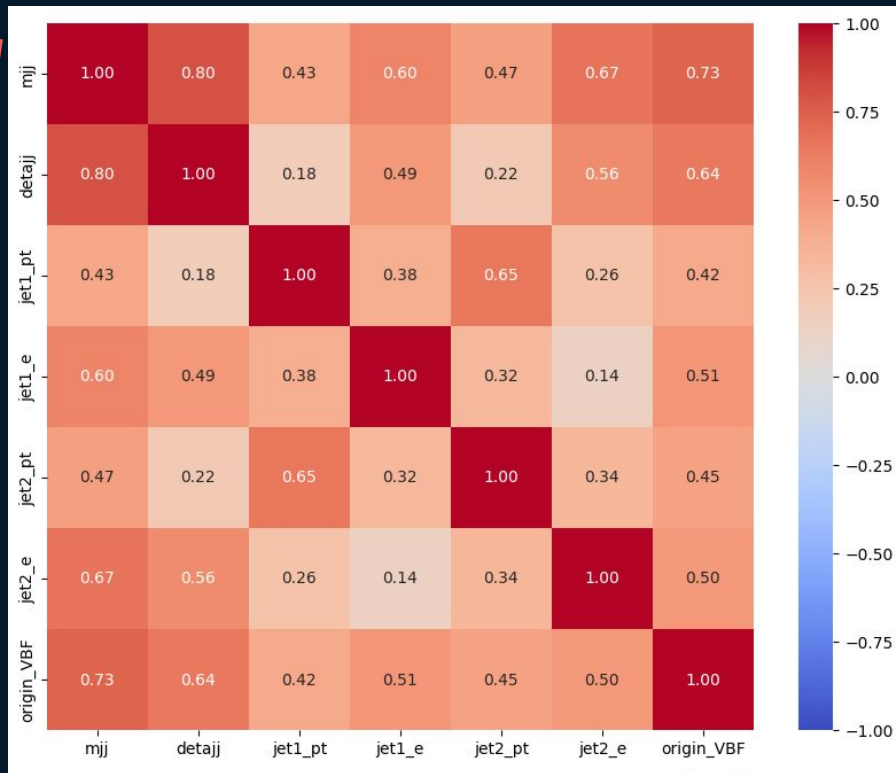


CORRELATION

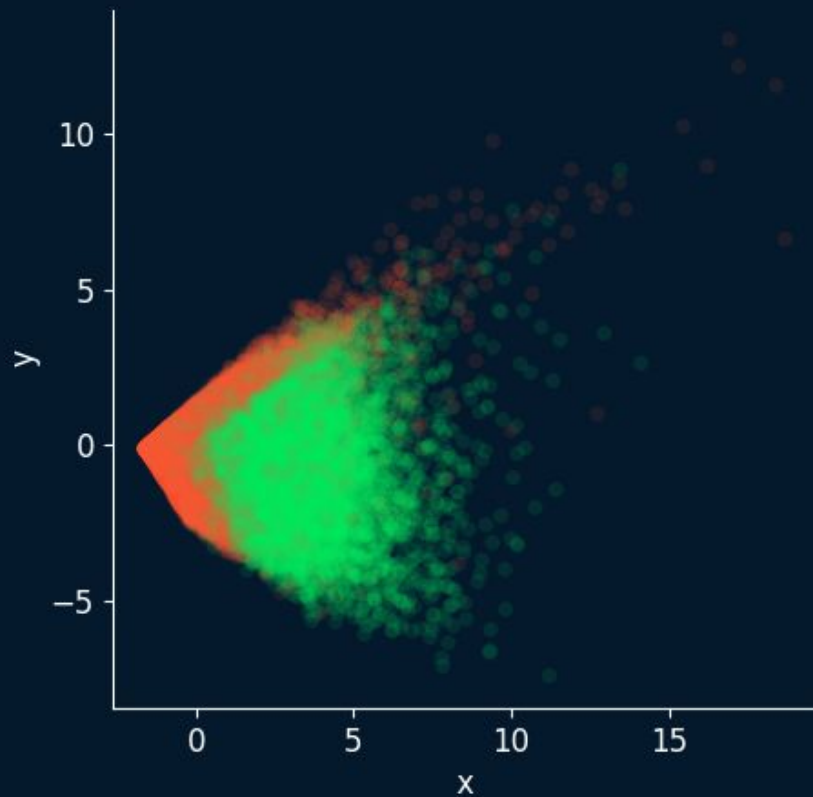


CORRELATION

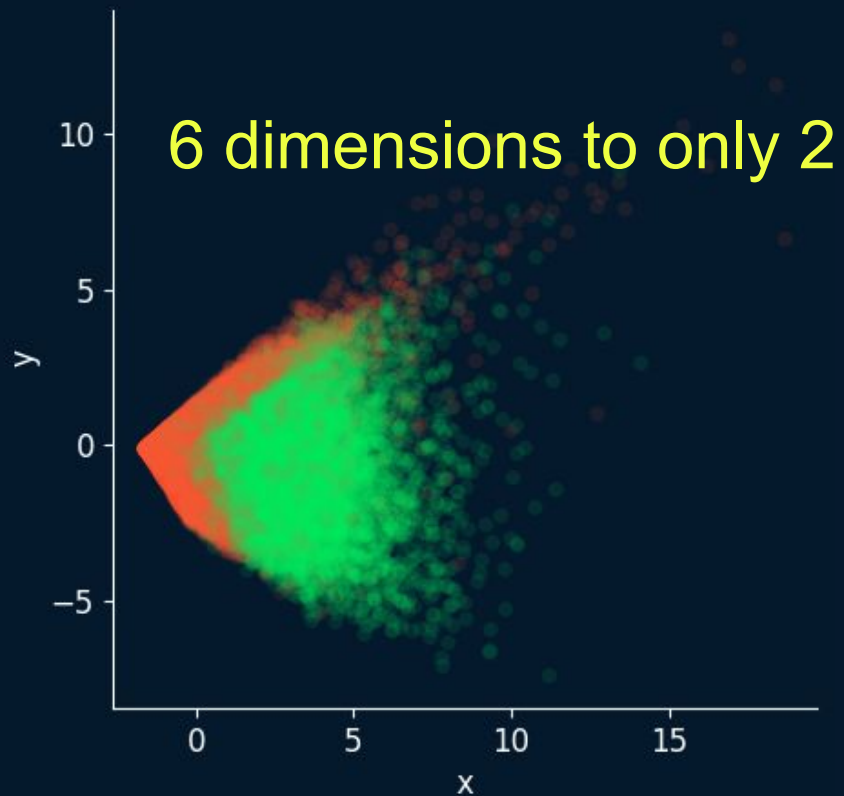
6 variables!



