Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines

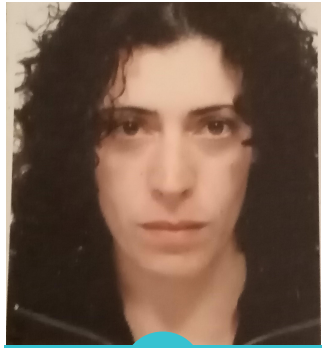
Big Data Analytics
A.A. 2020/21

MaLuCS
Luca Corbucci
Cinzia Lestini
Marco Giuseppe Marino
Simone Rossi

MaLuCS TEAM

LUCA CORBUCCI

Master Degree in Computer Science

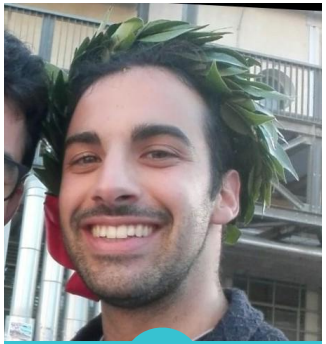


CINZIA LESTINI

Master Degree Data Science and Business Informatics

MARCO GIUSEPPE MARINO

Master Degree in Computer Science



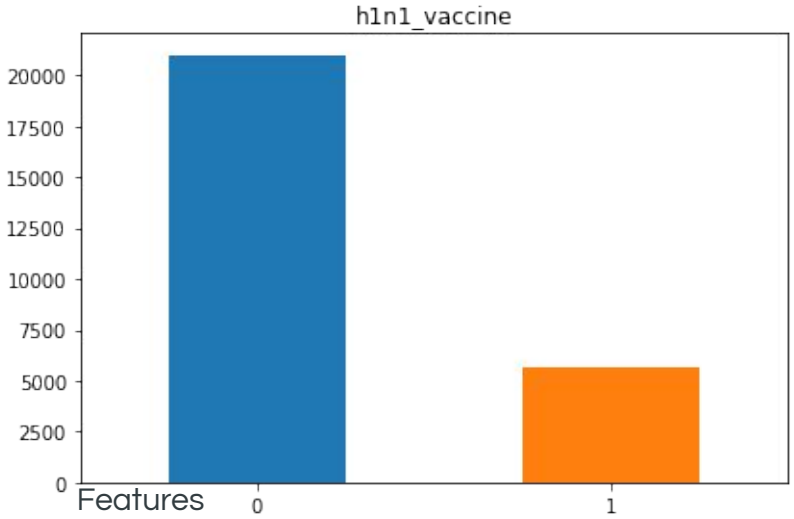
SIMONE ROSSI

Master Degree in Computer Science

INTRODUCTION

BALANCING

H1N1 VACCINE



26707

Rows in training dataset

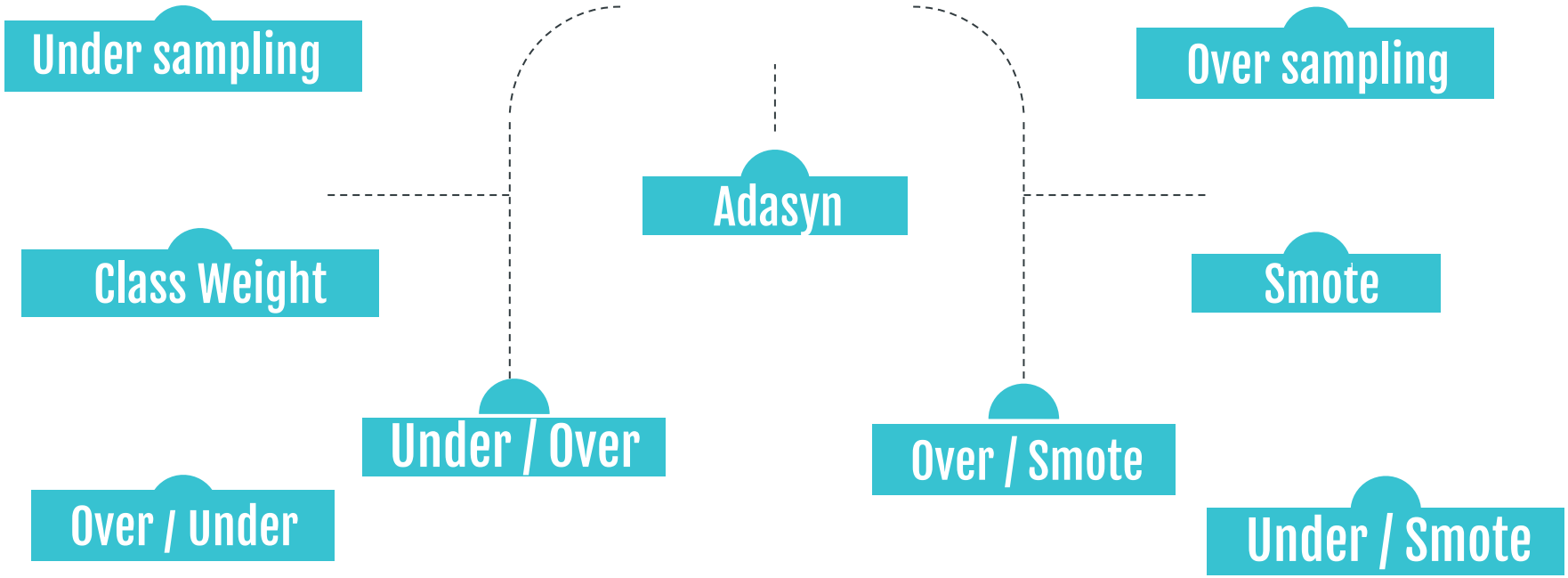
79% Class 0

21% Class 1

Consequences:

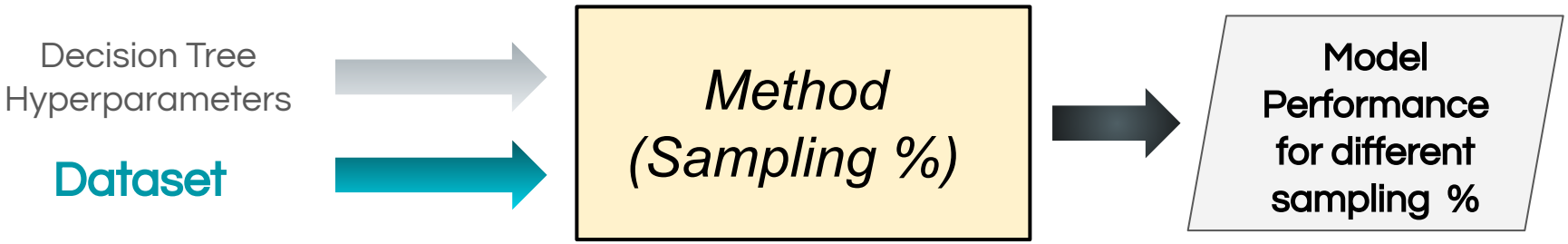
Lower performance class 1

BALANCING METHODOLOGIES USED



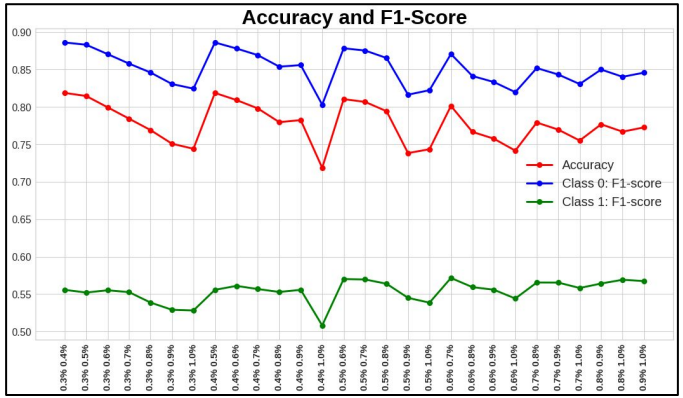
New Functions for Best sampling Percentages

ONE FOR EACH METHOD

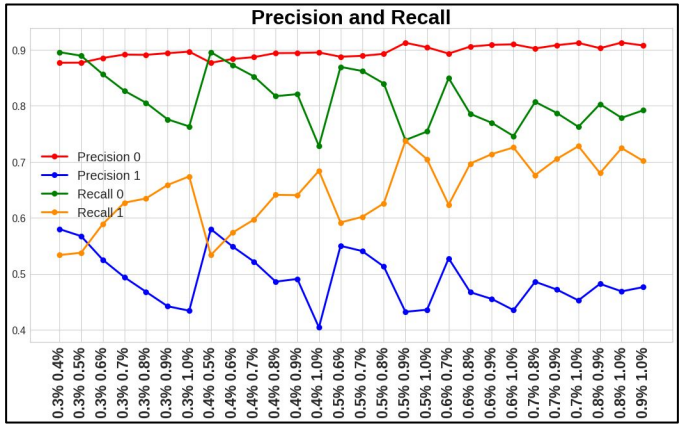


To compare the performance of the methods we used the **Decision Tree Classifier** as a model

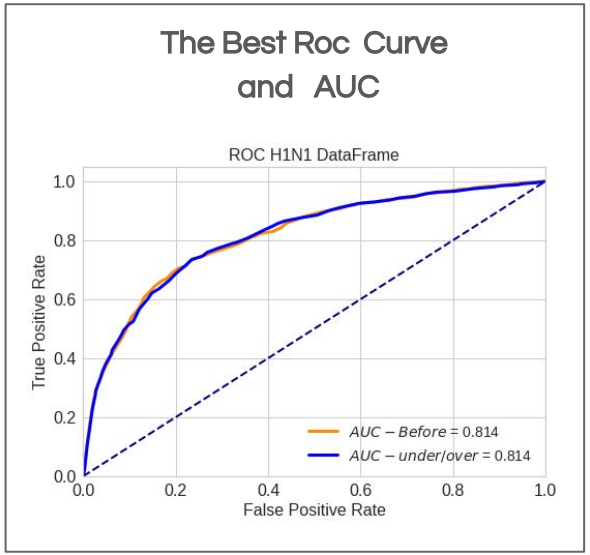
An example: Under/Over Sampling %



Accuracy and F1-Score



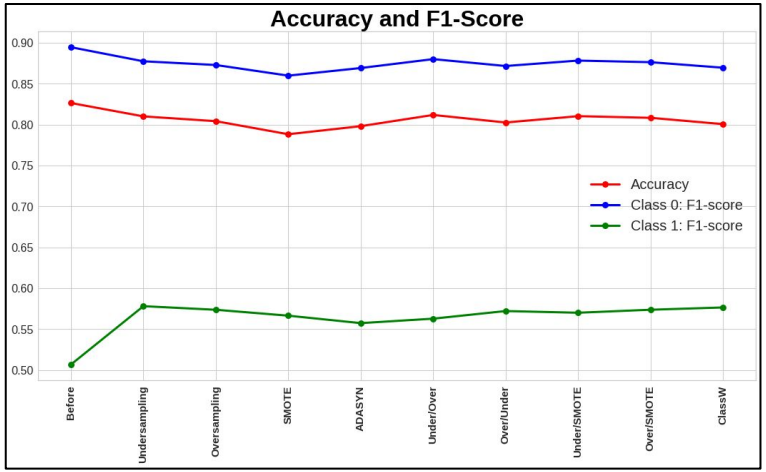
Precision and Recall



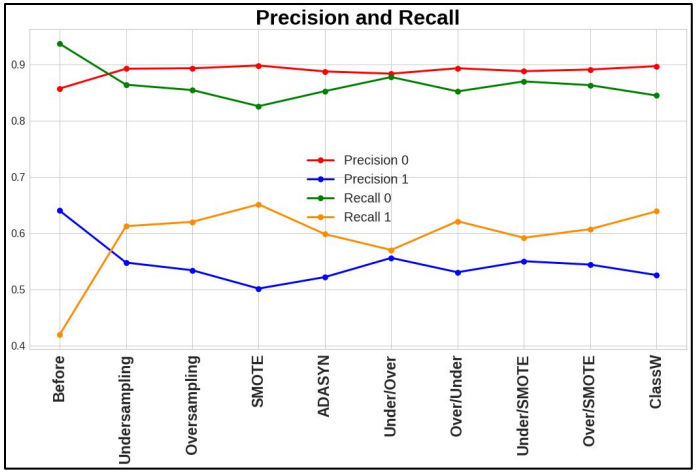
To compare the performance of the methods we used the **Decision Tree Classifier** as a model