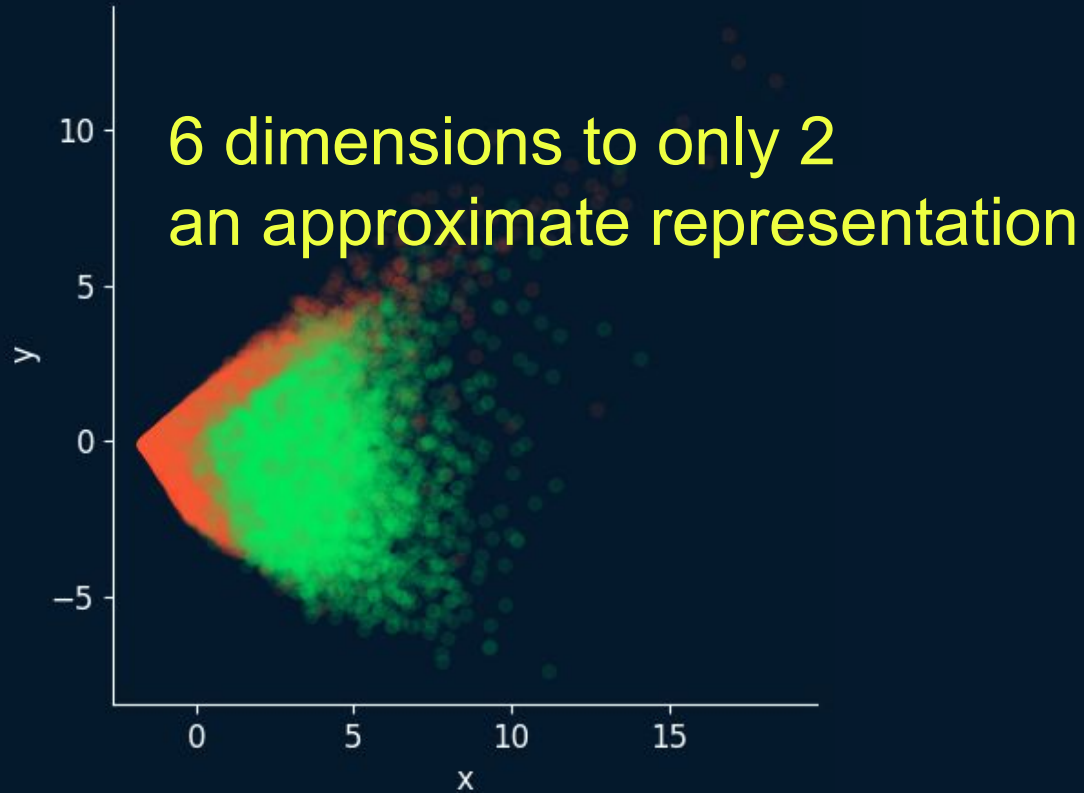
PCA REPRESENTATION



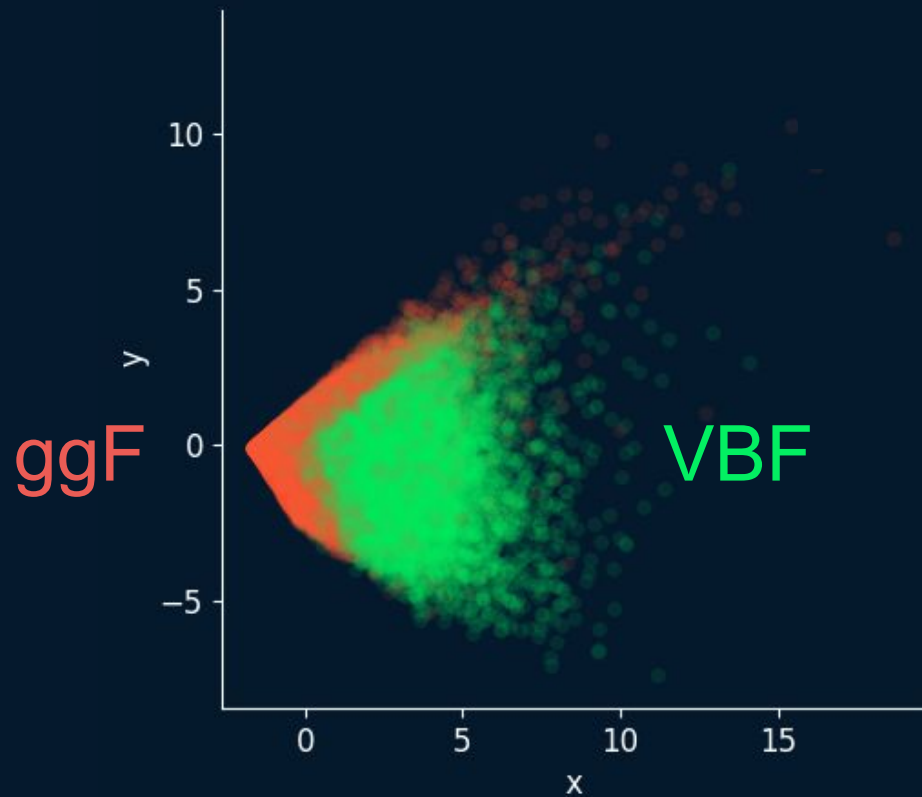
PCA REPRESENTATION



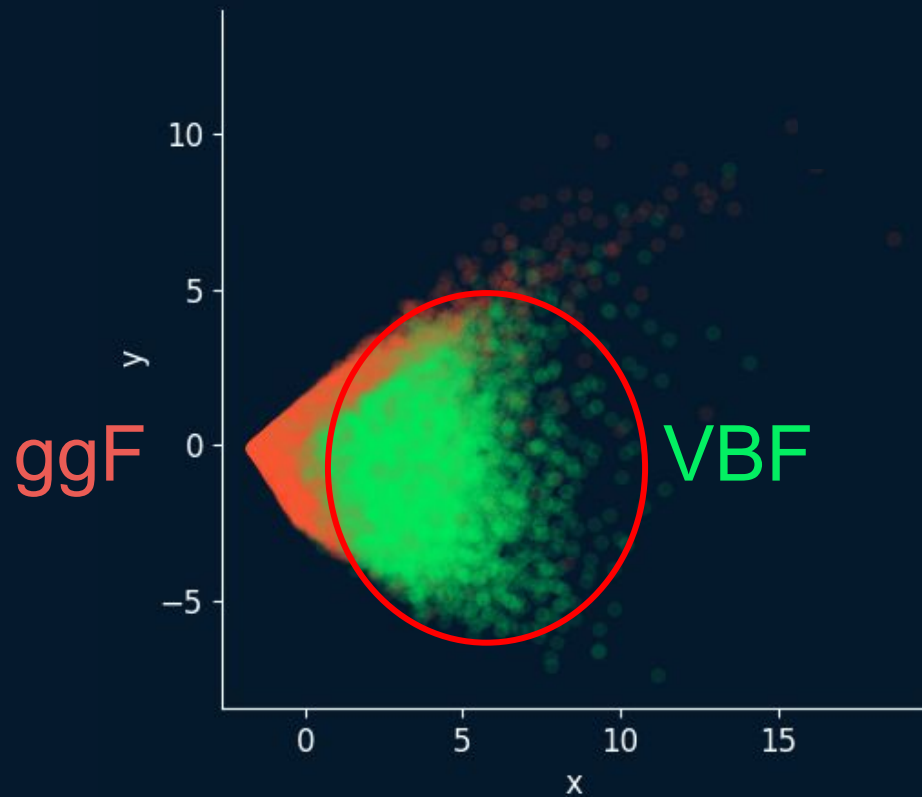
PCA REPRESENTATION



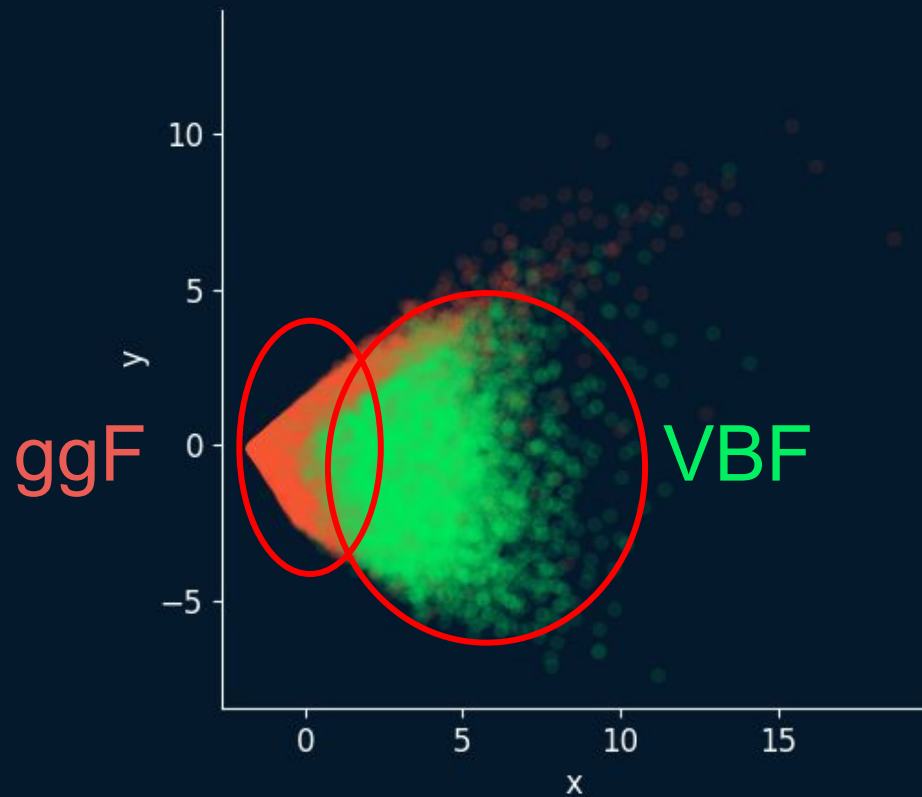
PCA REPRESENTATION



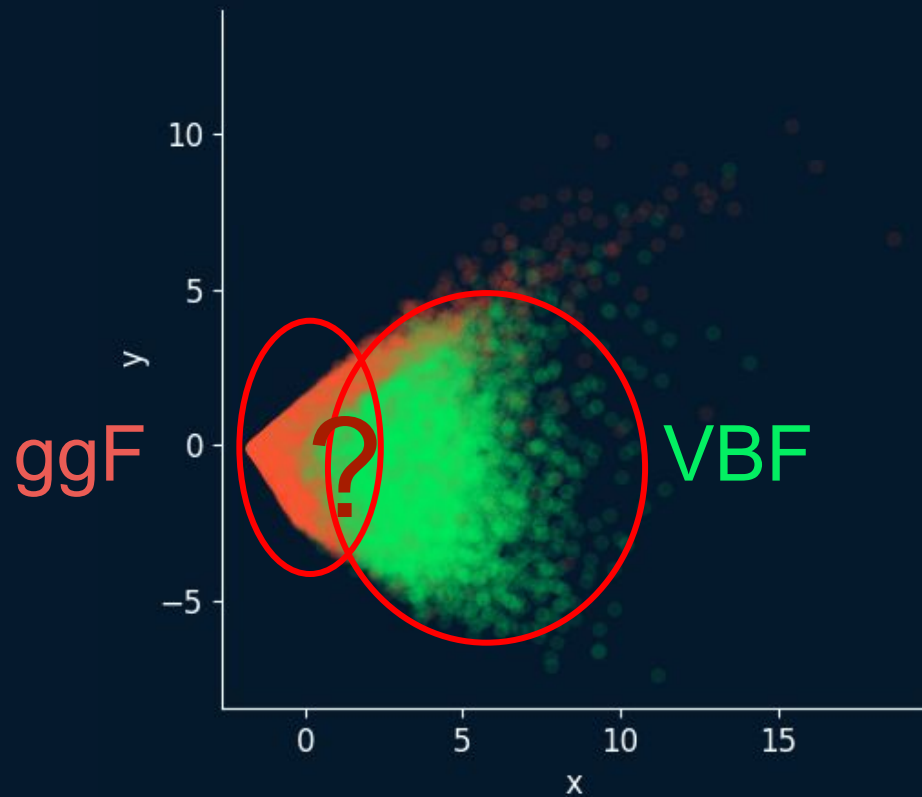
PCA REPRESENTATION



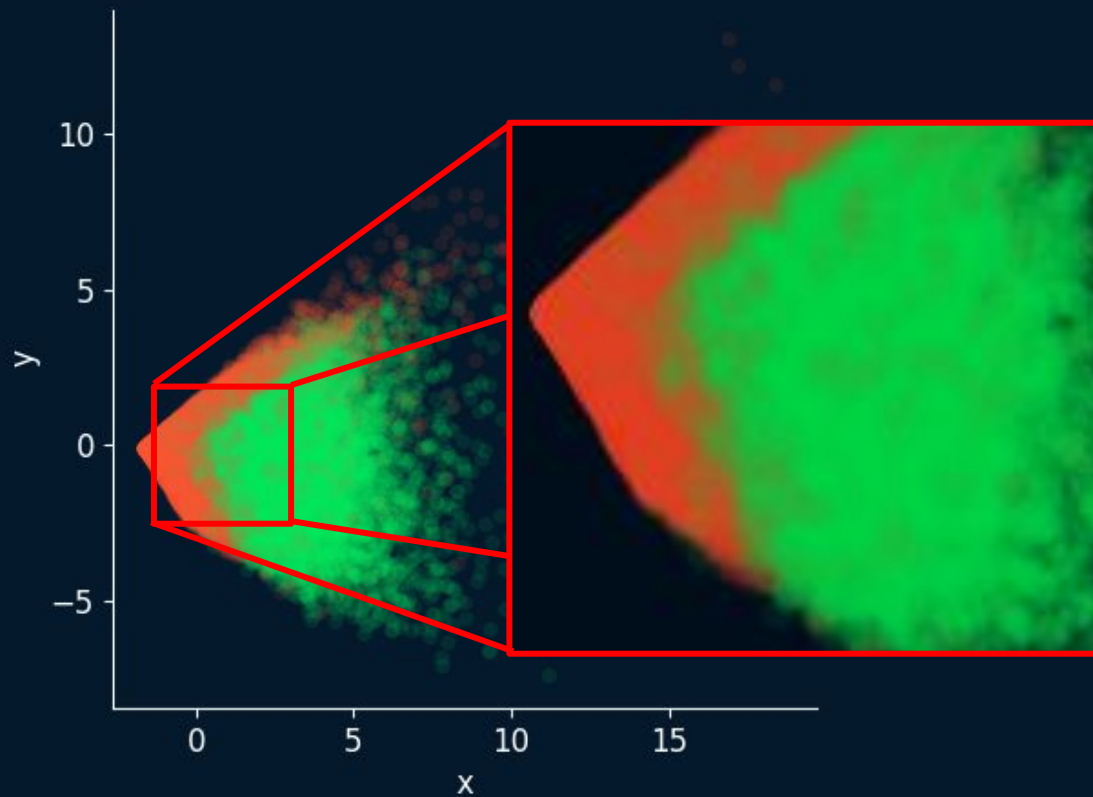
PCA REPRESENTATION



PCA REPRESENTATION



PCA REPRESENTATION



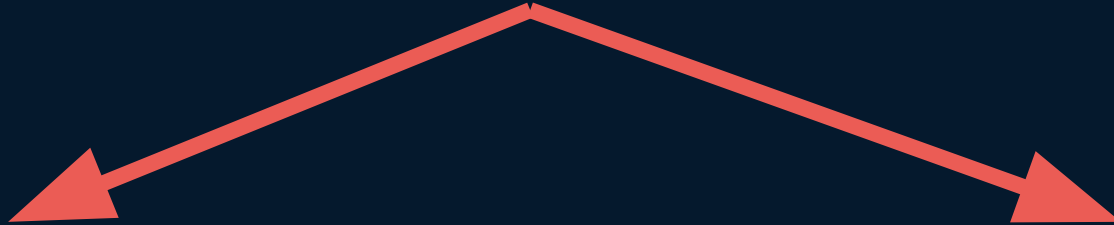
Preparation for ML

4 million events



$n_{jet30} > 1$

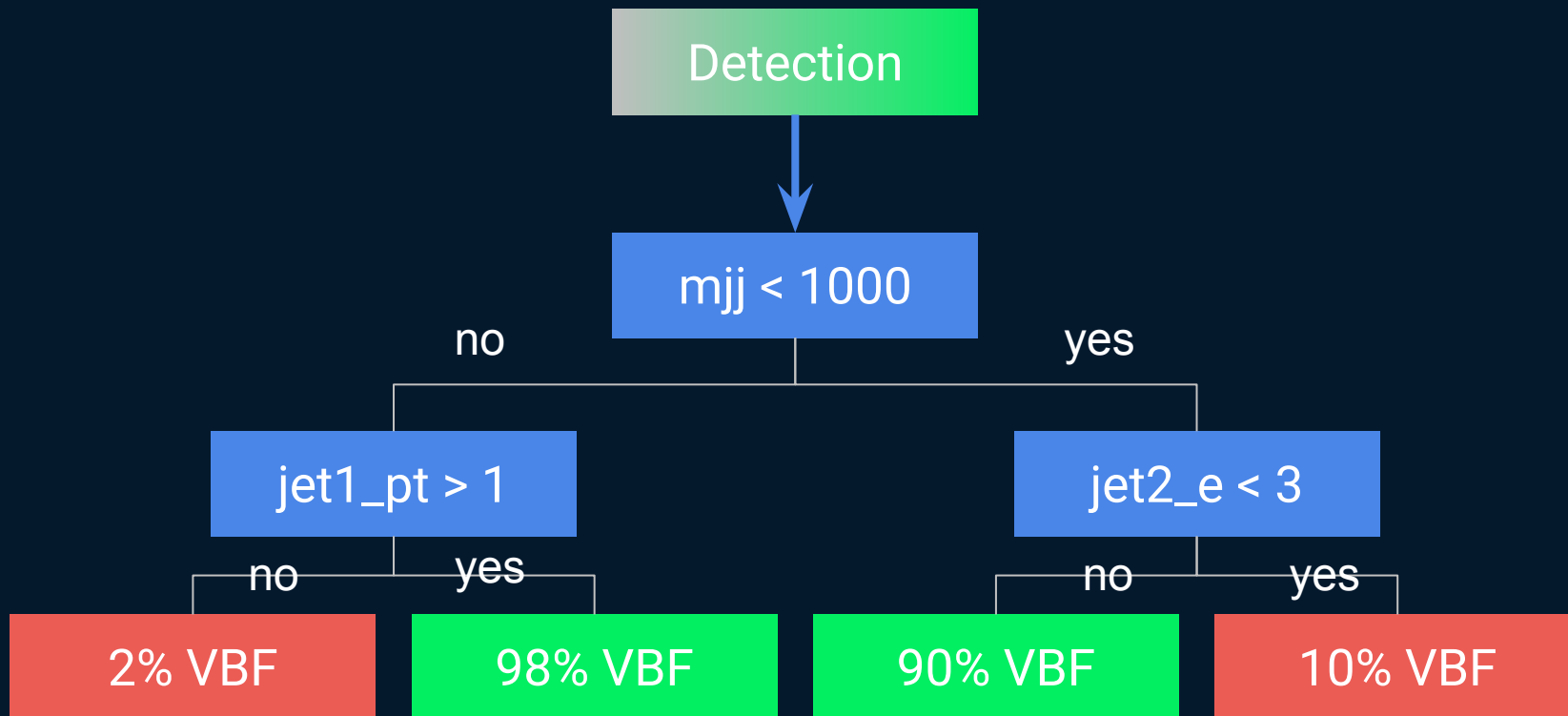
2 million events



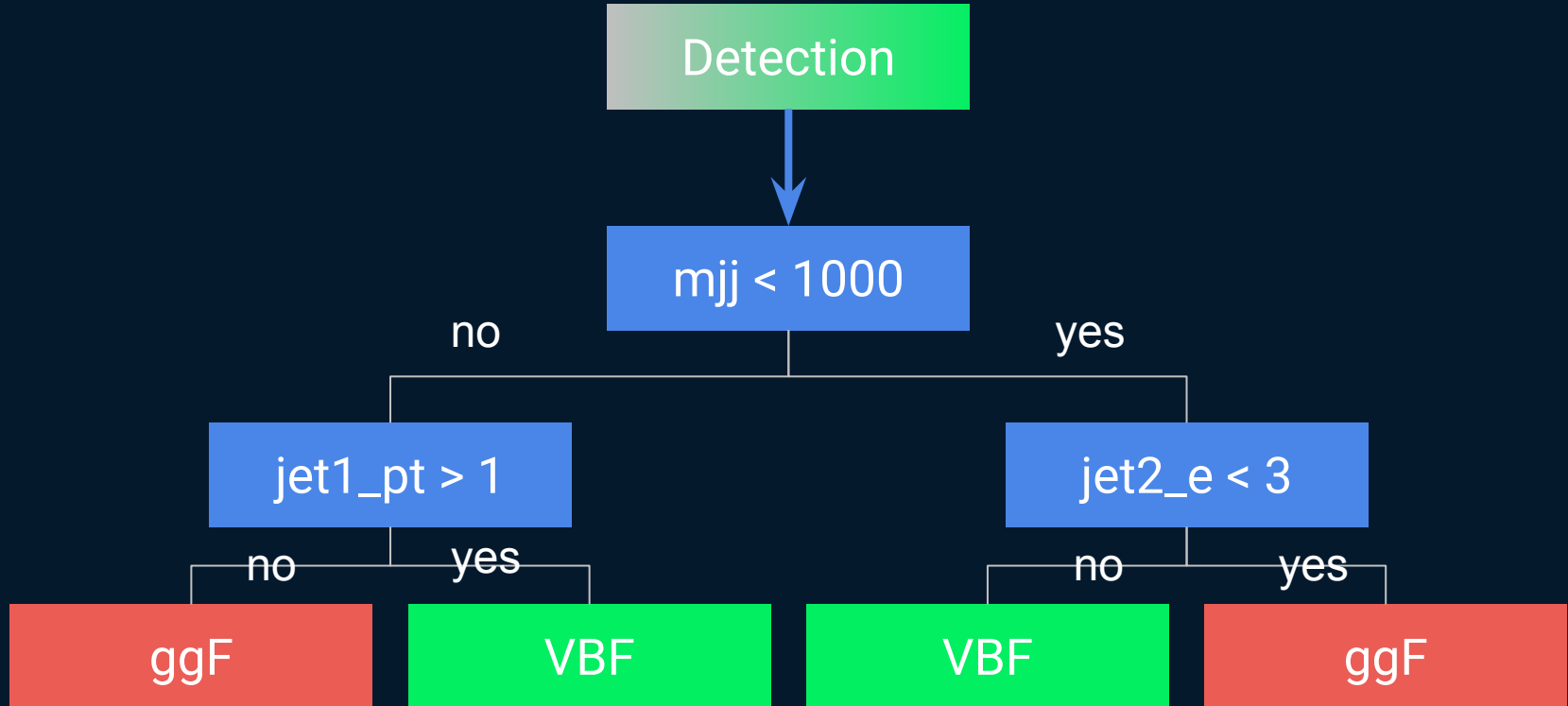
70% training set

30% testing set

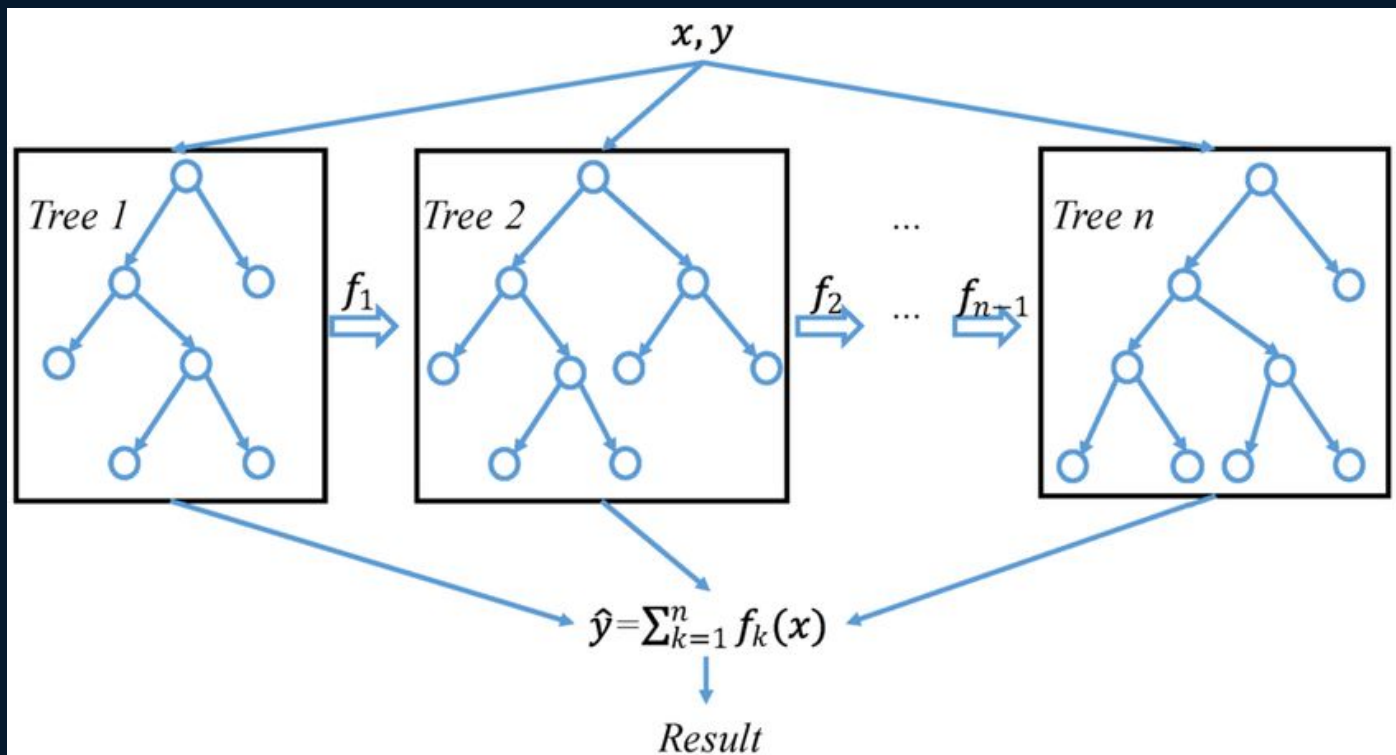