Compared Methods

Accuracy and F1-score



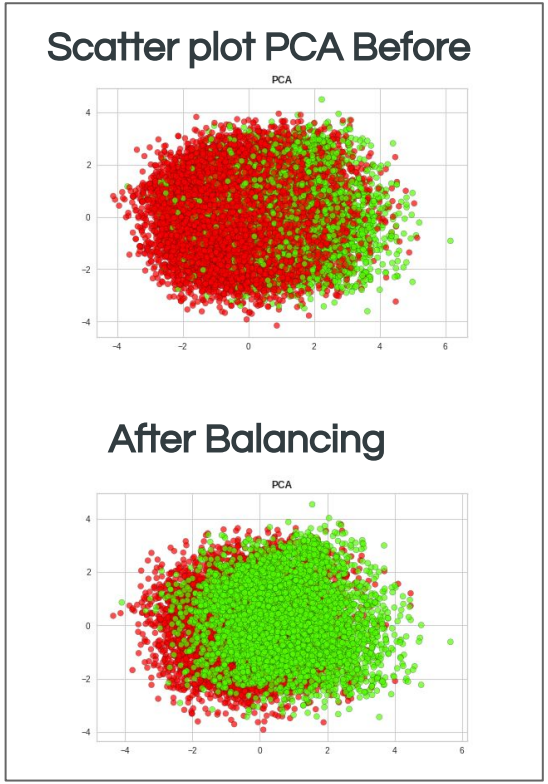
Precision and Recall



Best Method: Under Sampling 0,6

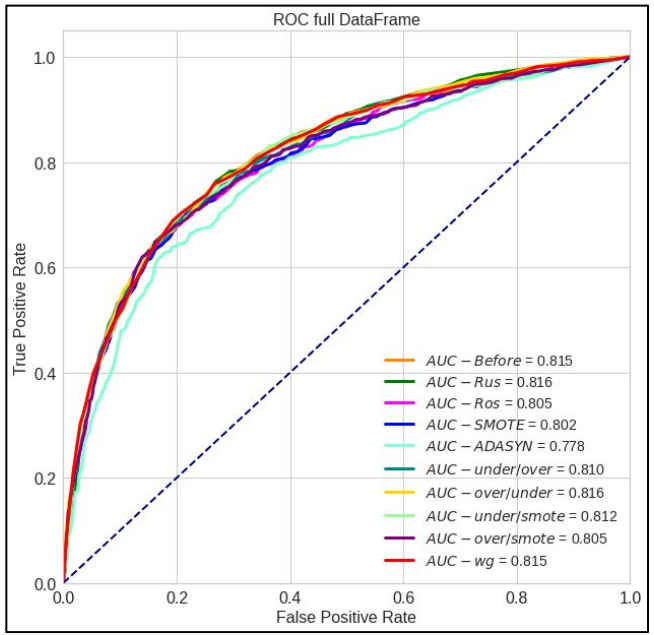
Accuracy	0,81
Auc	0.817
Best F1score class1	0,58

(Before was 0,51)



Compared Methods

Curva Roc and AUC



What's the Best Methods?

What's the Best Metrics?

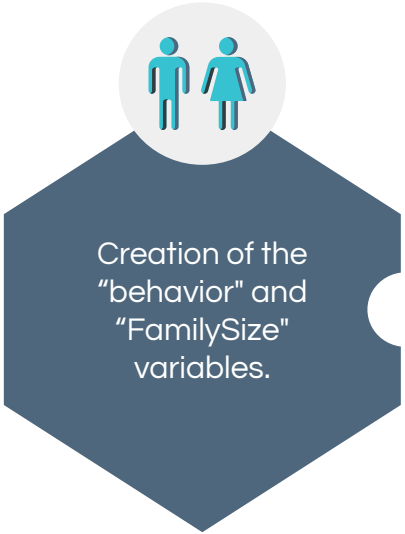
A Trade-off:

- Good Accuracy, and Auc
- The Best F1-score Class 1

DATA TRANSFORMATION ASSESSMENT

DATA TRANSFORMATION ASSESSMENT

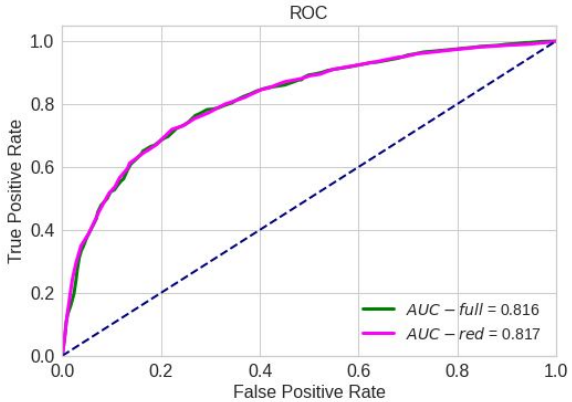
SPLITTING

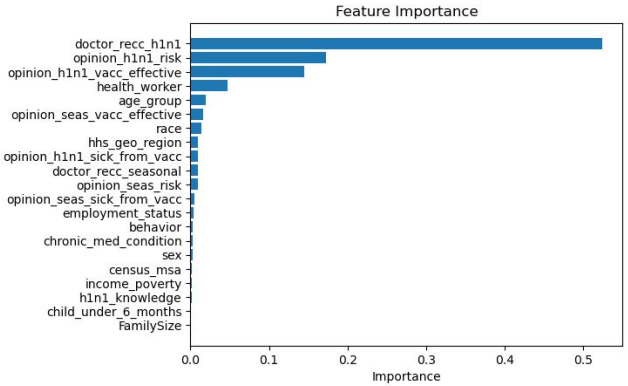


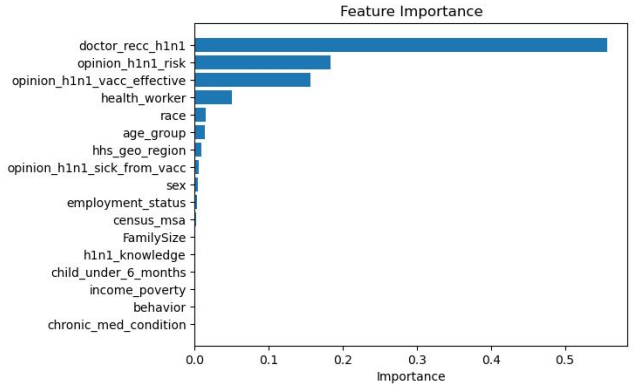
NEW FEATURES

SPLITTING: H1N1

Difference	Accuracy	F1:0	F1:1
Training	0.004	0.002	0.01
Test	0.007	0.002	-0.004

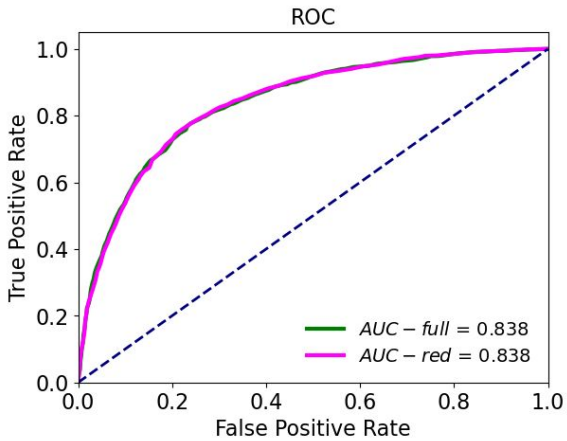


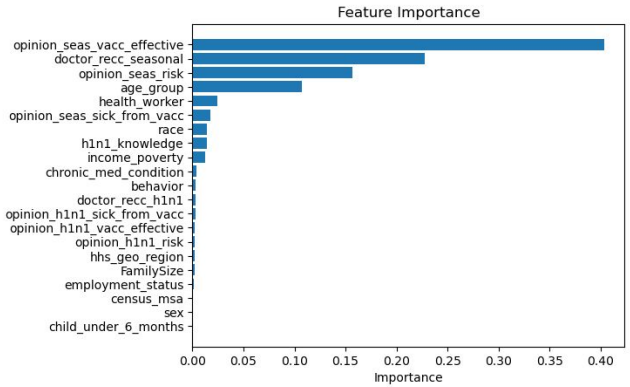


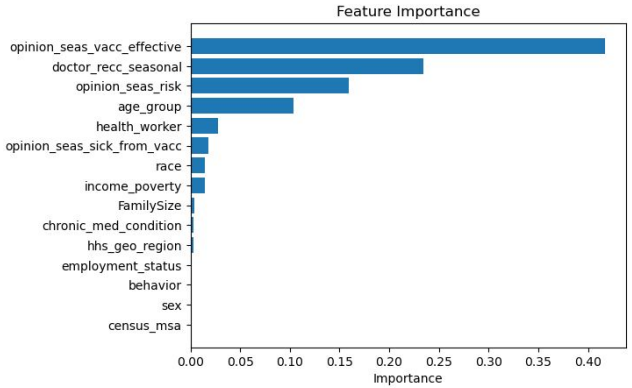


SPLITTING: Seasonal flu

Difference	Accuracy	F1:0	F1:1
Training	0.003	0.005	0.0008
Test	-0.003	-0.001	-0.0068

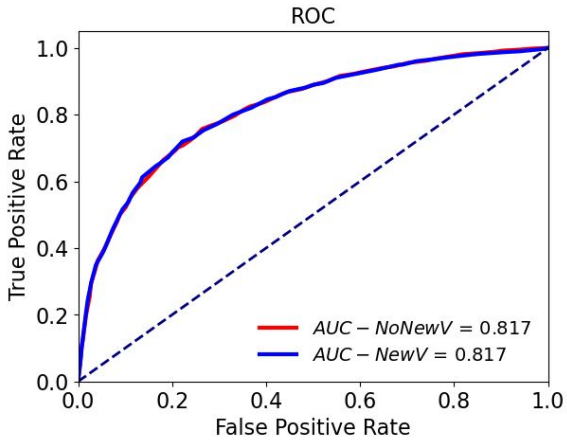


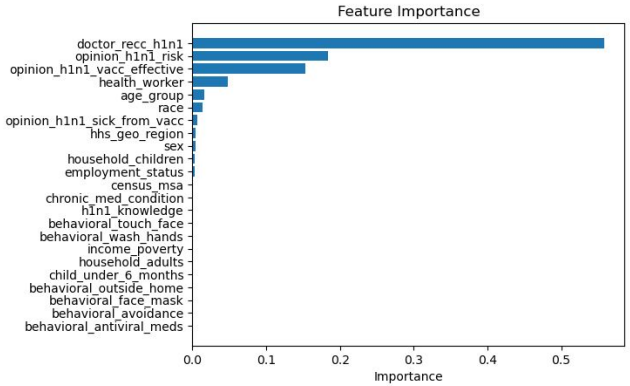


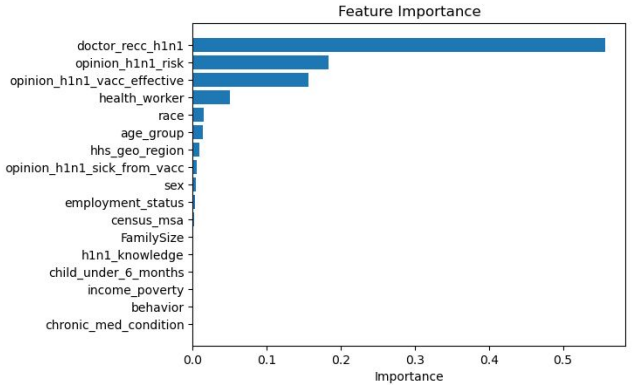


NEW FEATURES: H1N1

Difference	Accuracy	F1:0	F1:1
Training	-0.001	0.001	-0.01
Test	0.0008	0.001	-0.01

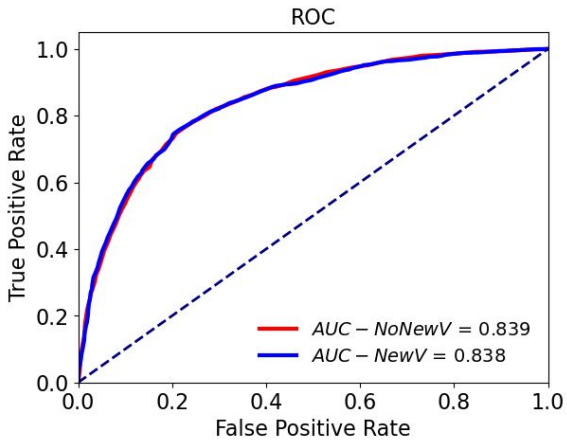


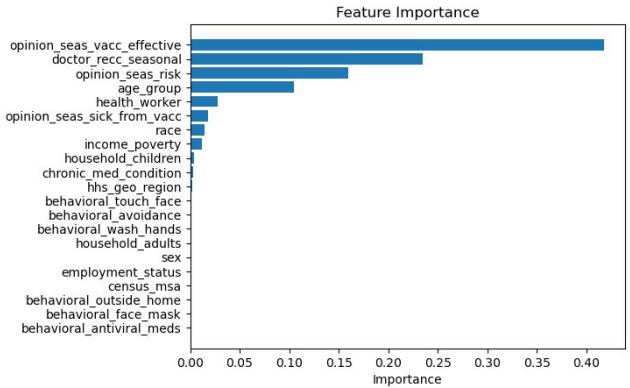


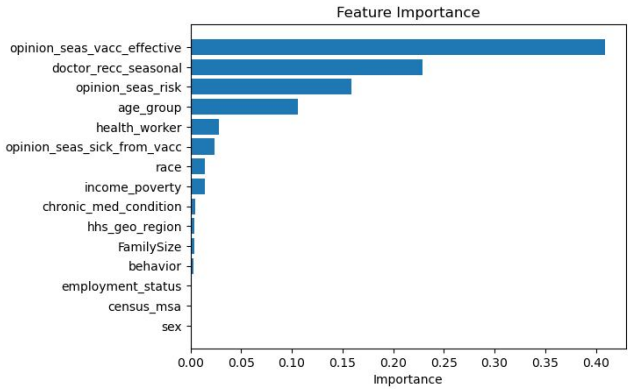


NEW FEATURES: Seasonal flu

Difference	Accuracy	F1:0	F1:1
Training	-0.0028	-0.005	0.0008
Test	0.0022	-0.0002	0.005







MODELLING

Which Models we used?

Decision Tree

Simple model.
Easy to explain [1]

Random Forest

Ensemble classifier, we wanted to improve the Decision Tree

XGBoost

Ensemble Classifier. We wanted to compare two ensemble classifier.

SVM

We wanted to make a comparison between Decision Tree and a more complex model

Logistic Regression

Simple and interpretable model [2]

Neural Network

Most complex models. Difficult interpretation.

[1] https://christophm.github.io/interpretable-ml-book/tree.html
[2] https://christophm.github.io/interpretable-ml-book/logistic.html

Hyperparameter Tuning

We split our dataset in 3 parts:
Train
Validation
Test

DATASET DIVISION

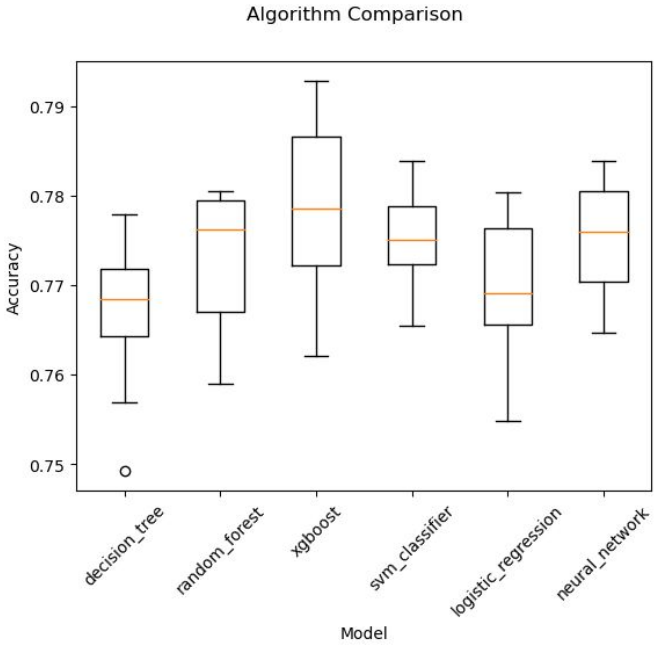
We used the scikit-learn Grid Search and Randomized Grid Search

LOOKING FOR HYPERPARAMETERS

We used the Accuracy as metric for the seasonal Dataset and F1-Score as metric for H1N1

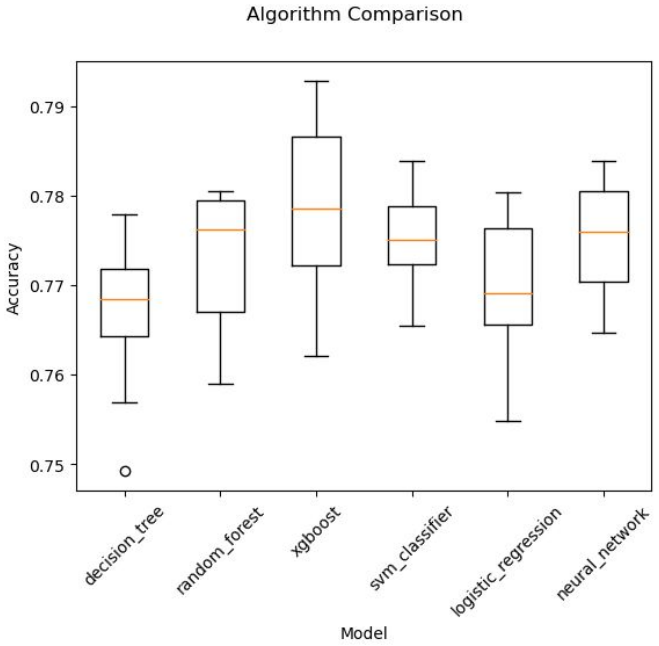
METRICS

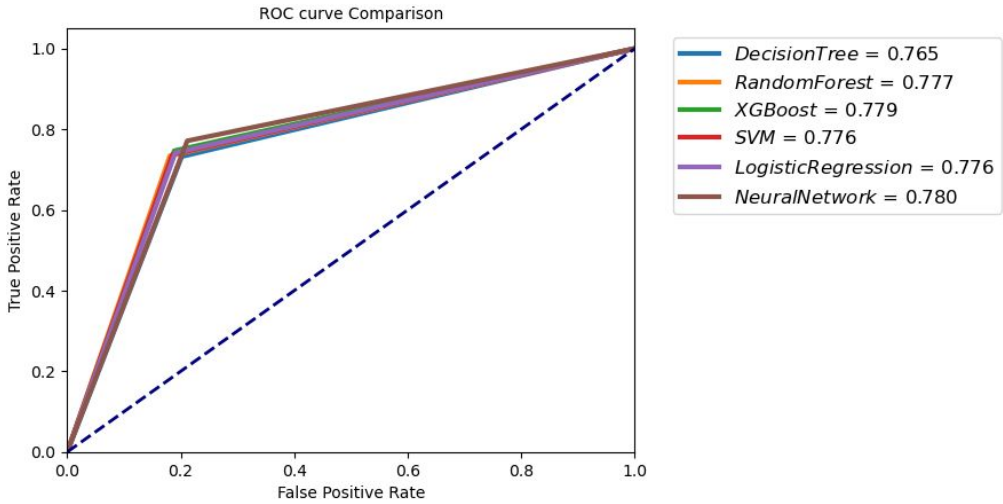
TESTING OUR MODELS – SEASONAL TARGET VARIABLE