DECISION TREES CLASSIFICATION



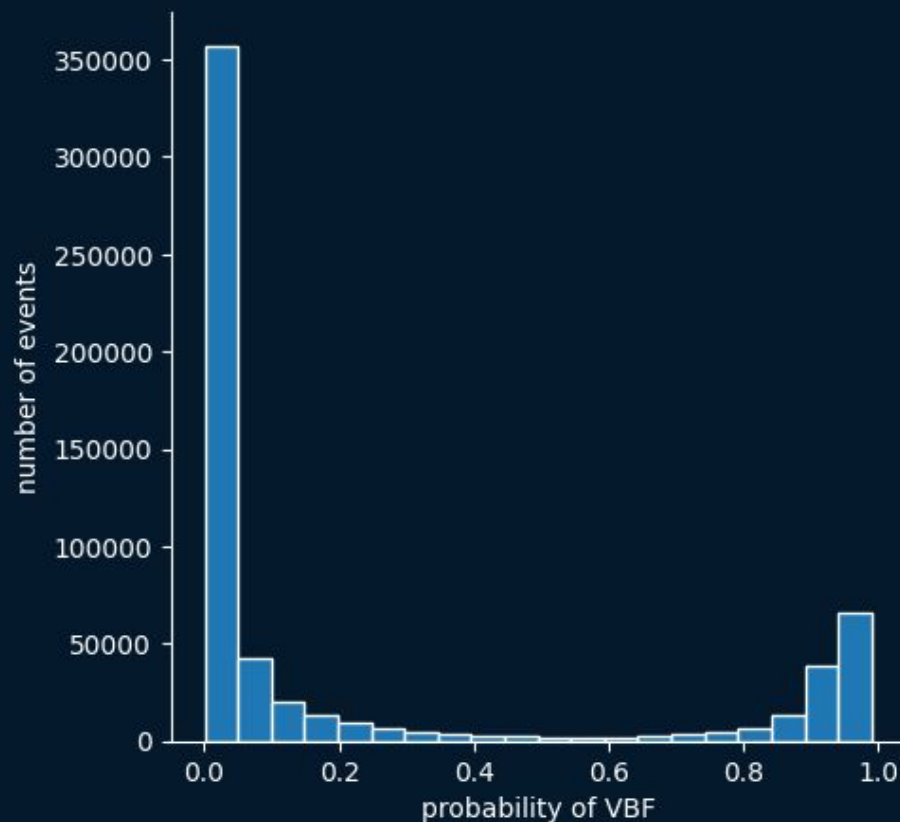
DECISION TREES CLASSIFICATION



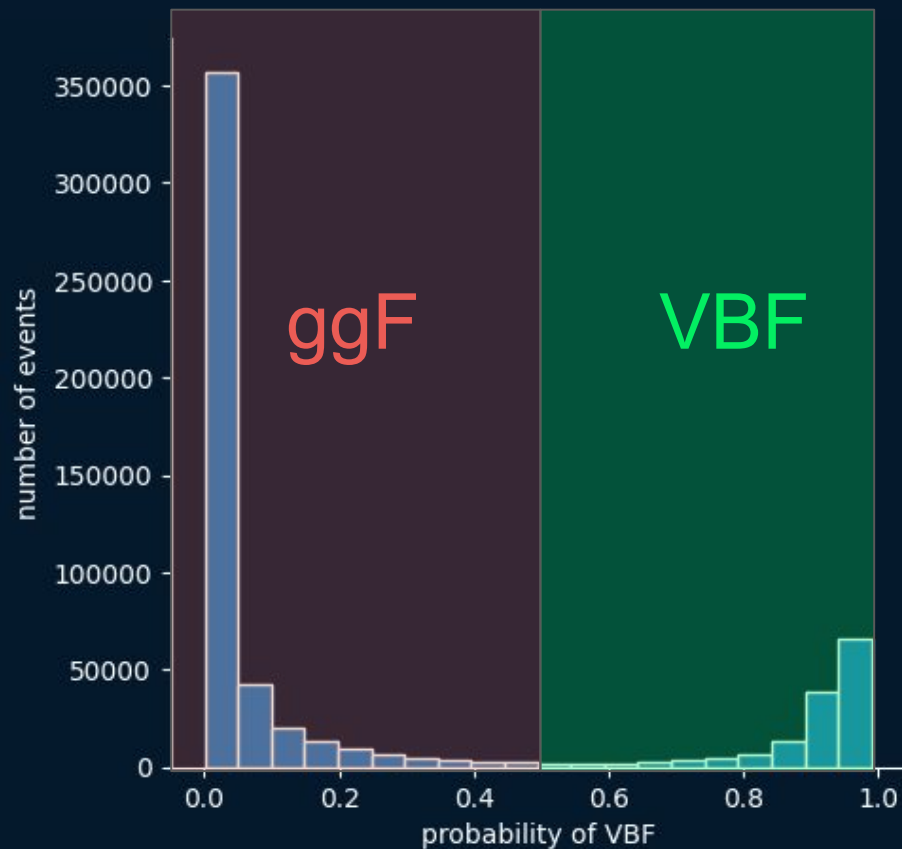
XGBoost



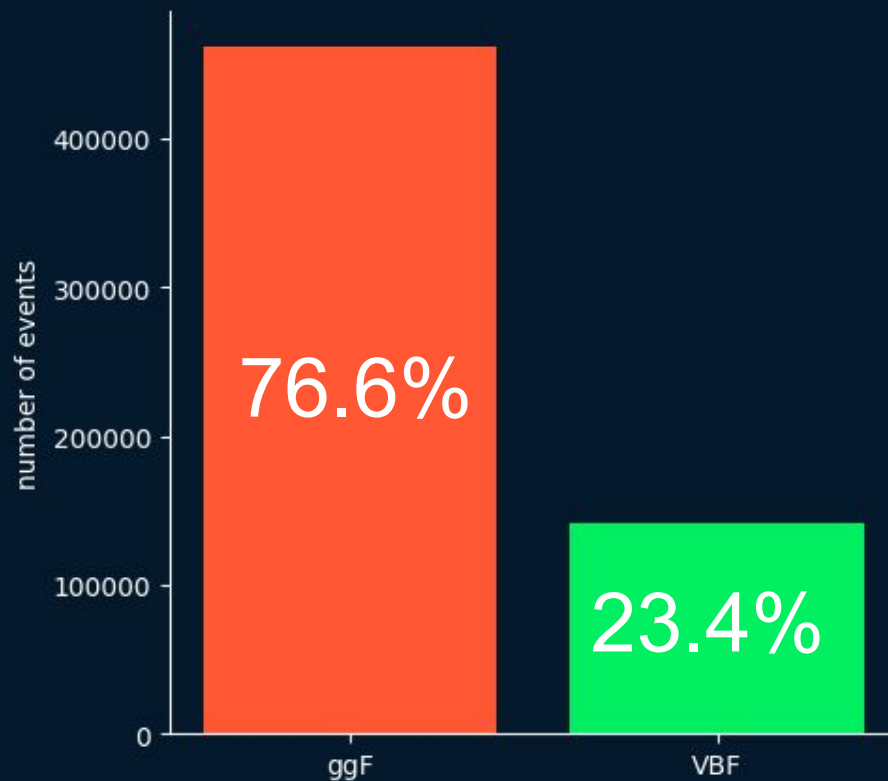
HISTOGRAM OF CLASSIFICATION OUTPUT



HISTOGRAM OF CLASSIFICATION OUTPUT



HISTOGRAM OF CLASSIFICATION OUTPUT



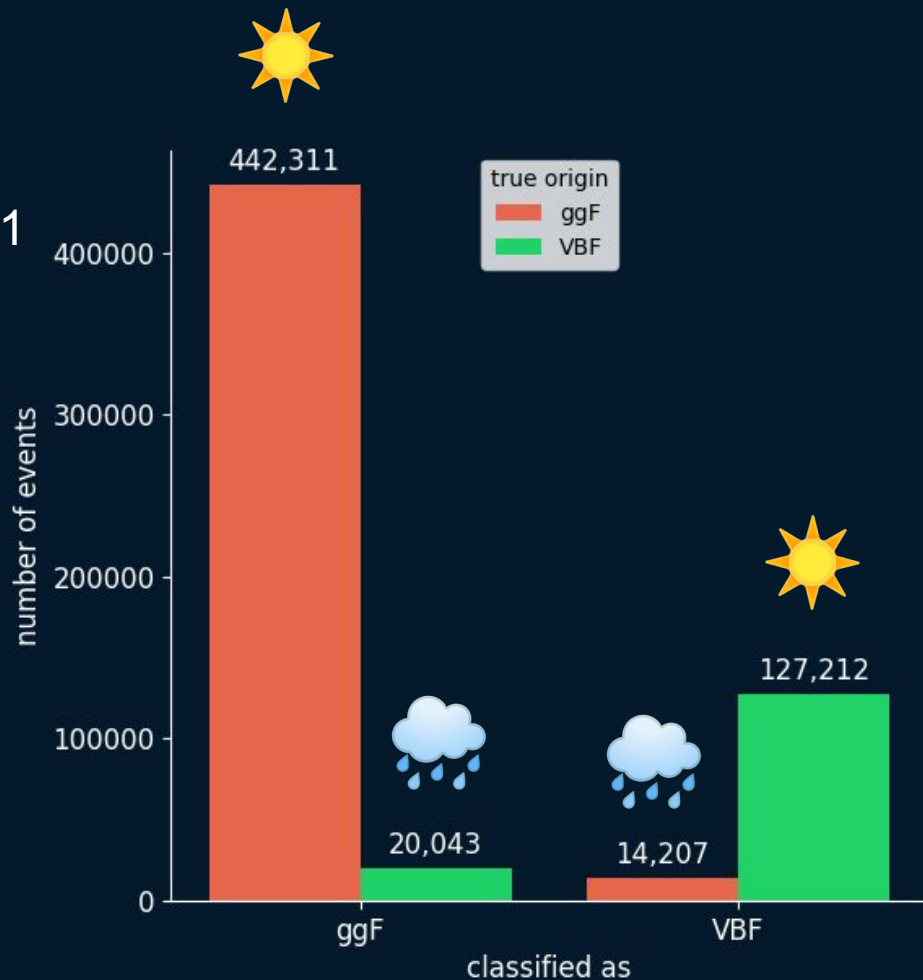
NEW CUTS

Just cut $n_{jet30} > 1$

and ML

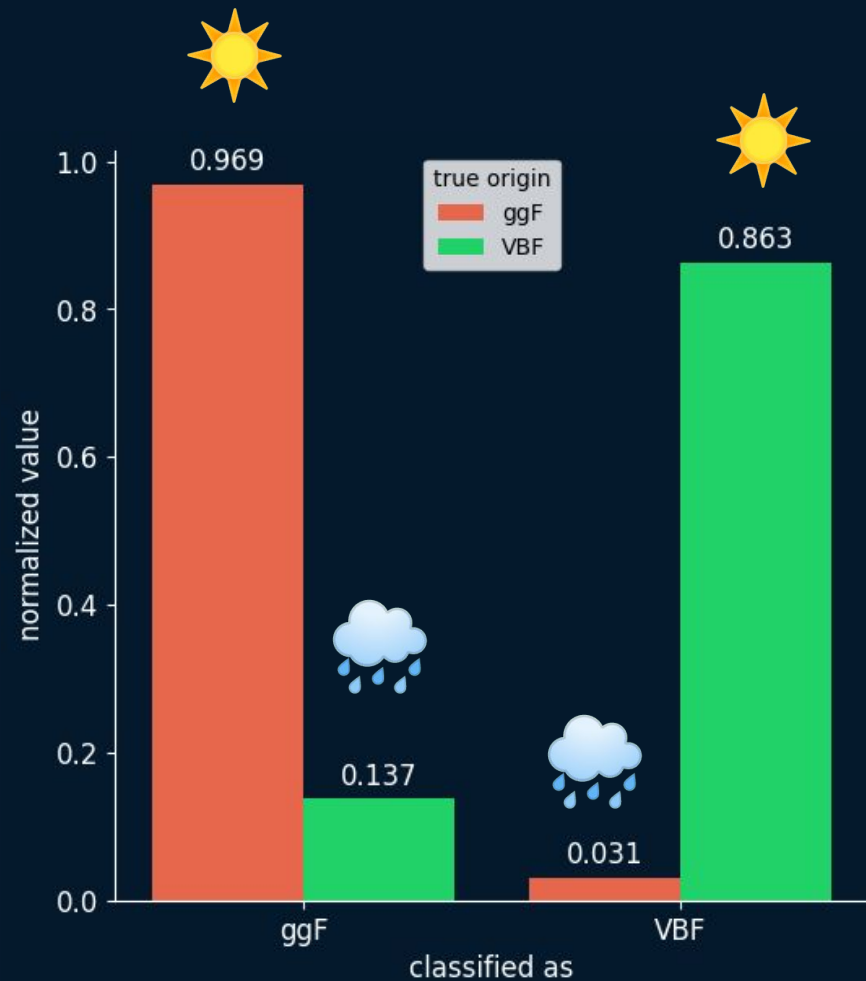
NEW CUTS

Just cut $n_{\text{jet}30} > 1$
and ML

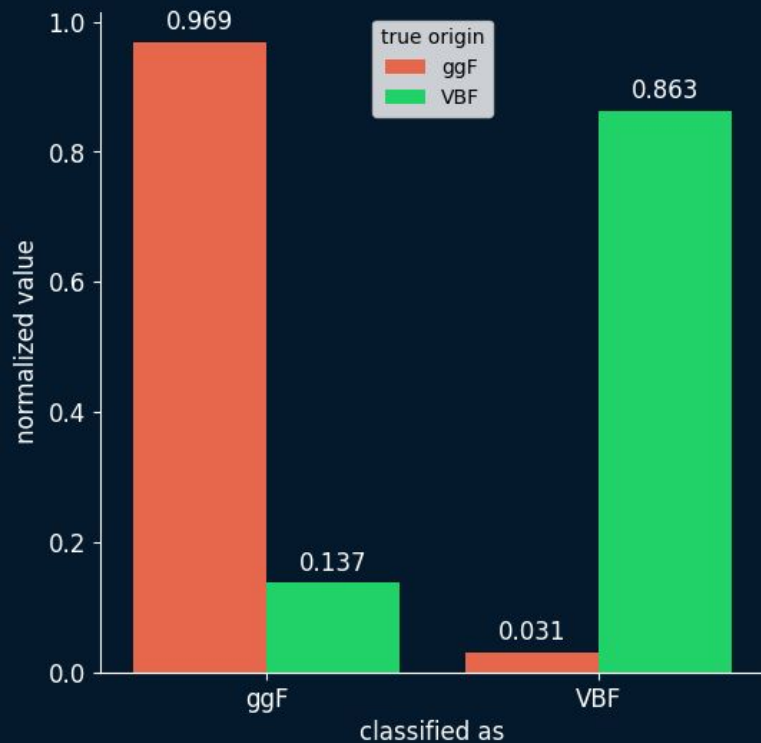
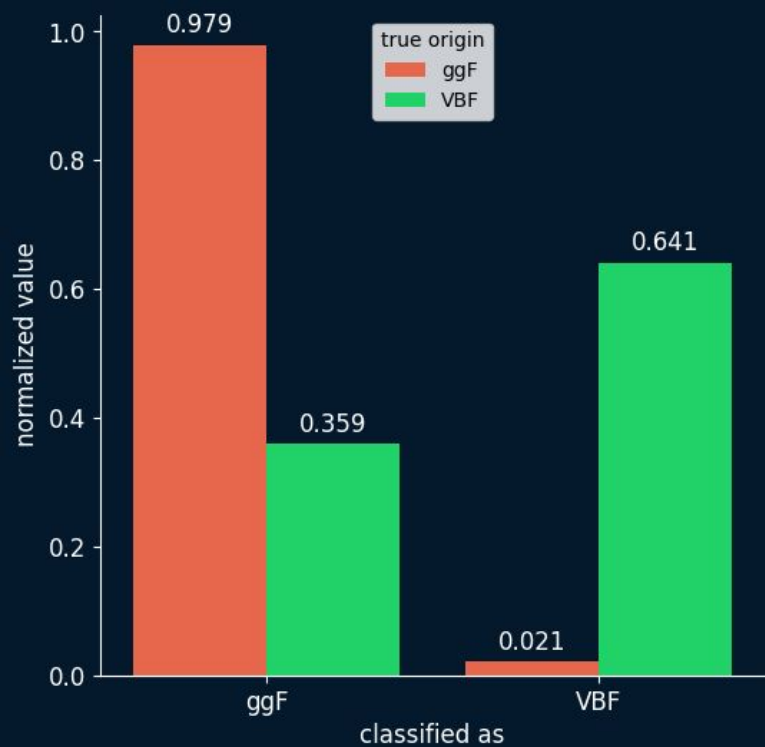


NEW CUTS

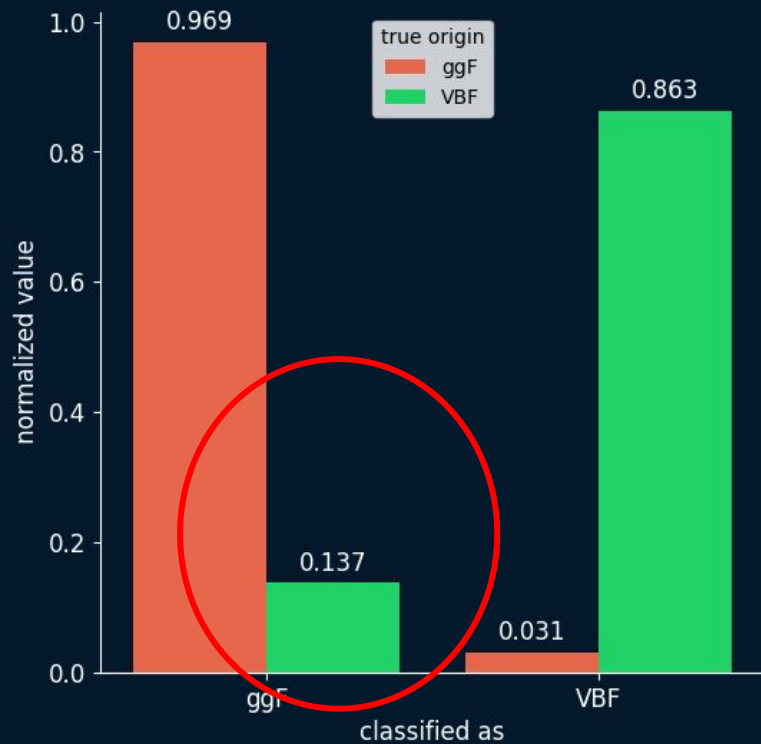
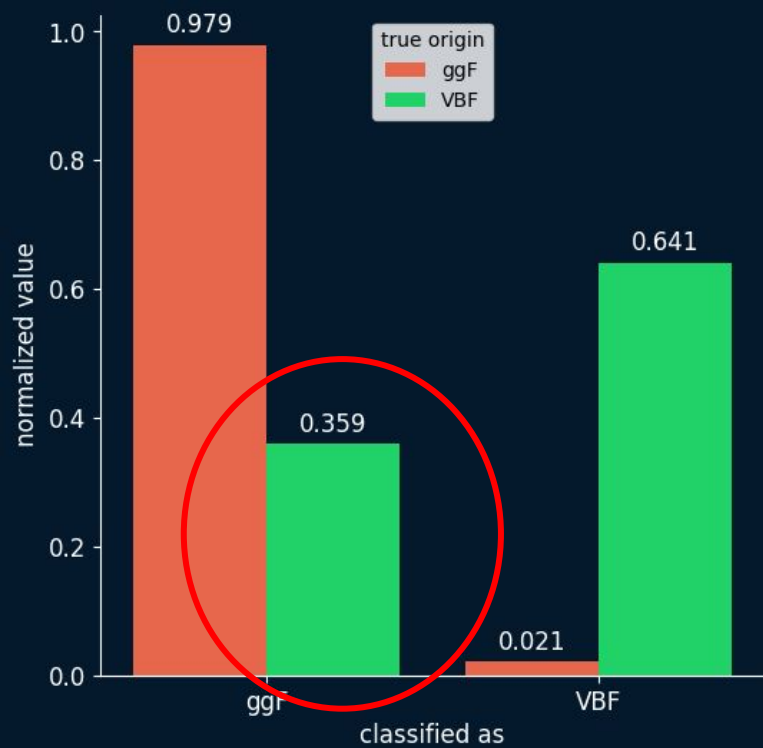
Just cut $n_{\text{jet}30} > 1$
and ML