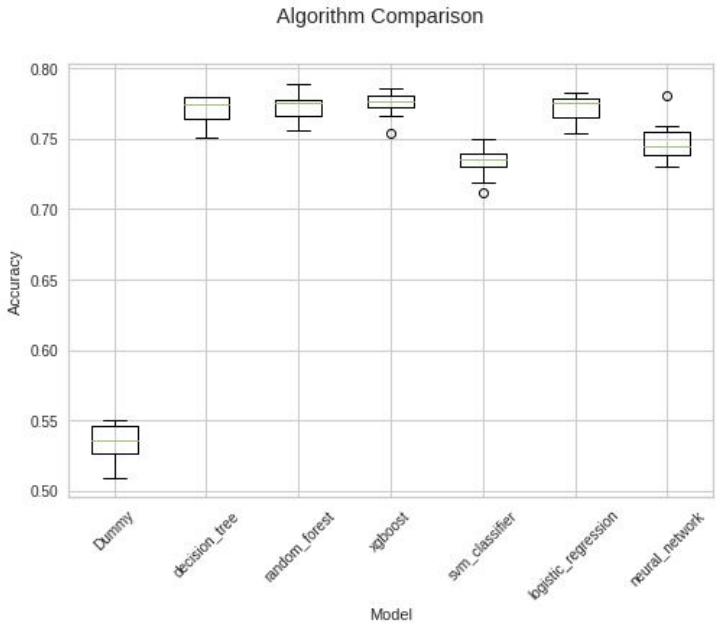
Model	Accuracy
Dummy model	0.50
Decision Tree	0.77
Random Forest	0.78
XGBoost	0.78
SVM	0.78
Logistic Regression	0.78
Neural Network	0.78

TESTING OUR MODELS – SEASONAL TARGET VARIABLE



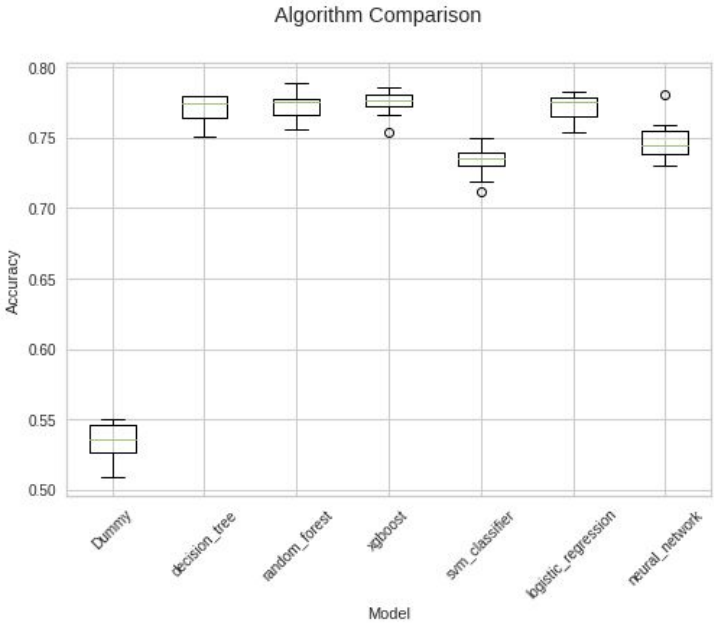


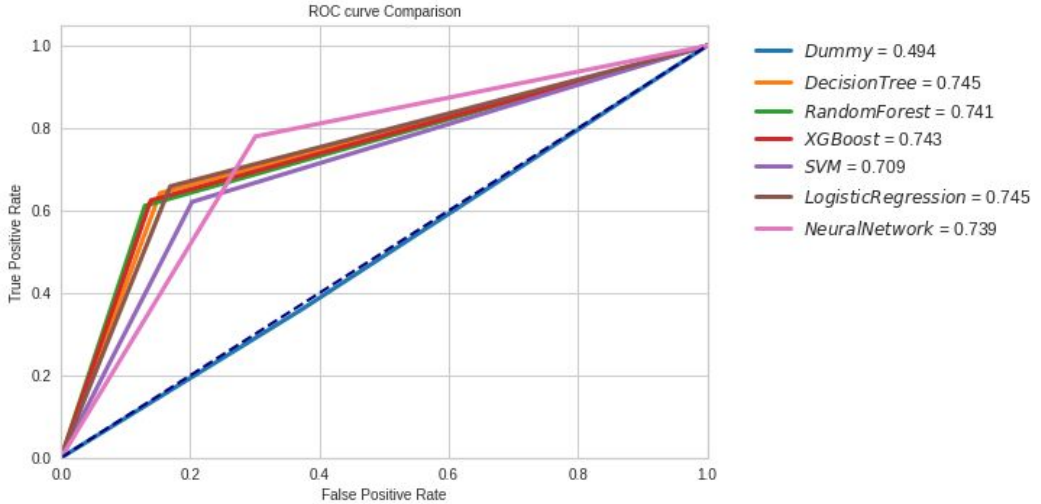
TESTING OUR MODELS - H1N1 TARGET VARIABLE



Model	Accuracy	F1-Score
Dummy model	0.58	0.29
Decision Tree	0.80	0.58
Random Forest	0.82	0.58
XGBoost	0.81	0.58
SVM	0.81	0.56
Logistic Regression	0.79	0.58
Neural Network	0.73	0.54

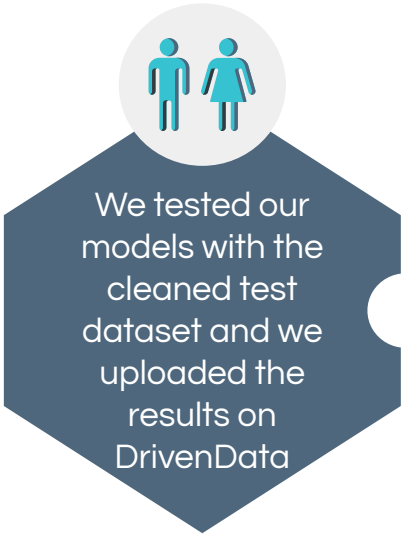
TESTING OUR MODELS – H1N1 TARGET VARIABLE