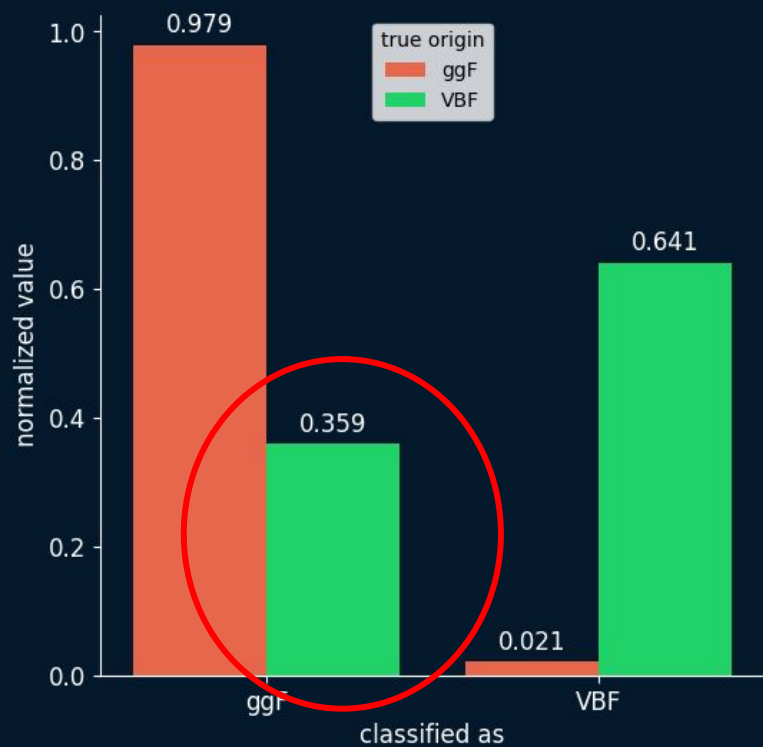
IMPROVEMENT



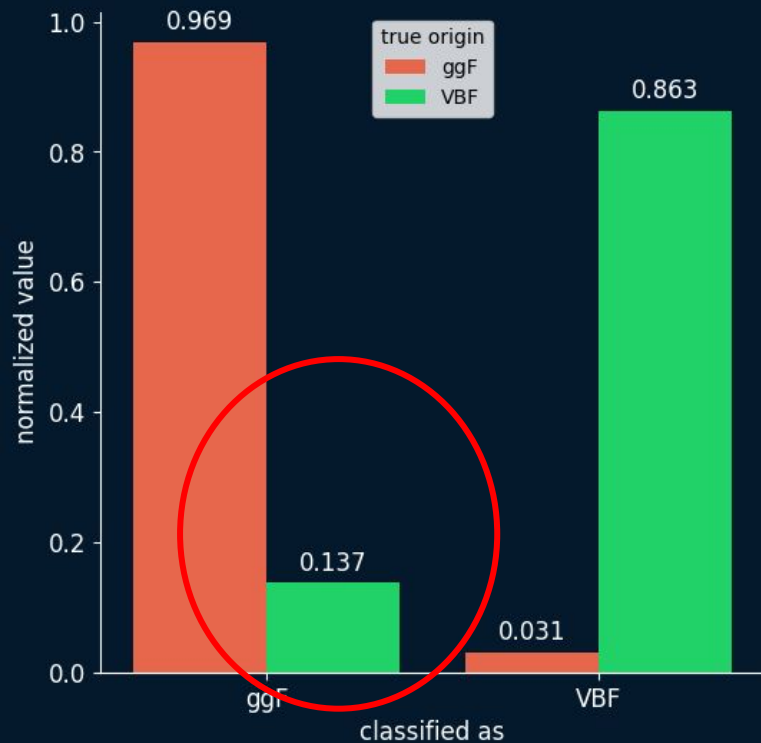
IMPROVEMENT



IMPROVEMENT

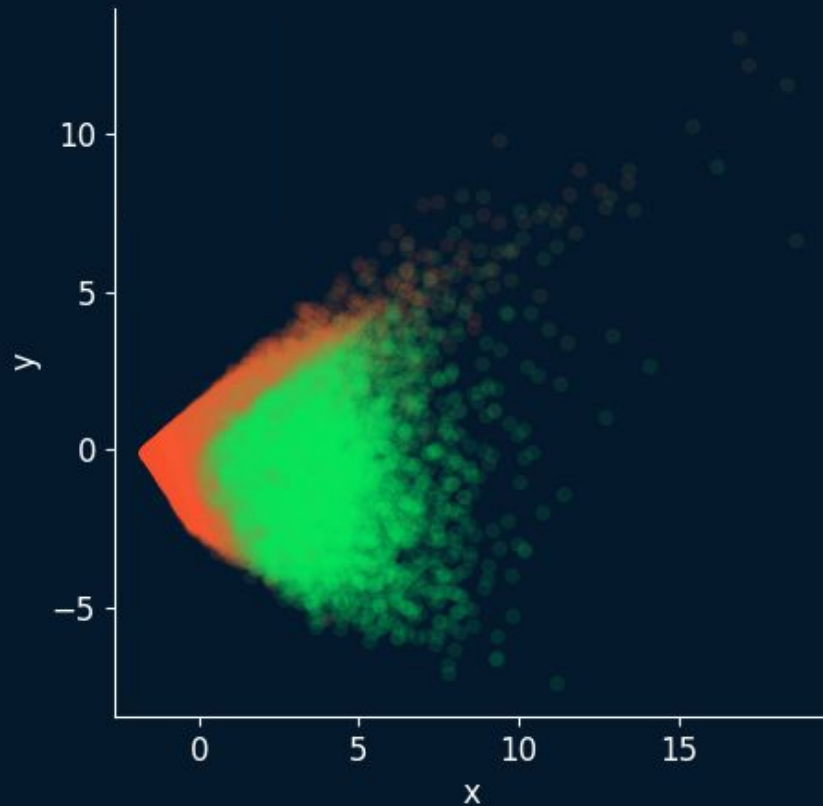


F1 score: 0.75

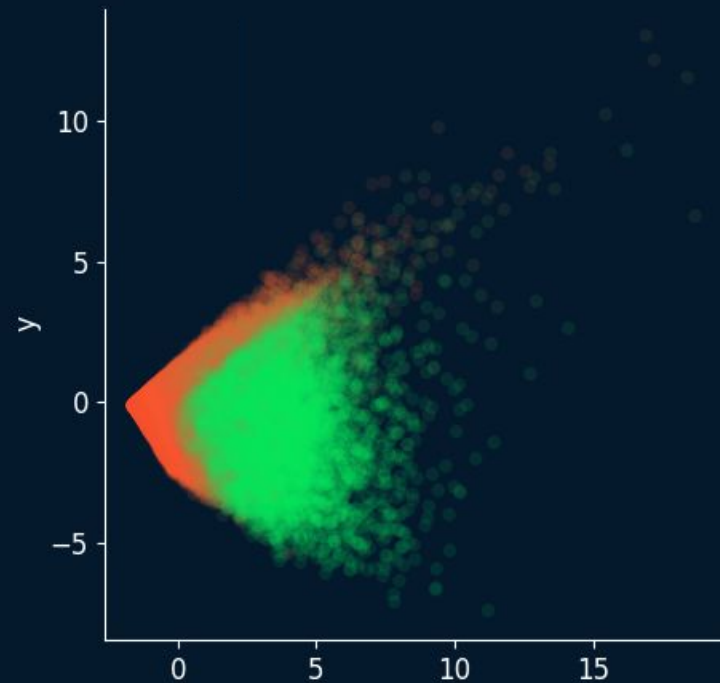
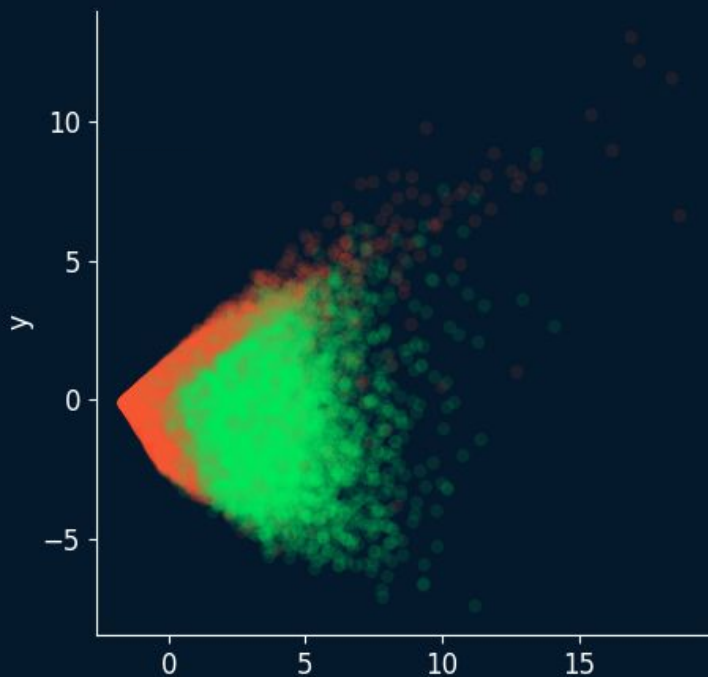


F1 score: 0.88

PREDICTED PCA REPRESENTATION



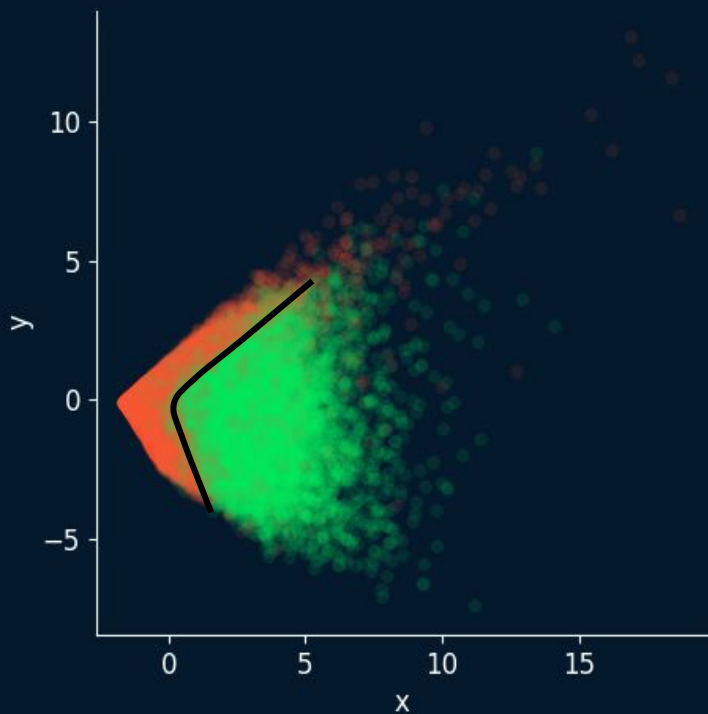
IMPROVEMENT



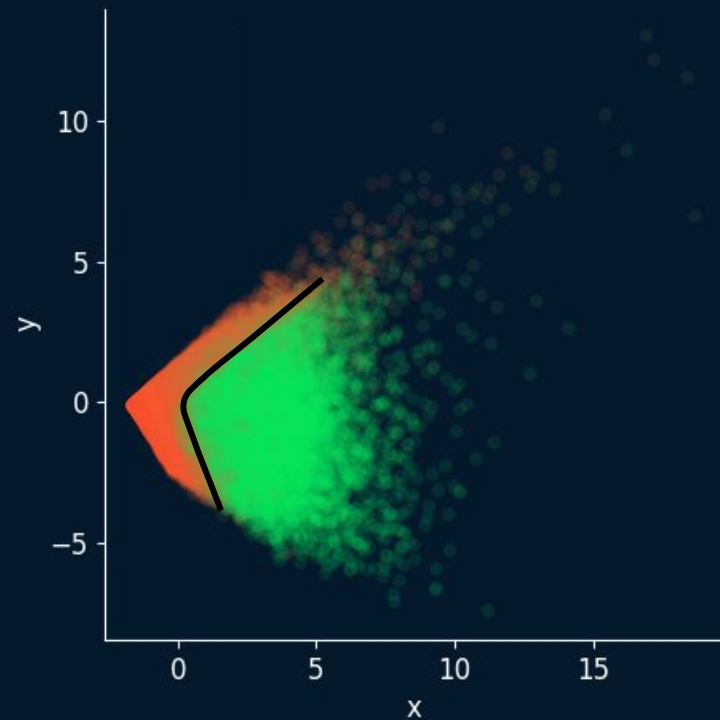
Real

Predicted

IMPROVEMENT

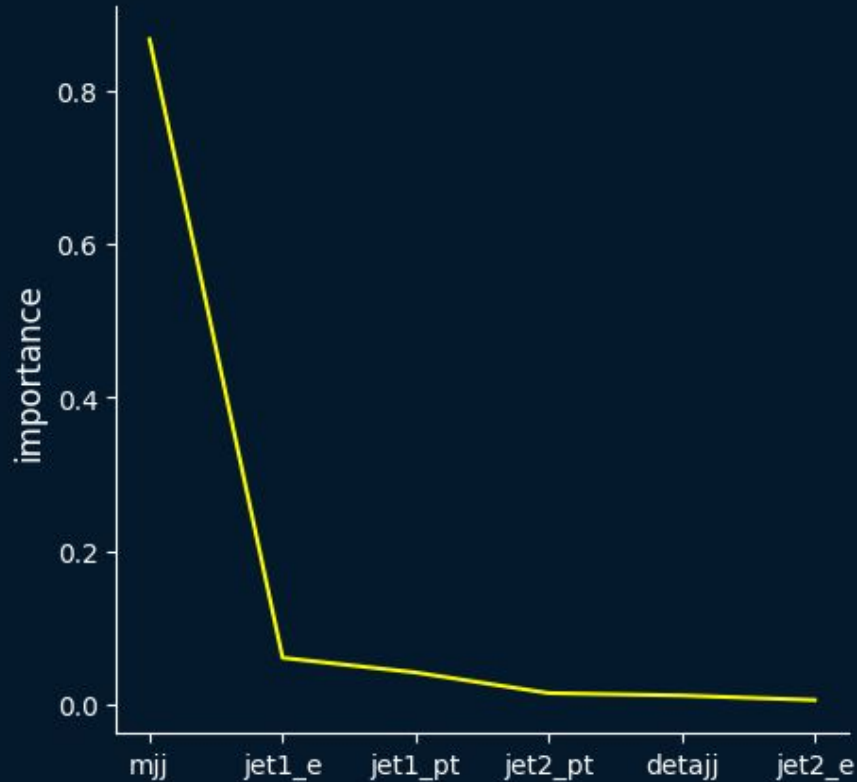


Real

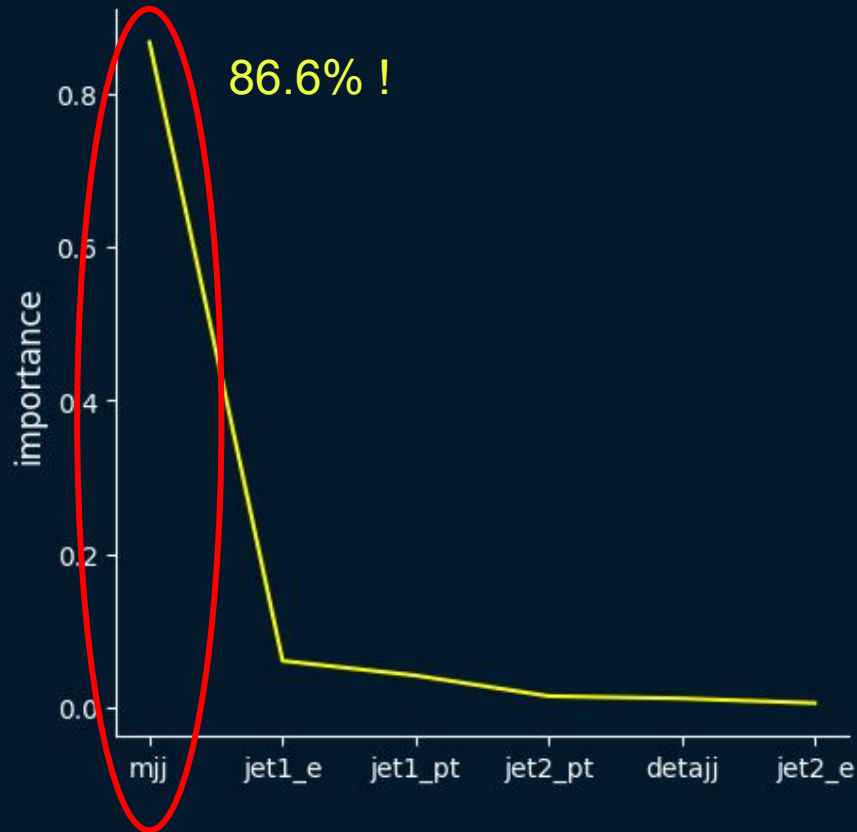


Predicted

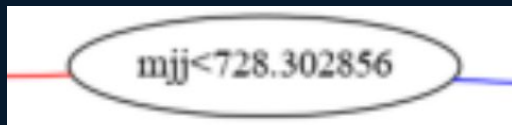
Most important variables in model



Most important variables in model

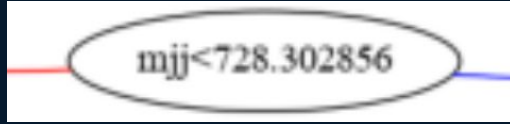


ML CUTS



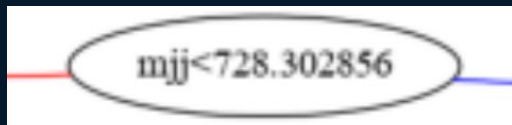
$m_{jj} < 728.302856$

ML CUTS

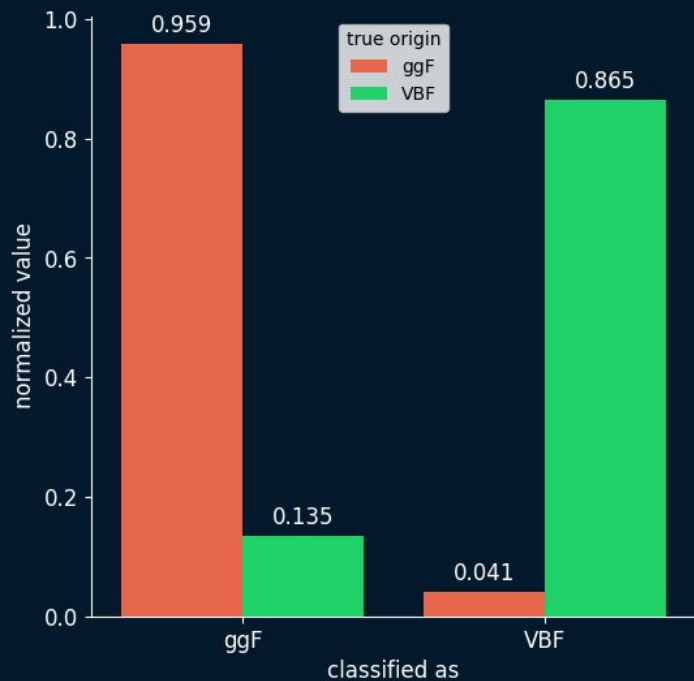


FIRST CUT IN
DECISION
TREE

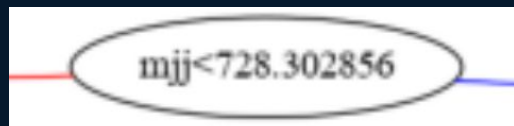
ML CUTS



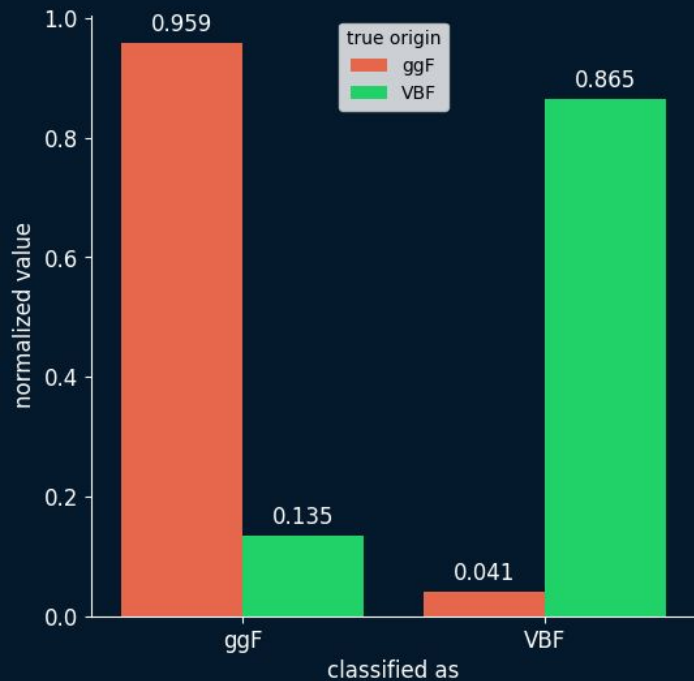
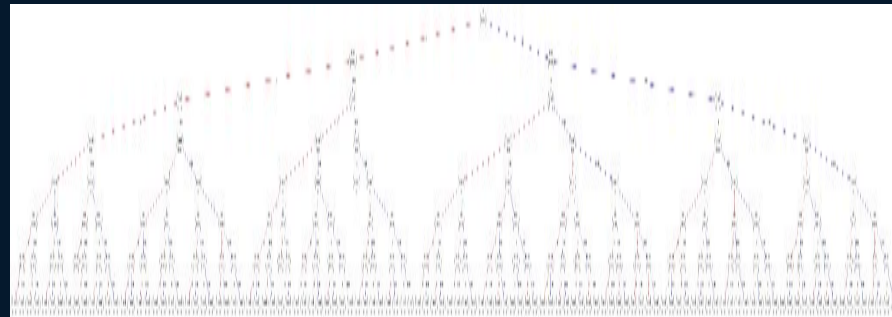
FIRST CUT IN
DECISION
TREE



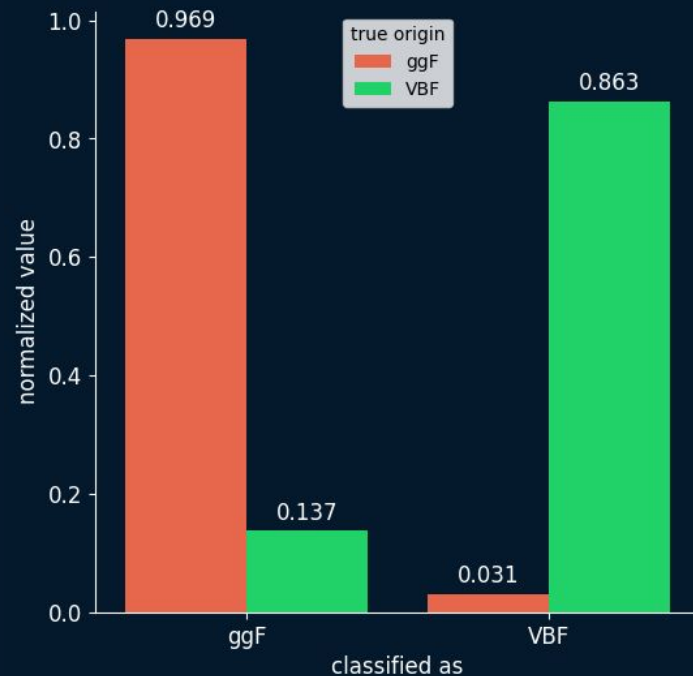
ML CUTS



FIRST CUT IN
DECISION
TREE



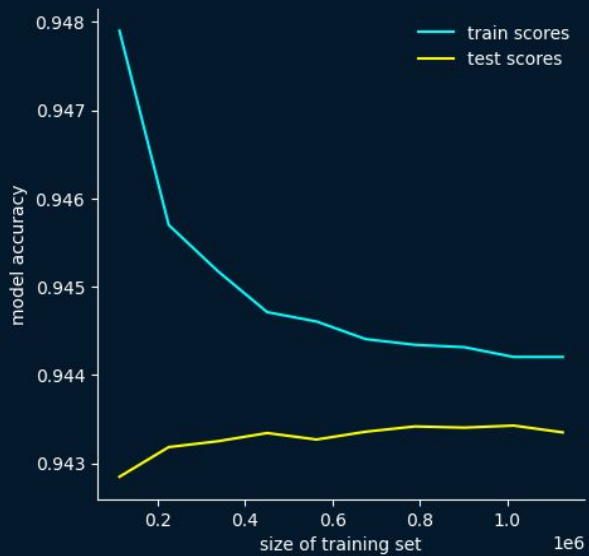
\approx



OVERFITTING?

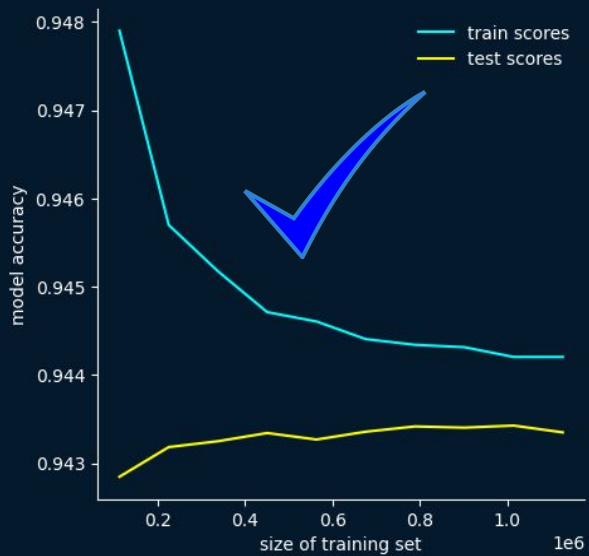
OVERFITTING?

Learning Curve



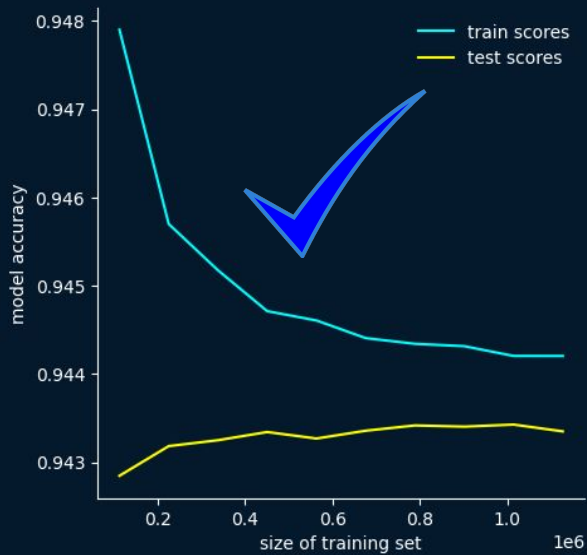
OVERFITTING?

Learning Curve

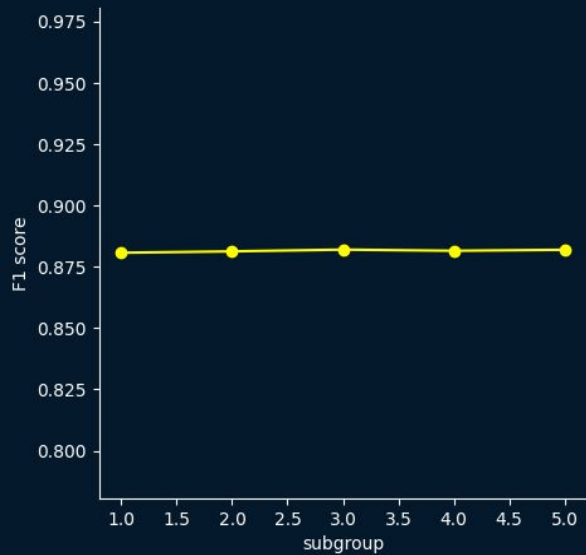


OVERFITTING?

Learning Curve

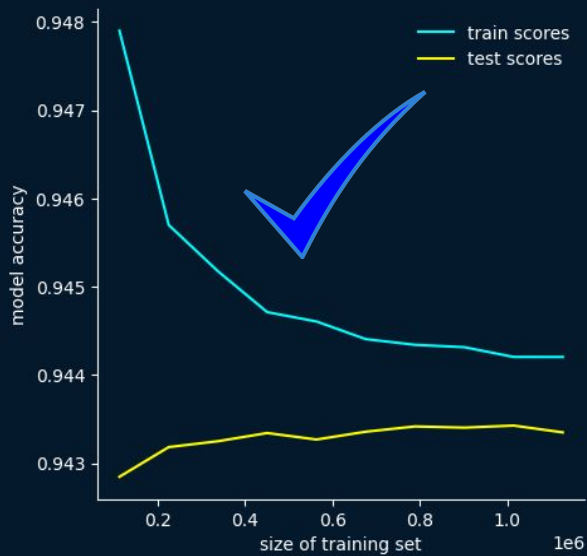


Cross Validation

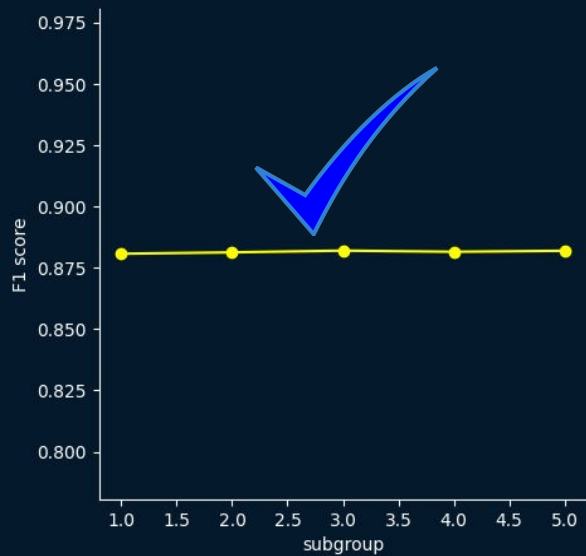


OVERFITTING?

Learning Curve

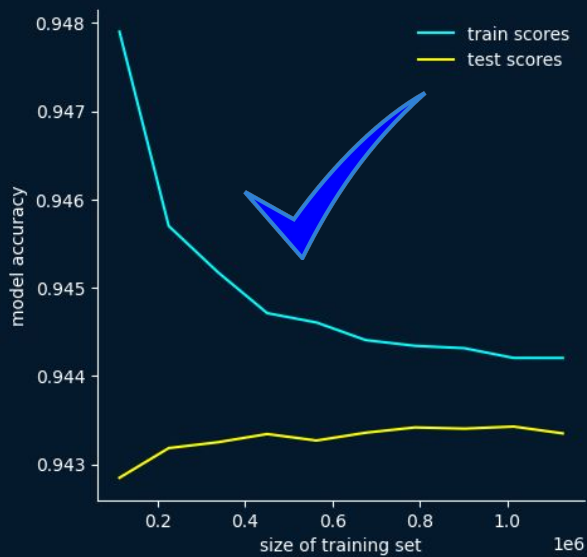


Cross Validation

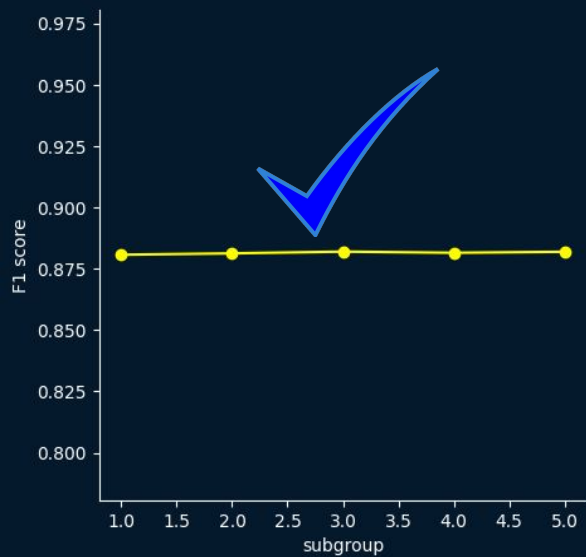


OVERFITTING?

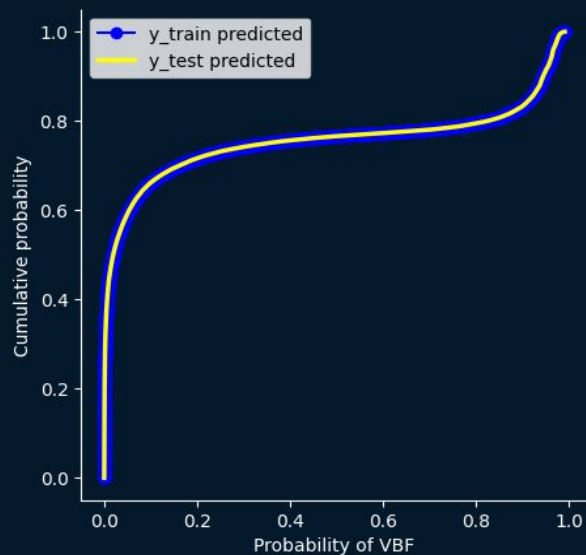
Learning Curve



Cross Validation

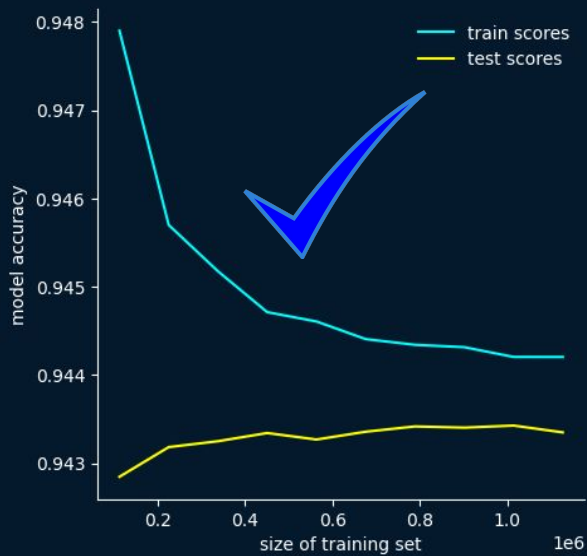


KS test

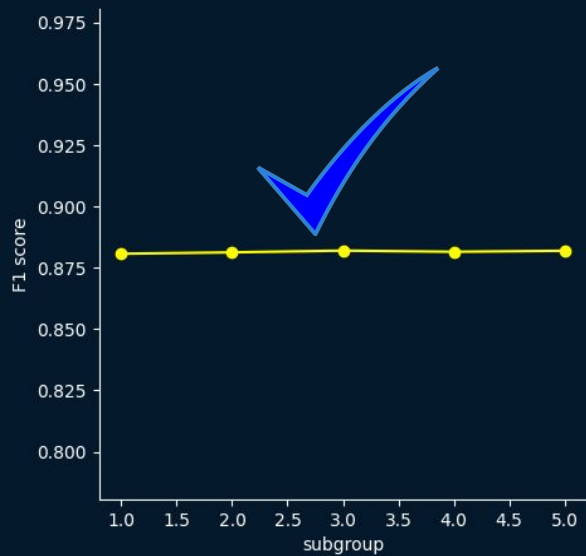


OVERFITTING?

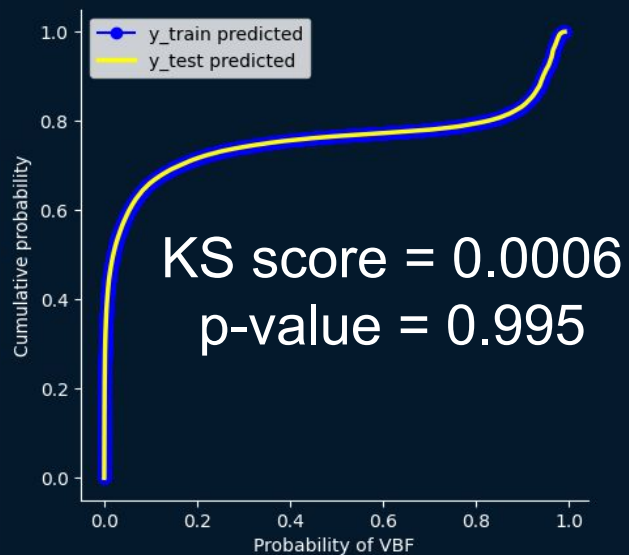
Learning Curve



Cross Validation

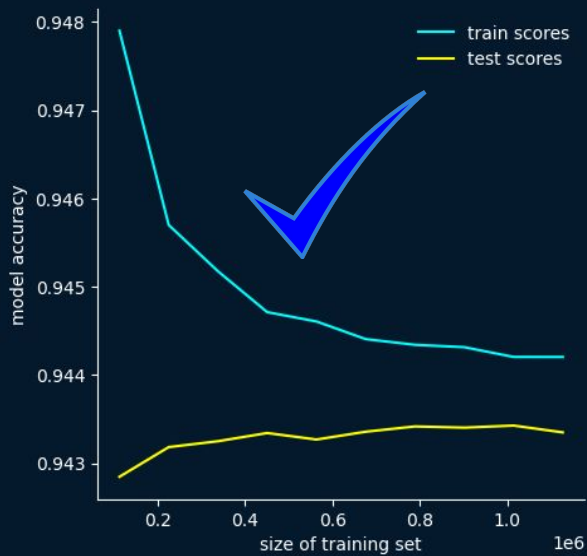


KS test

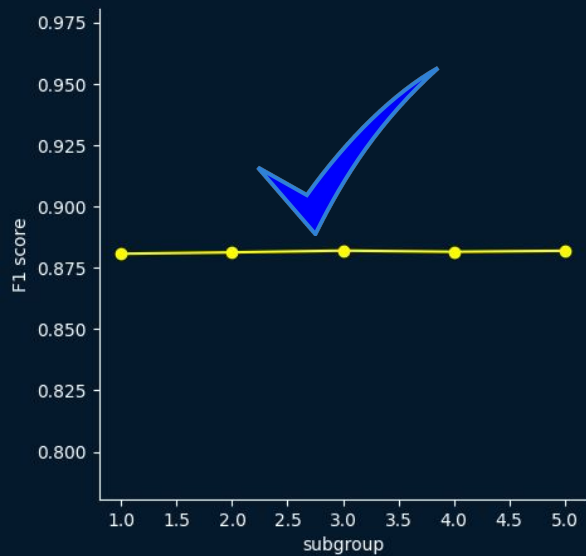


OVERFITTING?

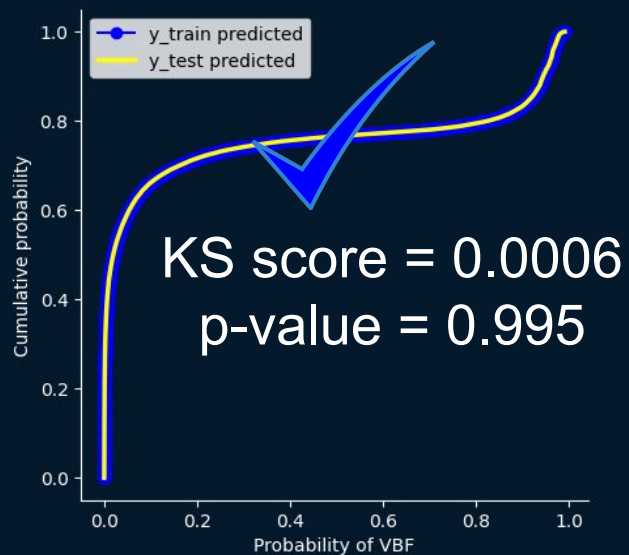
Learning Curve



Cross Validation



KS test



NEXT STEPS



NEXT STEPS



THANKS FOR YOUR
ATTENTION

THANKS FOR YOUR ATTENTION



https://github.com/tomilee09/tesis_pregrado