TESTING OUR MODELS ON DRIVENDATA

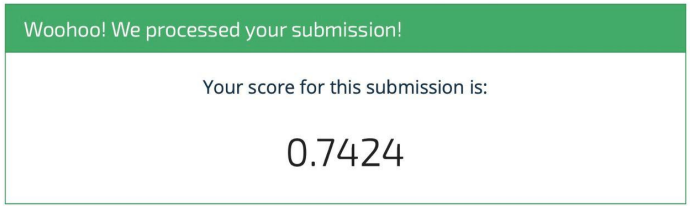
PIPELINE



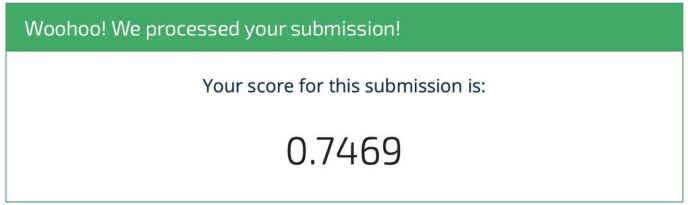
TEST

TESTING OUR MODELS ON DRIVENDATA

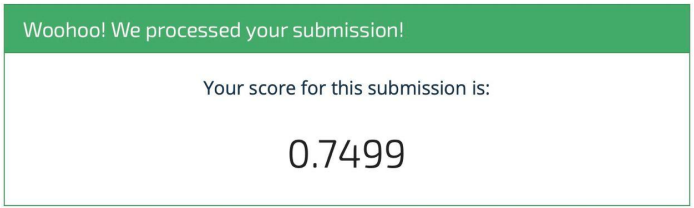
Decision Tree



Random Forest



Neural Network



1455
Competitors

373
Current Rank

0.8658
Best Public AUROC

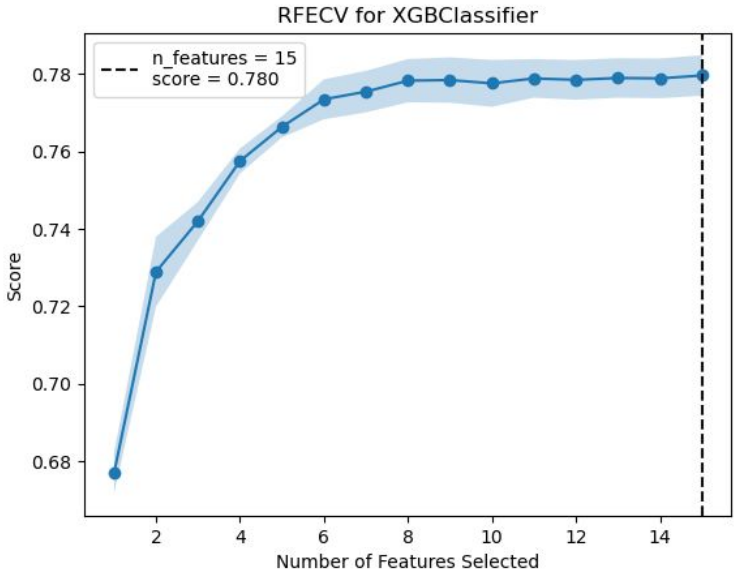
FEATURES SELECTION

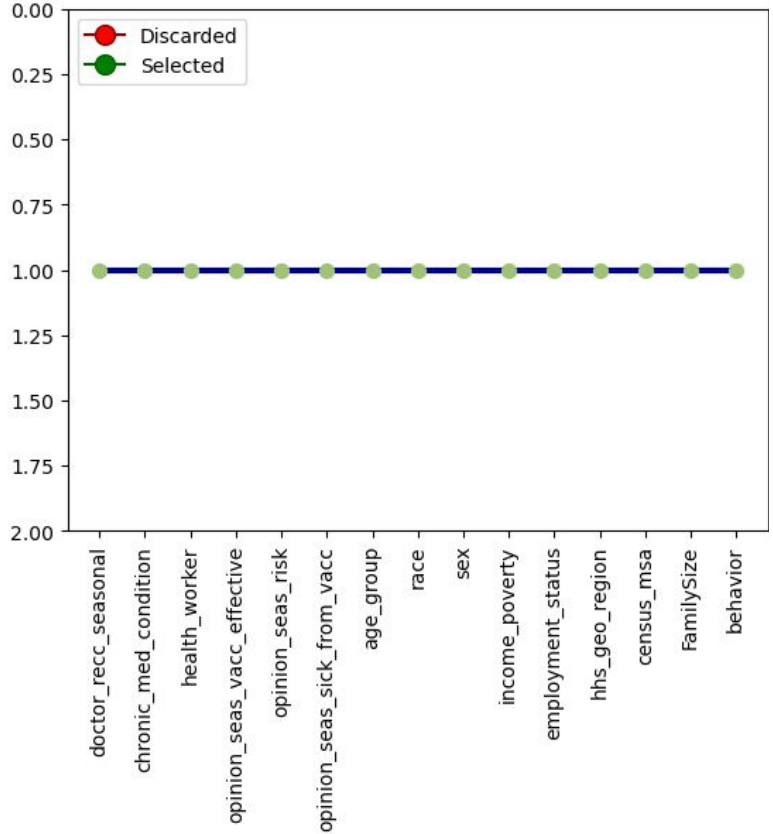
FEATURES SELECTION

SELECTION

FEATURES SELECTION: Seasonal flu XGBoost

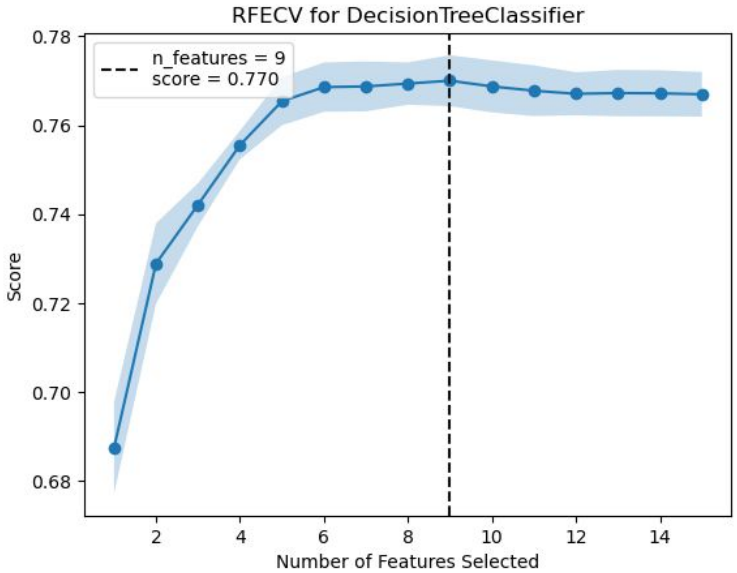
Accuracy	F1:0	F1:1
0.78	0.80	0.76

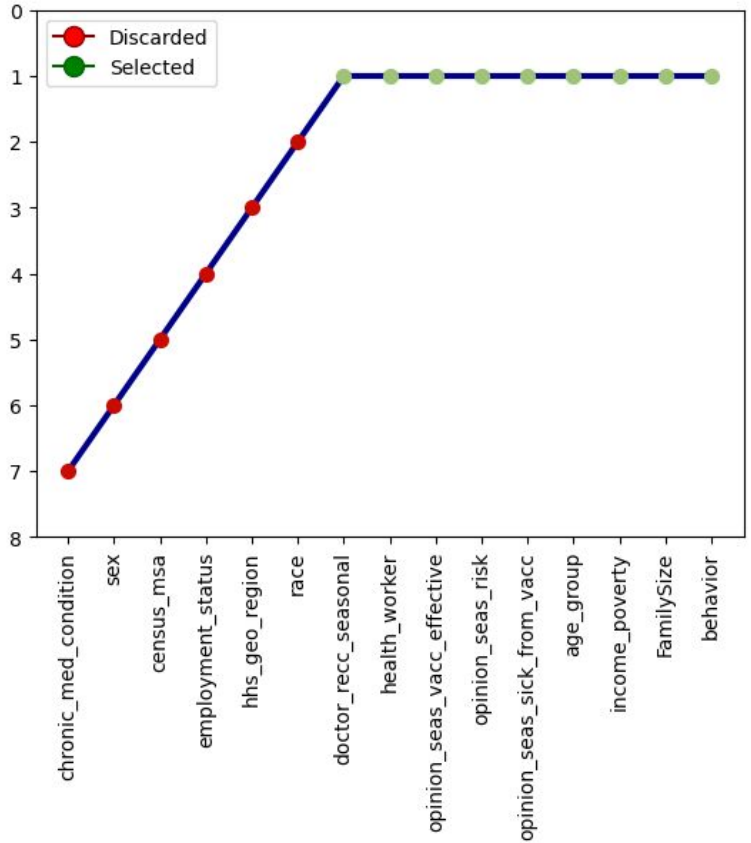




FEATURES SELECTION: Seasonal flu DECISION TREE

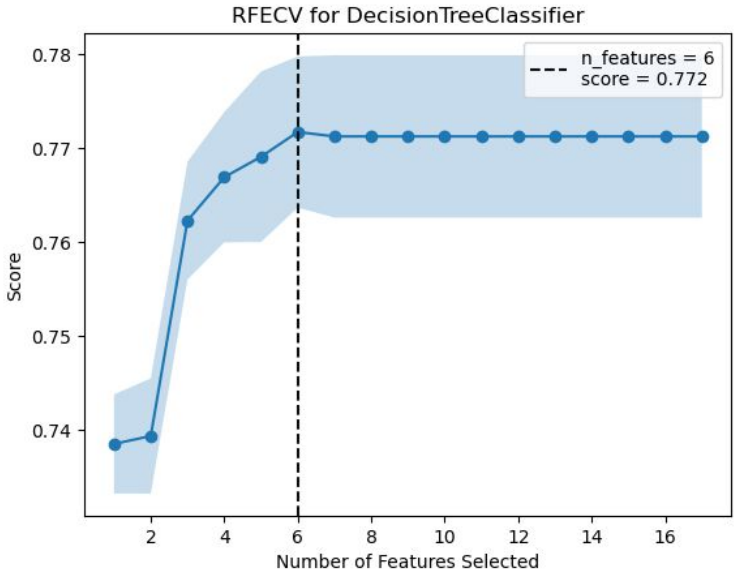
Accuracy	F1:0	F1:1
0.76	0.78	0.75

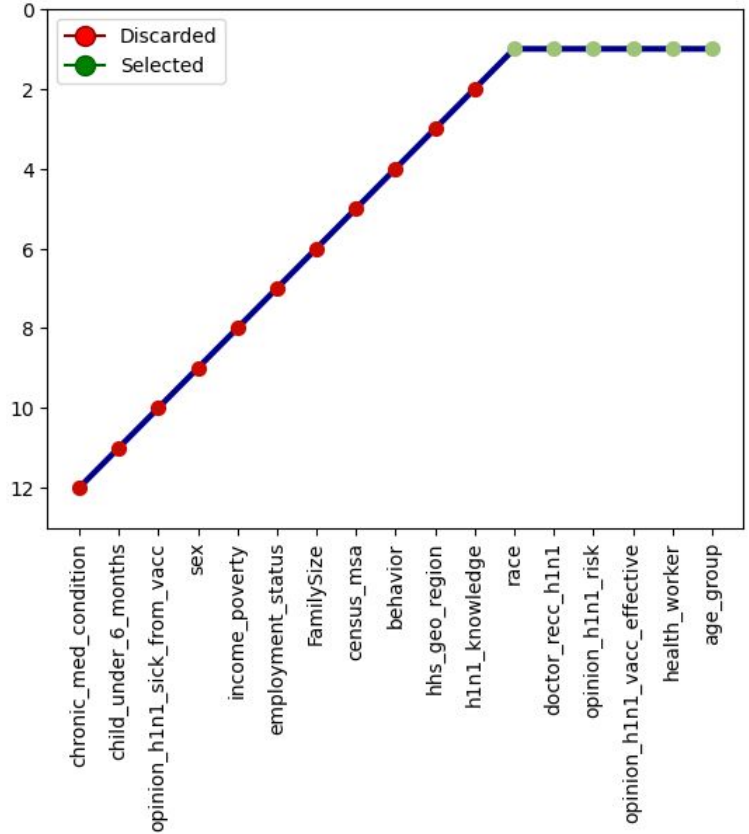




FEATURES SELECTION: H1N1 DECISION TREE

Accuracy	F1:0	F1:1
0.804	0.87	0.58





CONCLUSIONS

The second Milestone allowed us to develop and test several different models giving us the opportunity to reason about them in term of performances but also in term of interpretability.

THANKS!

Do you have any questions?

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.

REFERENCES

- U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics. The National 2009 H1N1 Flu Survey. Hyattsville, MD: Centers for Disease Control and Prevention, 2012.
- Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines