COVID-19 MORTALITY PREDICTION USING MACHINE LEARNING MODELS

: A COMPARITIVE STUDY

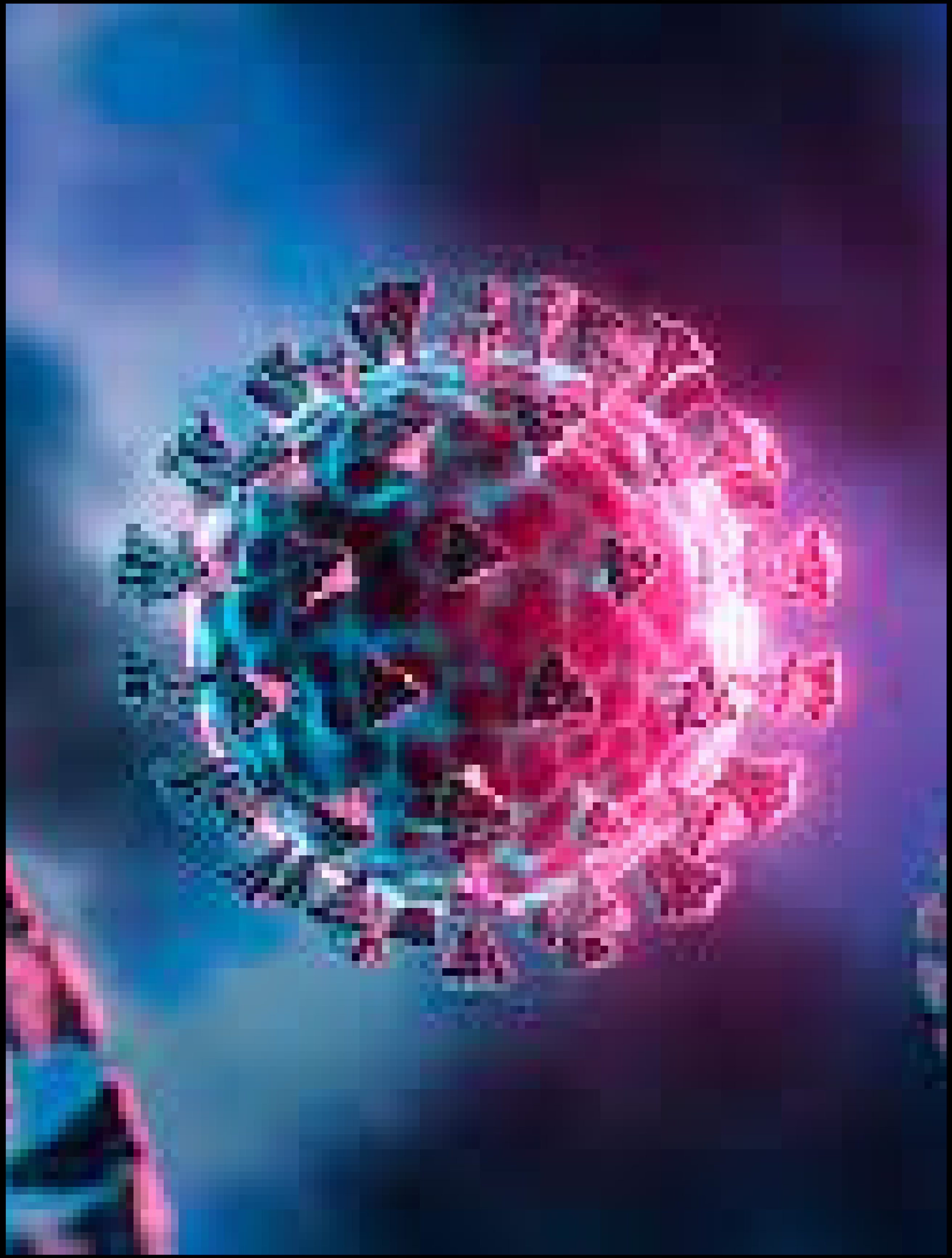
INTRODUCTION

COVID-19, caused by the highly infectious SARS-CoV-2 virus, has been declared a global public health emergency by the WHO. Epidemiological models have been deployed for outbreak prediction, peak estimation and mortality rate prediction.

The goal of this project is to provide assistance to medical units based on critical situations such as the following:

During the beginning of a pandemic wave, medical units generally do not have issues with medical infrastructures such as beds and medical supplies. In this scenario, we propose to use one of the mentioned models that have a better recall score (where a patient is considered critical even if there is a small chance that he might still recover later).

As the pandemic reaches a peak, medical supplies and infrastructure become sparse. In such scenarios, medical units can switch to a model with higher precision (to ensure the most needed patients are attended soon).



DATA PRE-PROCESSING

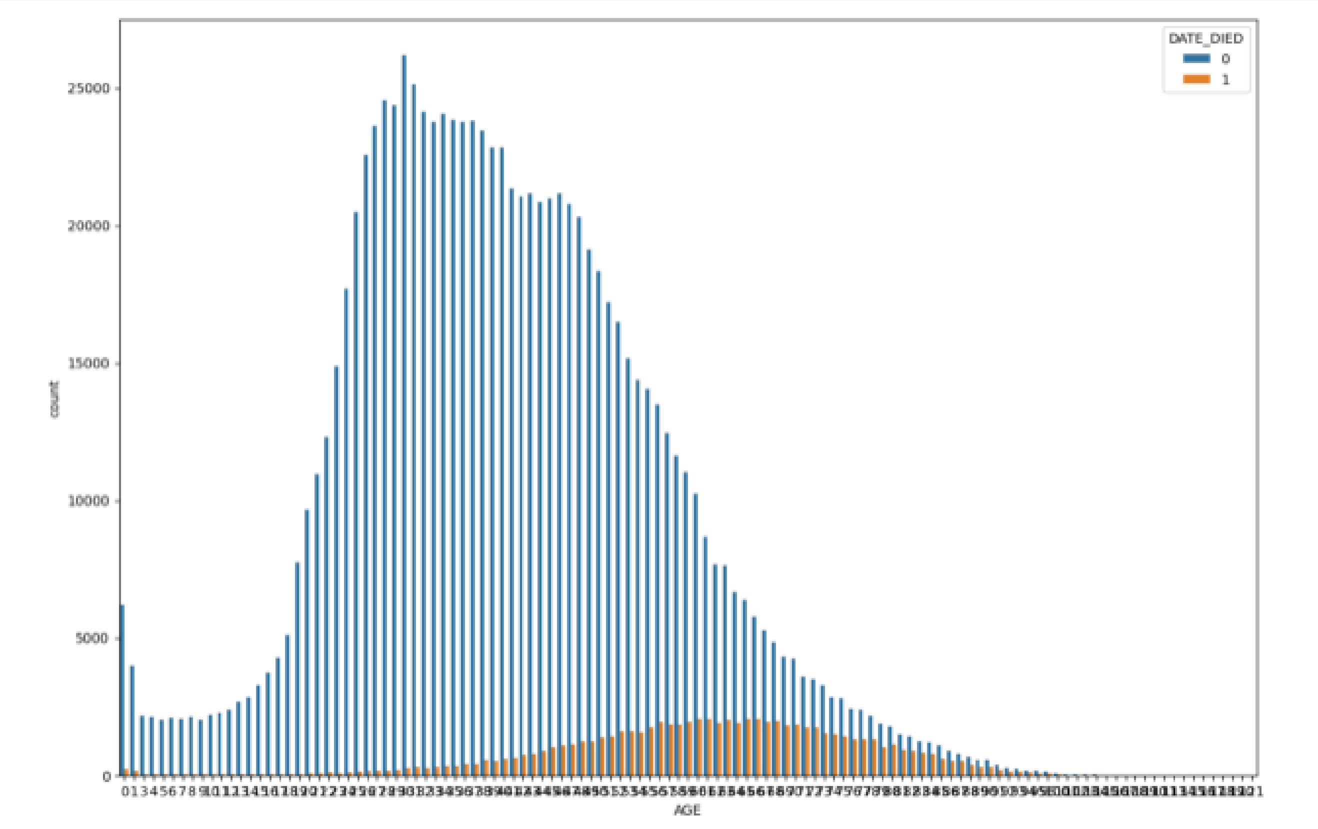
01 LOAD DATASET

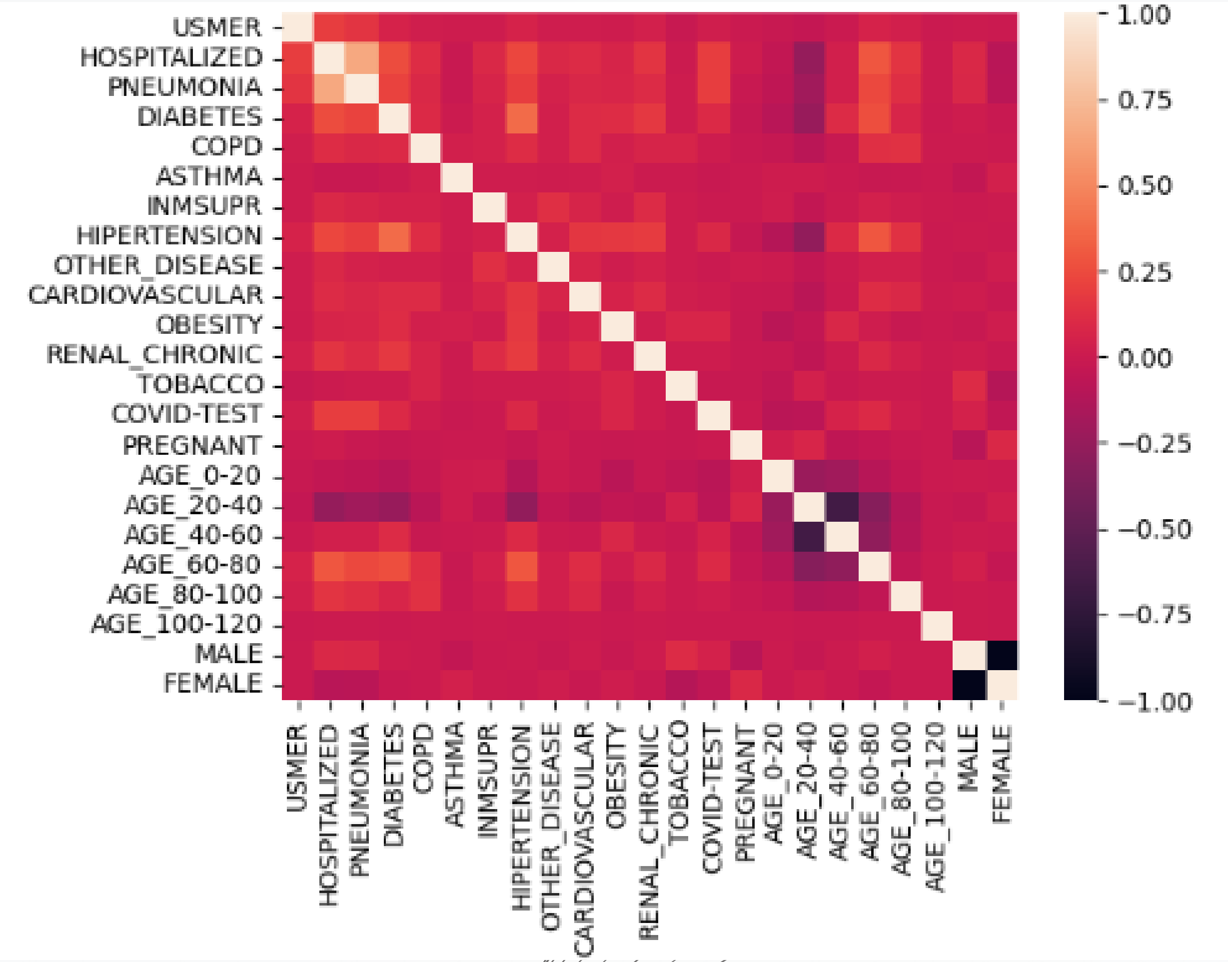
02 FINDING RELATIONS

03 PRE-PROCESSING

04 FINAL FORMATTING

	Total_missing_count	Percentage_missing
ICU	856032	81.64
INTUBED	855869	81.62
PREGNANT	527265	50.28
PNEUMONIA	16003	1.53
OTHER_DISEASE	5045	0.48
INMSUPR	3404	0.32
DIABETES	3338	0.32
TOBACCO	3220	0.31
HIPERTENSION	3104	0.30
RENAL_CHRONIC	3006	0.29
OBESITY	3032	0.29
CARDIOVASCULAR	3076	0.29
COPD	3003	0.29
ASTHMA	2979	0.28
AGE	345	0.03
MEDICAL_UNIT	0	0.00
DATE_DIED	0	0.00
PATIENT_TYPE	0	0.00
SEX	0	0.00
CLASIFFICATION_FINAL	0	0.00
USMER	0	0.00





FEATURE SELECTION AND DATA RE-SAMPLING

Not PCA

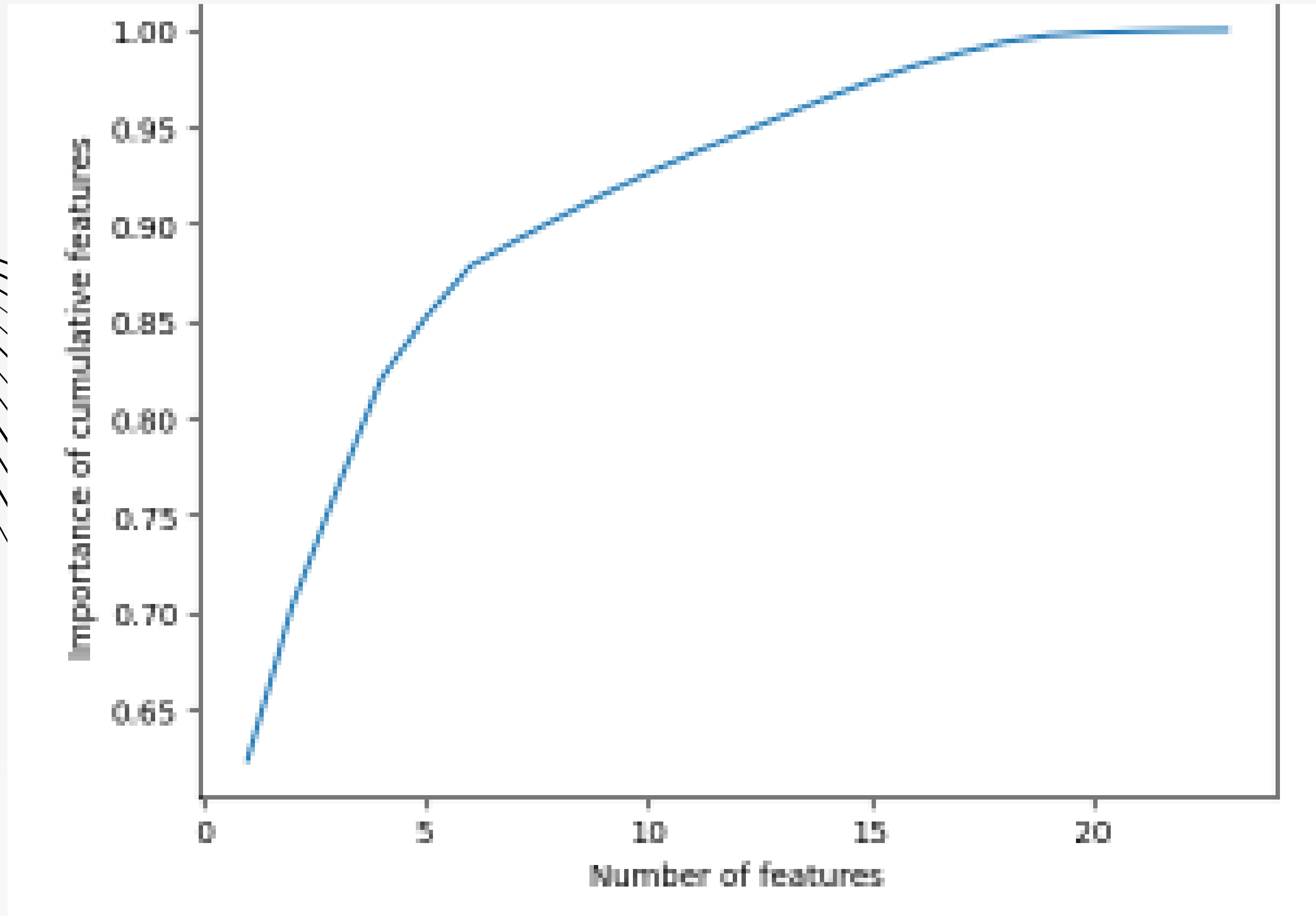
Principal component analysis works best when the feature set has continous values. In our case, all features are binary values and therefore PCA is ruled out

Decision Trees

We use Decision trees with a higher max_depth value so as get the feature_importance based on the gini index

Resampling

We use SMOTE and RandomUnderSampler modules provided by imblearn package to under sample the majority class and oversample the minority class




```
np.cumsum(list(important_features.values()))
```

```
array([0.62368734, 0.70295856, 0.76285882, 0.8281737 , 0.85182216,
       0.87815796, 0.89113641, 0.9033984 , 0.91524446, 0.92629333,
       0.93638829, 0.94590016, 0.95553603, 0.96485467, 0.9736132 ,
       0.9815581 , 0.98816025, 0.99380038, 0.99713923, 0.99850495,
       0.99927187, 0.99981512, 1.        ])
```

```
best_features = list(important_features.keys())[:np.argmax(np.cumsum(list(important_features.values())) >= 0.9)+1]
best_features
```

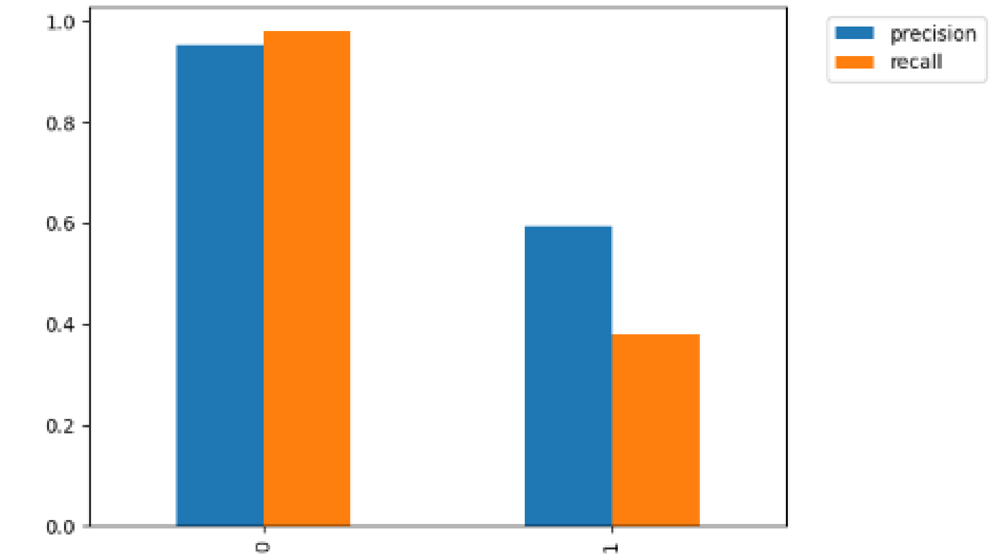
```
['HOSPITALIZED',
 'AGE_60-80',
 'PNEUMONIA',
 'COVID-TEST',
 'AGE_80-100',
 'AGE_40-60',
 'HIPERTENSION',
 'DIABETES']
```

MODELS

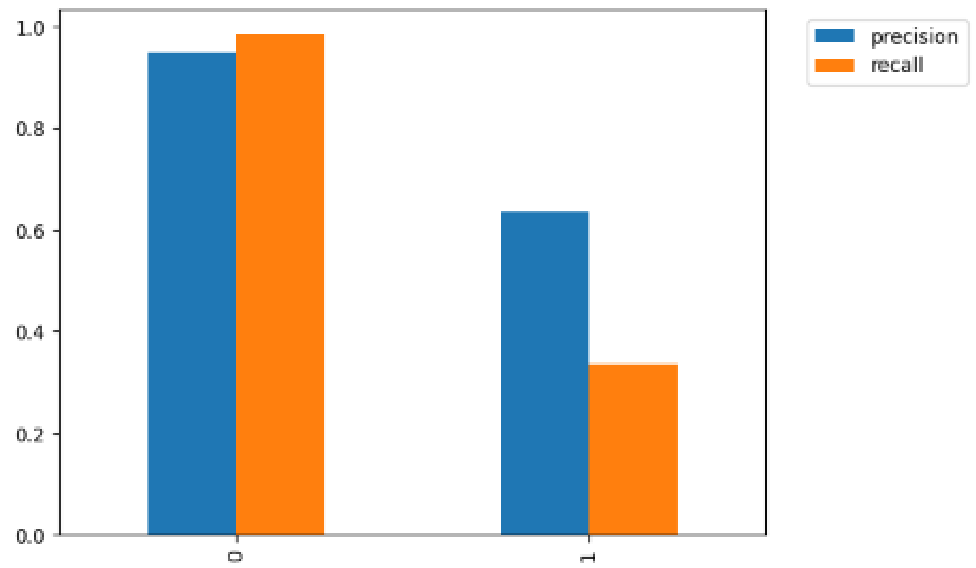
XG- BOOST

Unsampled training set

	0	1	accuracy	macro avg	weighted avg
precision	0.952452	0.595140	0.935755	0.773796	0.926319
recall	0.979589	0.380246	0.935755	0.679918	0.935755
f1-score	0.965830	0.464021	0.935755	0.714925	0.929129
support	236734.000000	18680.000000	0.935755	255414.000000	255414.000000



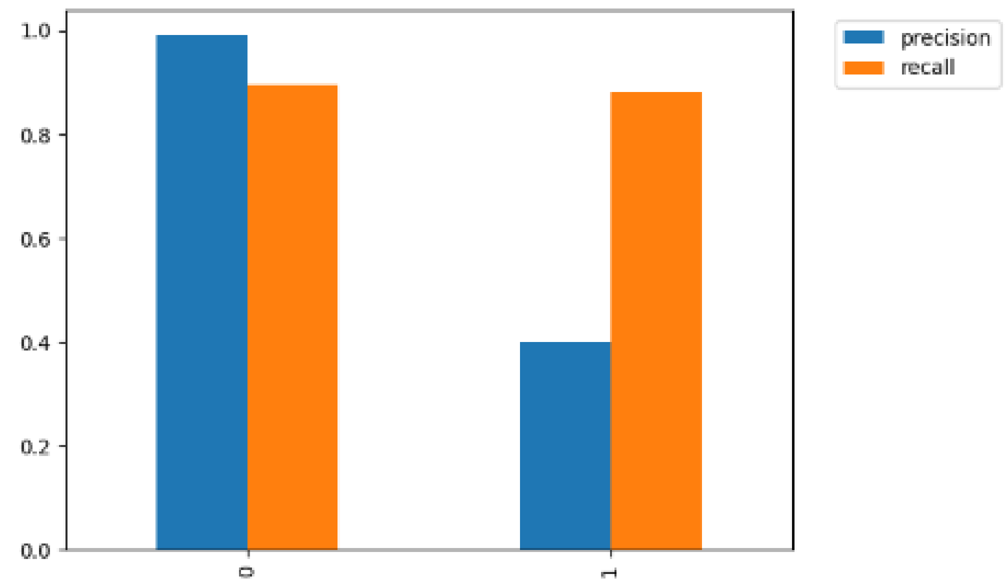
	0	1	accuracy	macro avg	weighted avg
precision	0.949559	0.634906	0.937341	0.792233	0.926547
recall	0.984704	0.337099	0.937341	0.660901	0.937341
f1-score	0.966813	0.440380	0.937341	0.703596	0.928311
support	236734.000000	18680.000000	0.937341	255414.000000	255414.000000



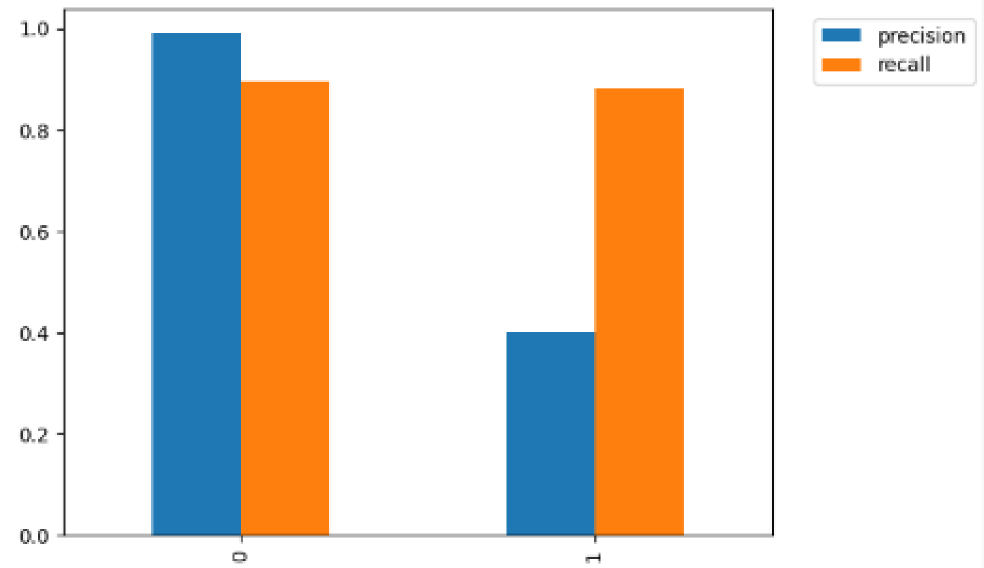
Unsampled training set with reduced feature set

Re-sampled training set

	0	1	accuracy	macro avg	weighted avg
precision	0.989671	0.399098	0.894262	0.694385	0.946479
recall	0.895262	0.881585	0.894262	0.888423	0.894262
f1-score	0.940102	0.549455	0.894262	0.744779	0.911532
support	236734.000000	18680.000000	0.894262	255414.000000	255414.000000



	0	1	accuracy	macro avg	weighted avg
precision	0.989671	0.399098	0.894262	0.694385	0.946479
recall	0.895262	0.881585	0.894262	0.888423	0.894262
f1-score	0.940102	0.549455	0.894262	0.744779	0.911532
support	236734.000000	18680.000000	0.894262	255414.000000	255414.000000



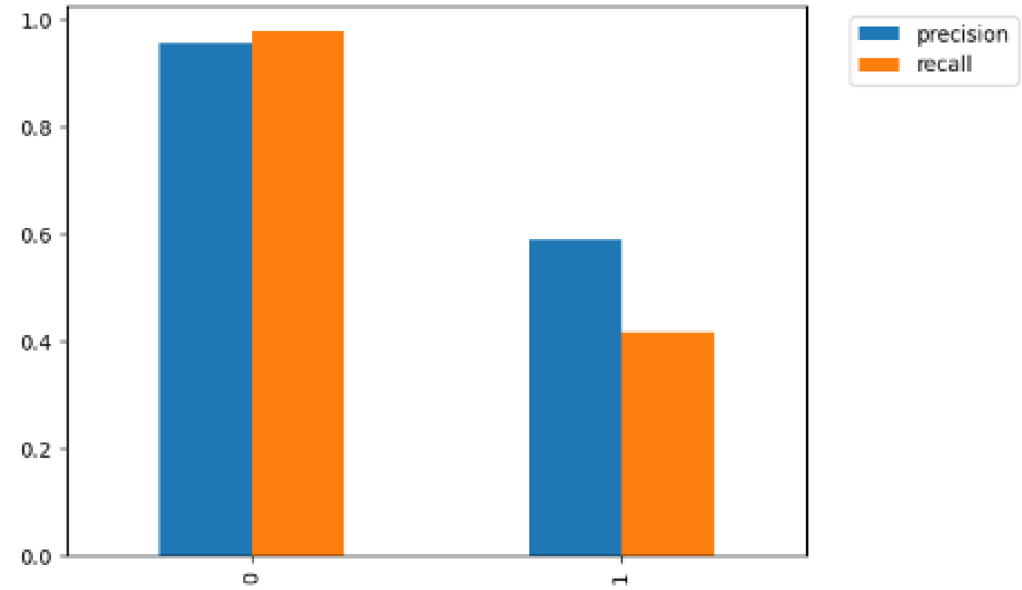
Resampled Training set with reduced features

MODELS

LOGISTIC REGRESSION

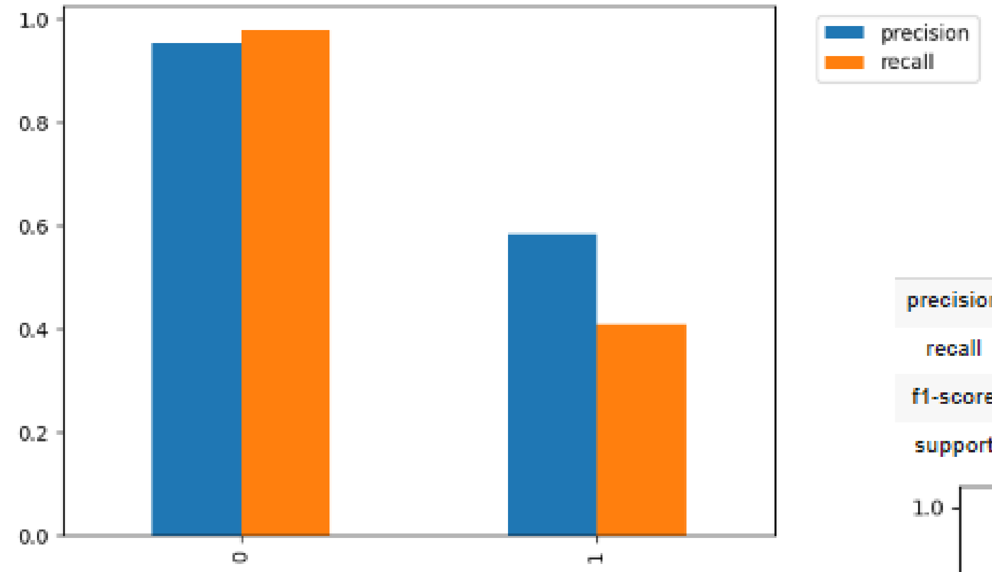
Unsampled training set

	0	1	accuracy	macro avg	weighted avg
precision	0.955089	0.590108	0.936194	0.772599	0.928396
recall	0.977105	0.417719	0.936194	0.697412	0.936194
f1-score	0.965972	0.489170	0.936194	0.727571	0.931100
support	236734.000000	18680.000000	0.936194	255414.000000	255414.000000

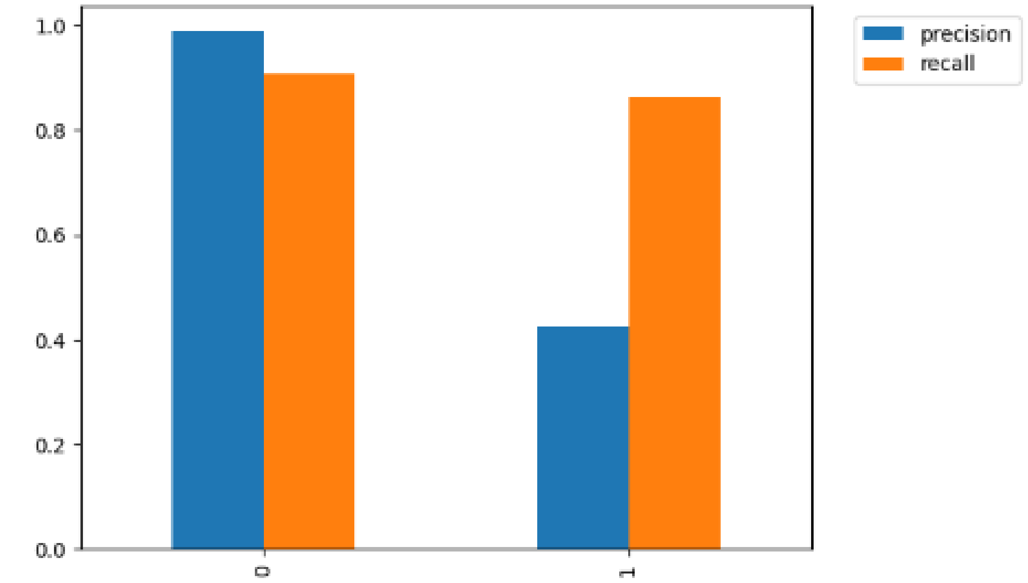


Unsampled training set with reduced features

	0	1	accuracy	macro avg	weighted avg
precision	0.954538	0.584860	0.935571	0.769699	0.927501
recall	0.977021	0.410278	0.935571	0.693649	0.935571
f1-score	0.965648	0.482255	0.935571	0.723952	0.930295
support	236734.000000	18680.000000	0.935571	255414.000000	255414.000000

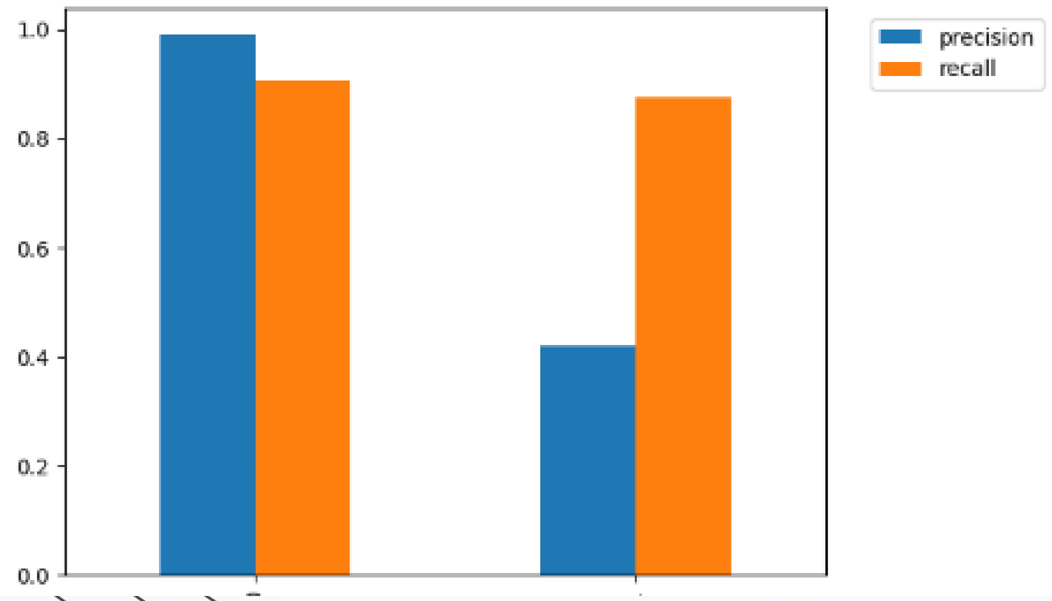


	0	1	accuracy	macro avg	weighted avg
precision	0.988261	0.426061	0.90495	0.707161	0.947144
recall	0.908239	0.863276	0.90495	0.885758	0.904950
f1-score	0.946562	0.570539	0.90495	0.758550	0.919061
support	236734.000000	18680.000000	0.90495	255414.000000	255414.000000



Resampled Training set with reduced features

	0	1	accuracy	macro avg	weighted avg
precision	0.989136	0.420452	0.902676	0.704794	0.947544
recall	0.904935	0.874036	0.902676	0.889486	0.902676
f1-score	0.945164	0.567777	0.902676	0.756471	0.917563
support	236734.000000	18680.000000	0.902676	255414.000000	255414.000000



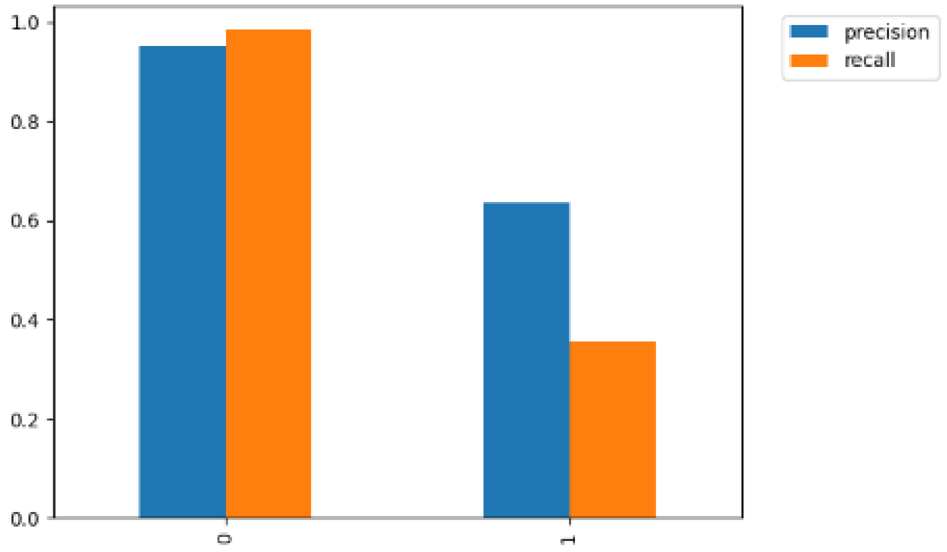
Re-sampled training set

MODELS

DECISION-TREE

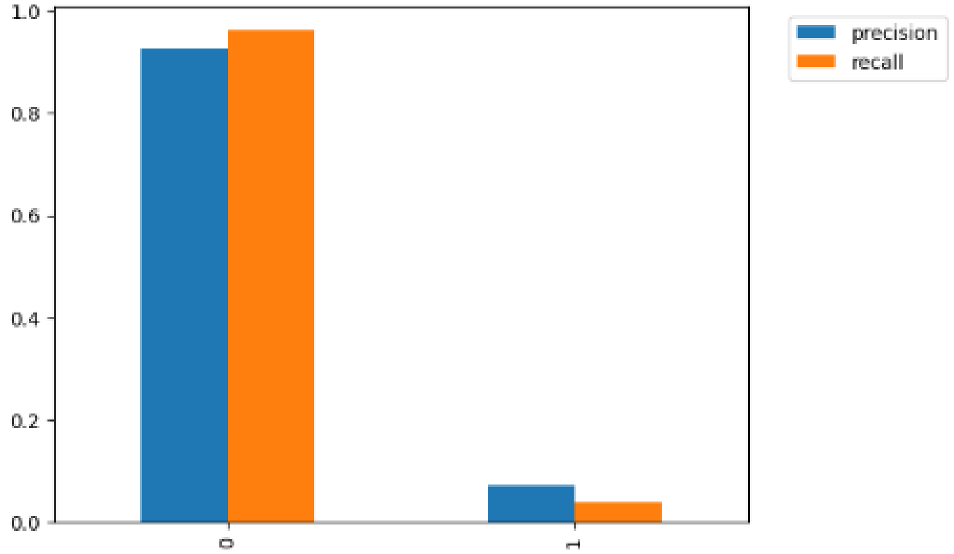
Unsampled training set

	0	1	accuracy	macro avg	weighted avg
precision	0.950329	0.635708	0.937345	0.793019	0.927093
recall	0.983767	0.355227	0.937345	0.669497	0.937345
f1-score	0.966759	0.455773	0.937345	0.711266	0.929019
support	236550.000000	18864.000000	0.937345	255414.000000	255414.000000



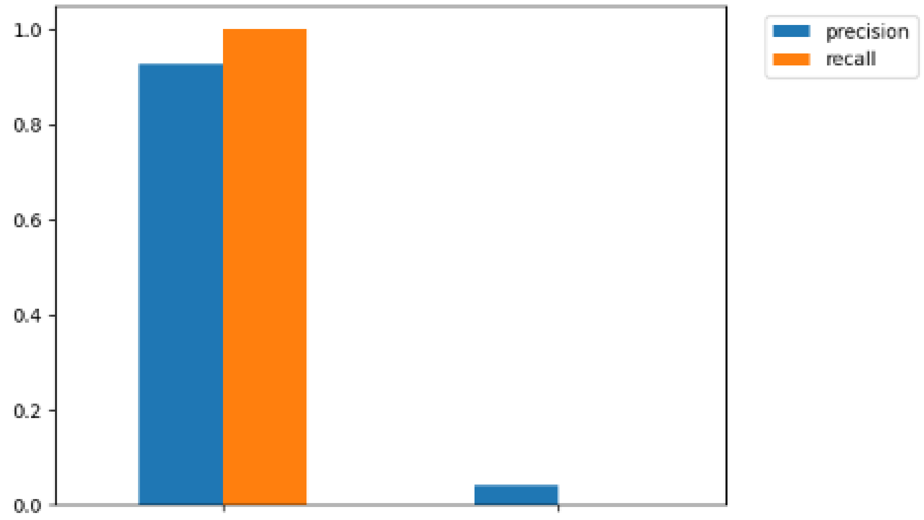
Unsampled training set with reduced feature set

	0	1	accuracy	macro avg	weighted avg
precision	0.926116	0.073185	0.892989	0.499651	0.863122
recall	0.961133	0.038486	0.892989	0.499809	0.892989
f1-score	0.943300	0.050445	0.892989	0.496872	0.877357
support	236550.000000	18864.000000	0.892989	255414.000000	255414.000000



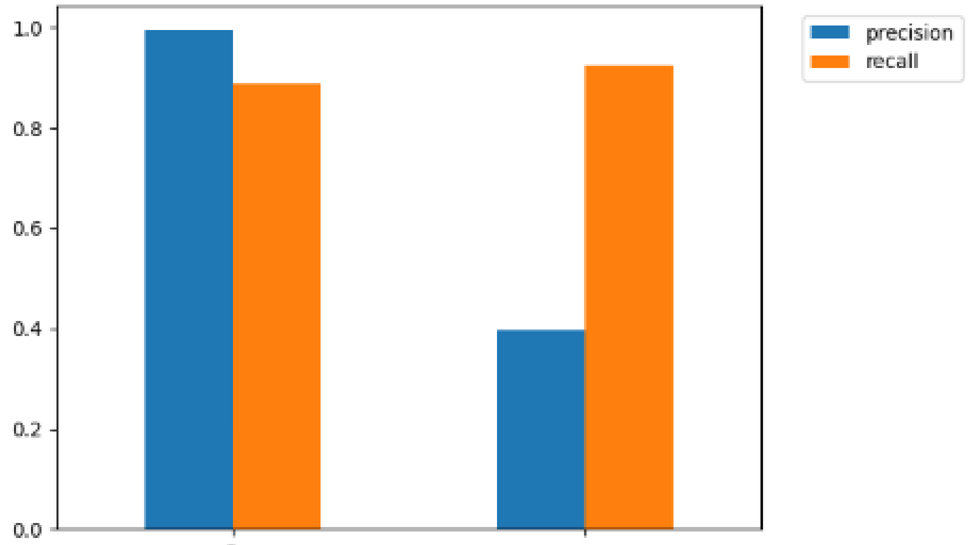
Re-sampled training set

	0	1	accuracy	macro avg	weighted avg
precision	0.926140	0.041667	0.926057	0.483904	0.860816
recall	0.999903	0.000053	0.926057	0.499978	0.926057
f1-score	0.961609	0.000106	0.926057	0.480858	0.890596
support	236550.000000	18864.000000	0.926057	255414.000000	255414.000000



Resampled Training set with reduced features

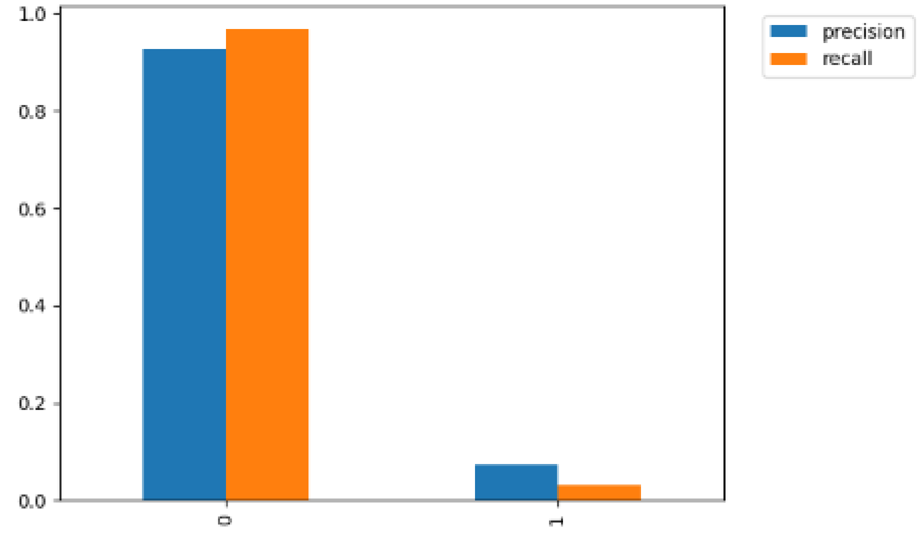
	0	1	accuracy	macro avg	weighted avg
precision	0.993298	0.395359	0.889983	0.694329	0.949136
recall	0.887195	0.924936	0.889983	0.906066	0.889983
f1-score	0.937253	0.553940	0.889983	0.745597	0.908943
support	236550.000000	18864.000000	0.889983	255414.000000	255414.000000



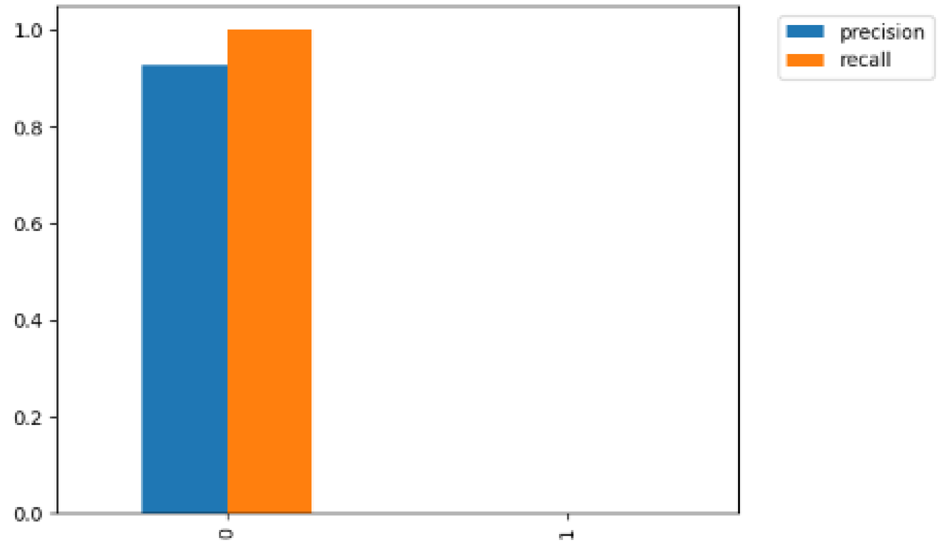
MODELS

Class-Specific Ensemble Model

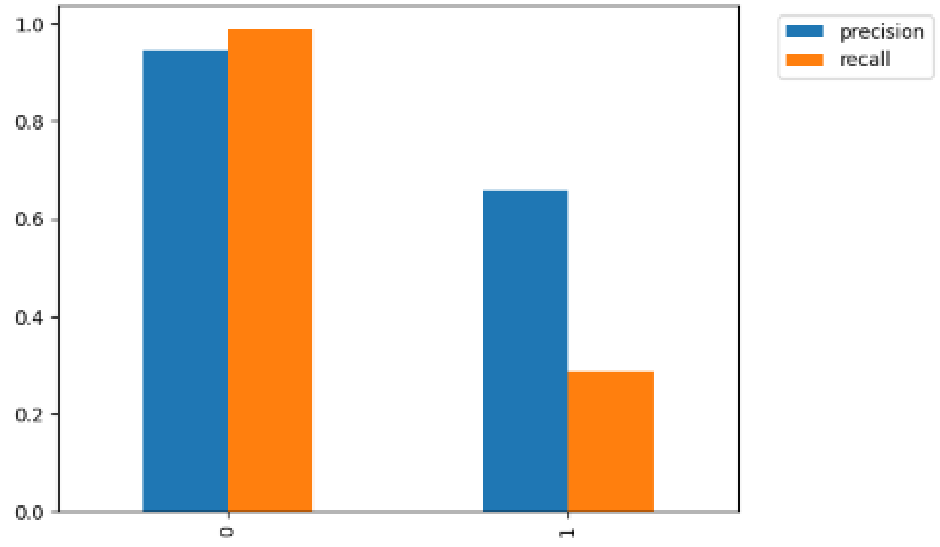
	0	1	accuracy	macro avg	weighted avg
precision	0.926107	0.072796	0.897836	0.499452	0.863084
recall	0.966832	0.032655	0.897836	0.499743	0.897836
f1-score	0.946031	0.045085	0.897836	0.495558	0.879490
support	236550.000000	18864.000000	0.897836	255414.000000	255414.000000



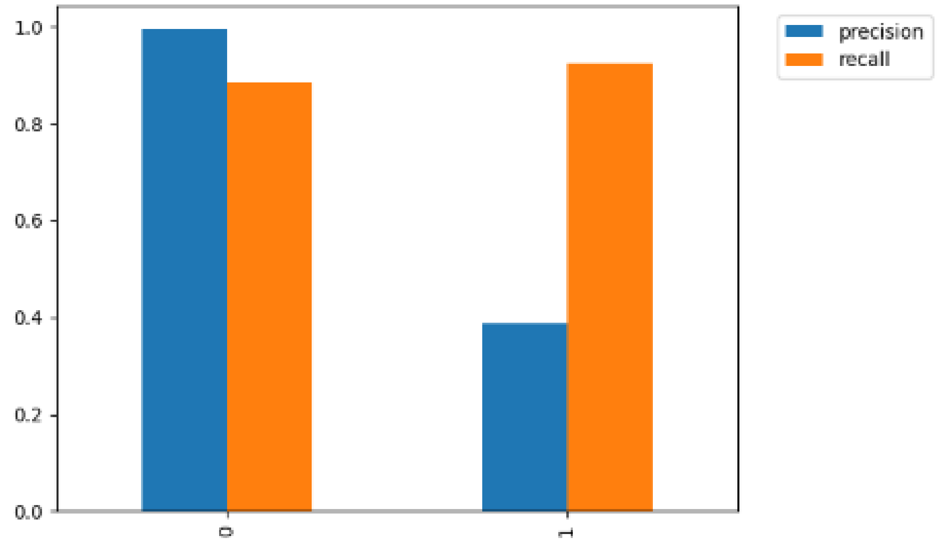
	0	1	accuracy	macro avg	weighted avg
precision	0.926143	0.0	0.926143	0.463072	0.857742
recall	1.000000	0.0	0.926143	0.500000	0.926143
f1-score	0.961656	0.0	0.926143	0.480828	0.890631
support	236550.000000	18864.0	0.926143	255414.000000	255414.000000

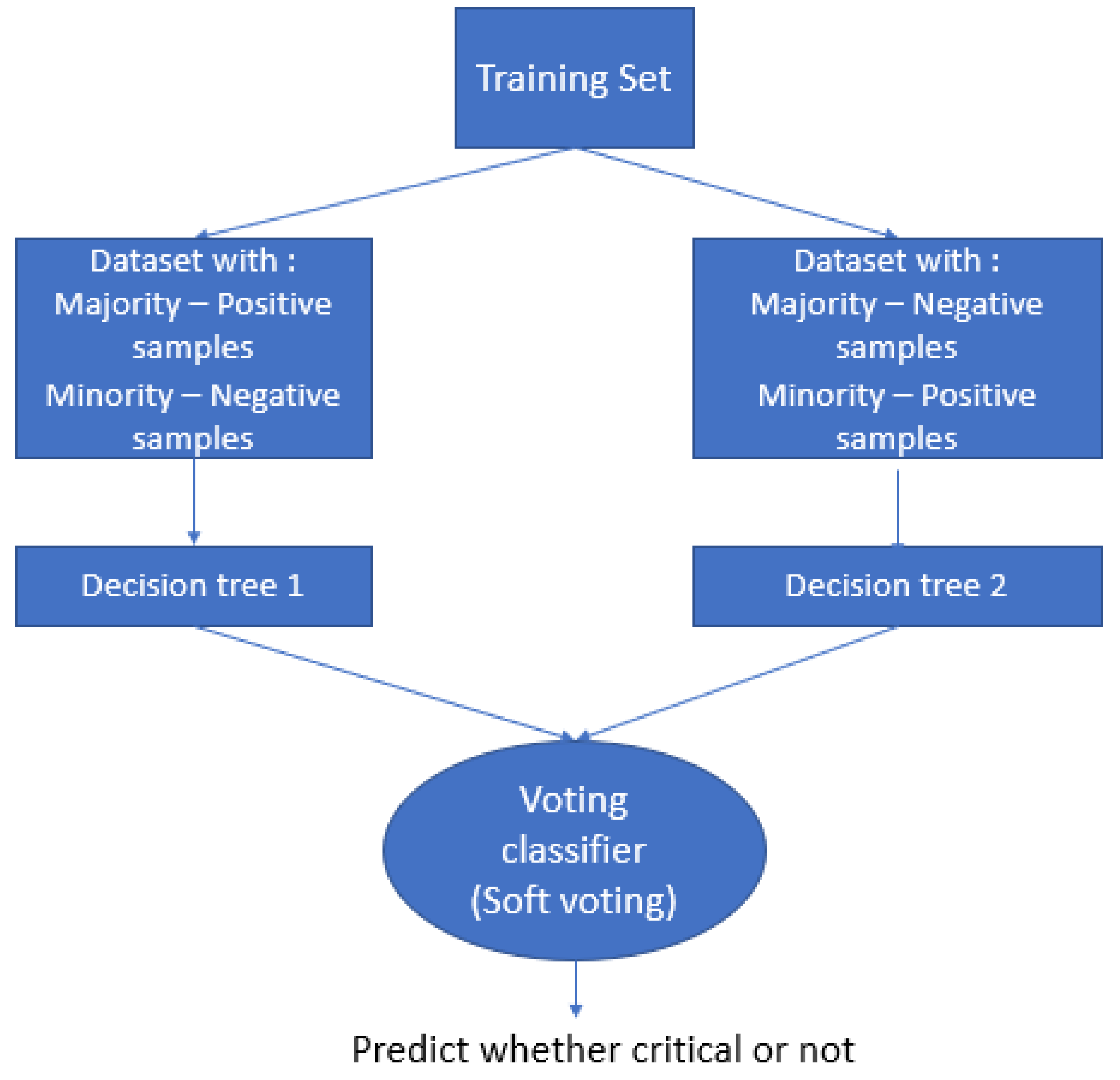


	0	1	accuracy	macro avg	weighted avg
precision	0.945528	0.659216	0.93635	0.802372	0.924382
recall	0.988205	0.286101	0.93635	0.637153	0.936350
f1-score	0.966396	0.399024	0.93635	0.682710	0.924492
support	236550.000000	18864.000000	0.93635	255414.000000	255414.000000



	0	1	accuracy	macro avg	weighted avg
precision	0.993250	0.388589	0.886983	0.690920	0.948592
recall	0.883978	0.924671	0.886983	0.904325	0.886983
f1-score	0.935434	0.547214	0.886983	0.741324	0.906761
support	236550.000000	18864.000000	0.886983	255414.000000	255414.000000

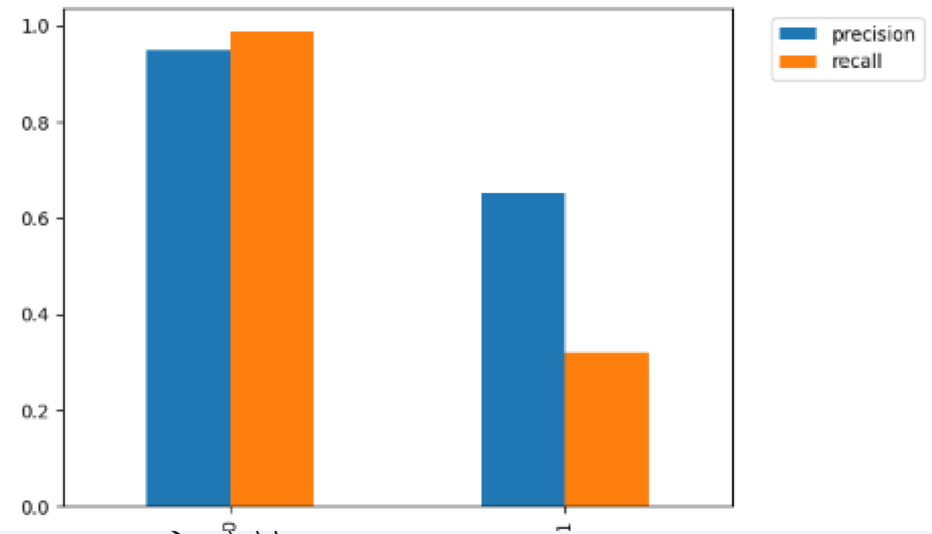




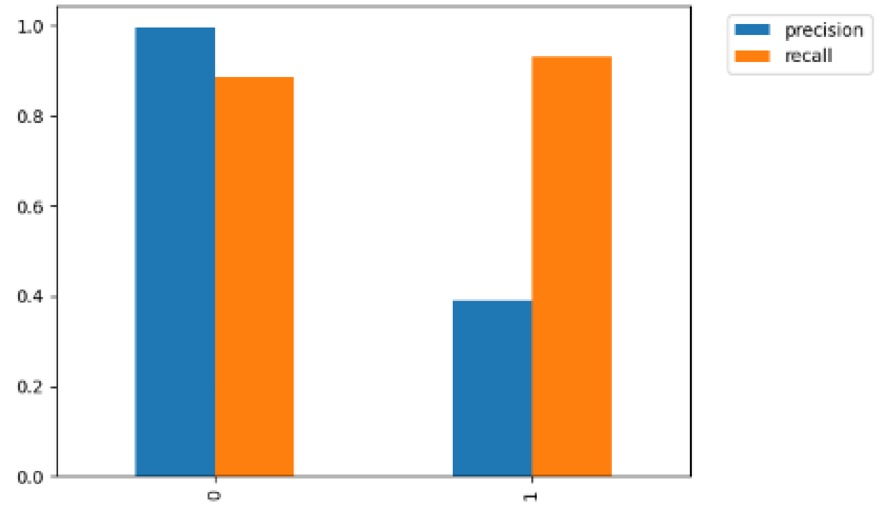
MODELS

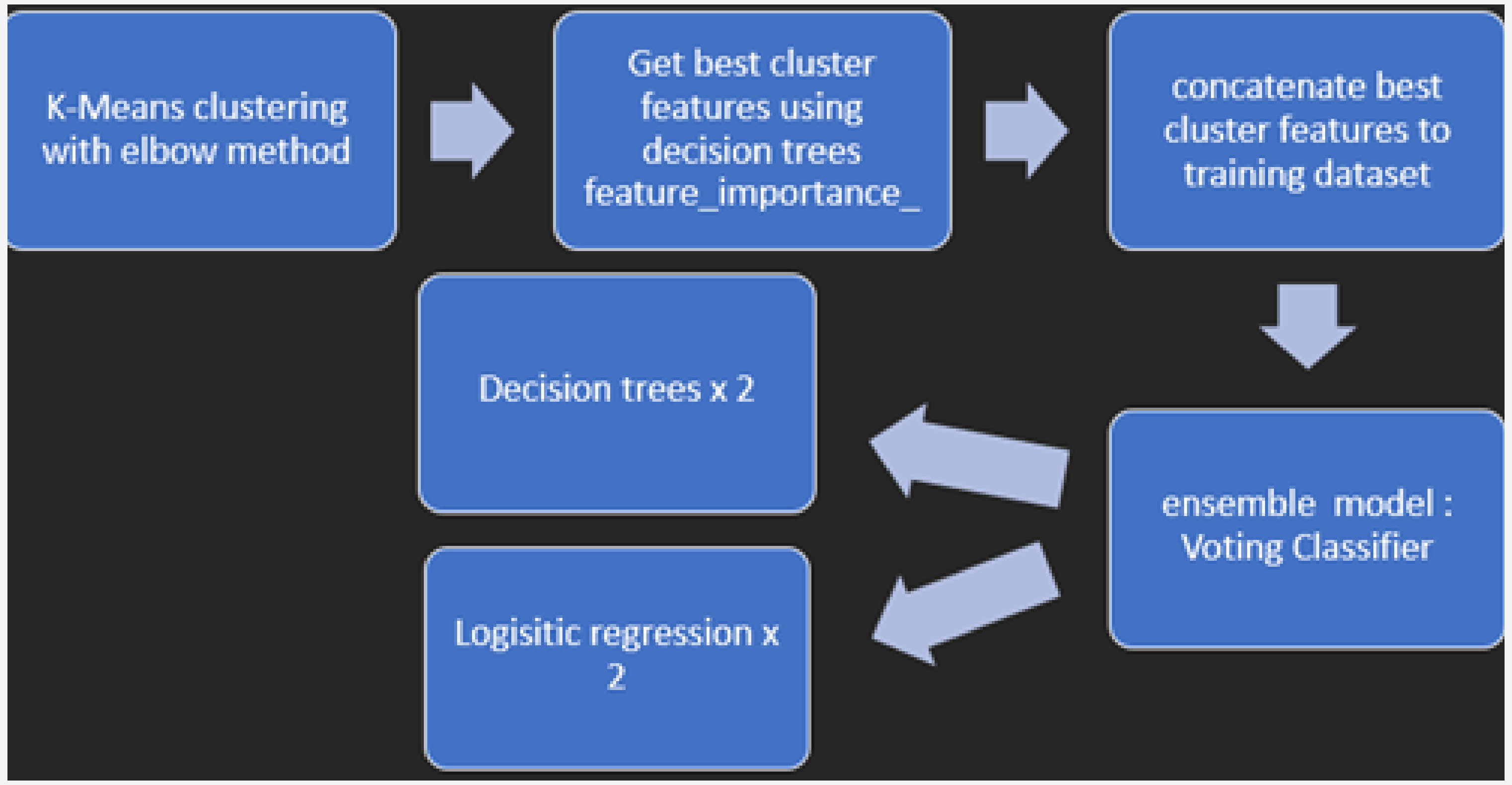
Hybrid Ensemble Model

	0	1	accuracy	macro avg	weighted avg
precision	0.947923	0.650495	0.937098	0.799209	0.925956
recall	0.986265	0.320558	0.937098	0.653411	0.937098
f1-score	0.966714	0.429474	0.937098	0.698094	0.927035
support	236550.000000	18864.000000	0.937098	255414.000000	255414.000000

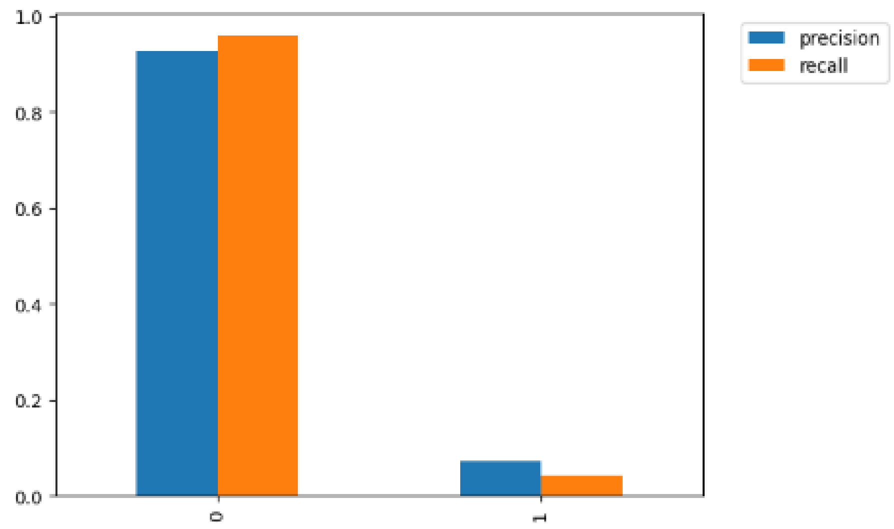


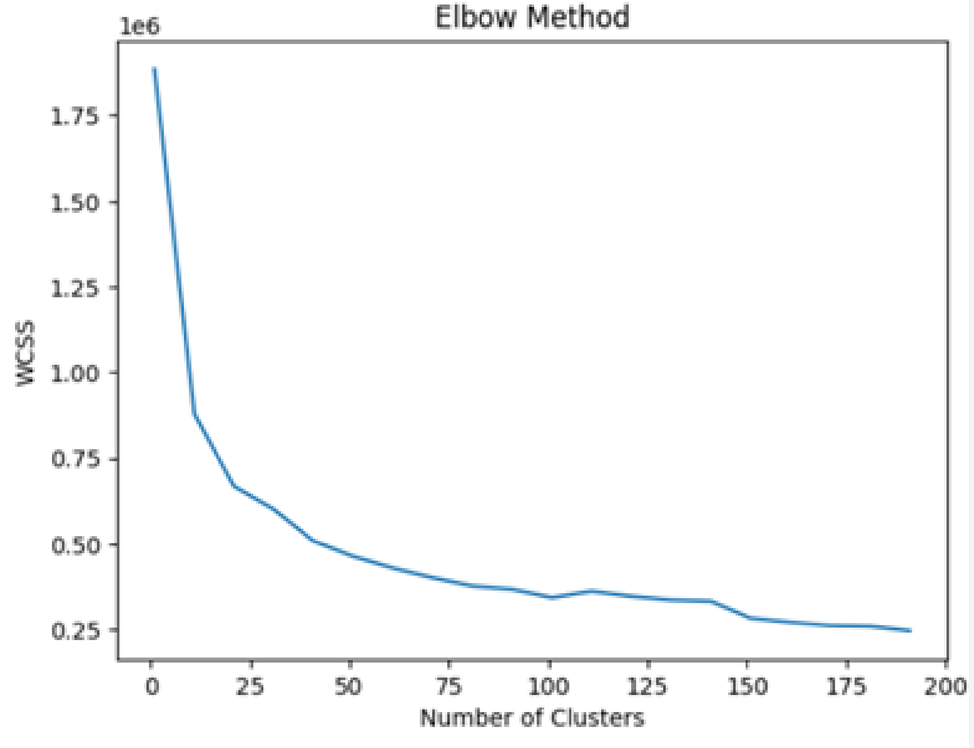
	0	1	accuracy	macro avg	weighted avg
precision	0.993650	0.390731	0.887763	0.692190	0.949121
recall	0.884464	0.929124	0.887763	0.906794	0.887763
f1-score	0.935883	0.550117	0.887763	0.743000	0.907392
support	236550.000000	18864.000000	0.887763	255414.000000	255414.000000



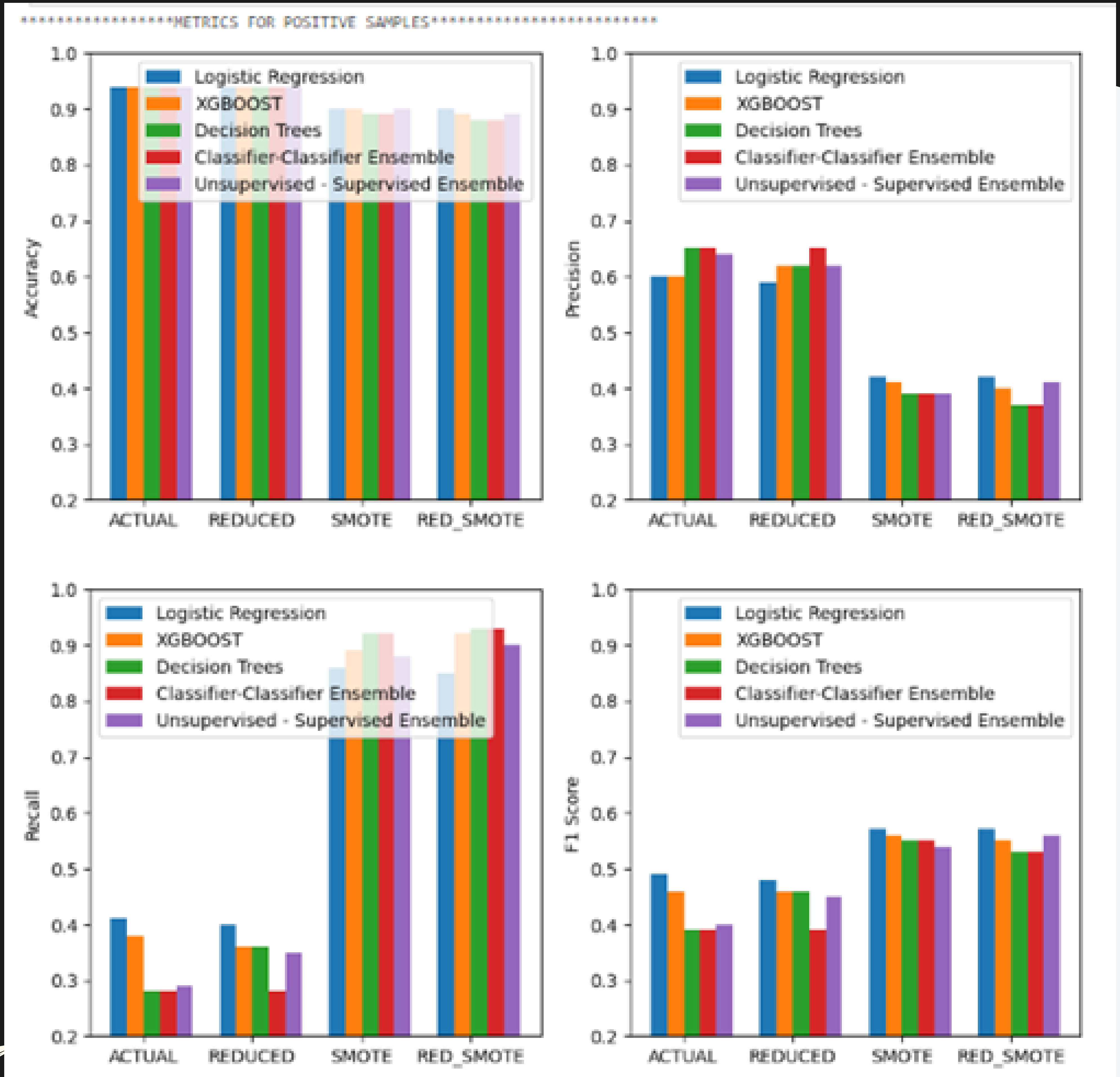


	0	1	accuracy	macro avg	weighted avg
precision	0.926058	0.071912	0.889971	0.498985	0.862973
recall	0.957662	0.041137	0.889971	0.499399	0.889971
f1-score	0.941595	0.052335	0.889971	0.496965	0.875917
support	236550.000000	18864.000000	0.889971	255414.000000	255414.000000





STATISTICS


*****************METRICS FOR POSITIVE SAMPLES*************************
Logistic Regression
XGBOOST
Decision Trees
Classifier-Classifier Ensemble
Unsupervised - Supervised Ensemble
Accuracy
Precision
Recall
F1 Score
ACTUAL
REDUCED
SMOTE
RED_SMOTE
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

OUR TEAM

Anirudh S Bhargav

- *Data exploration and Preprocessing*
- *Modelling*
- *Hyperparameter tuning for Hybrid and class-specific ensemble*
- *Literature survey*
- *GitHub*
- *IEEE report*

Irlanki Sandeep

- *Data exploration and Preprocessin*
- *Modelling*
- *Hyperparameter tuning for Logestic regression*
- *Literature survey*
- *Plotting metrics*
- *IEEE report*

Rahul Chauhan

- *Data exploration and Preprocessing*
- *Modelling*
- *Hyperparameter tuning for Decision tree*
- *Literature survey*
- *Presentation*
- *IEEE report*

Rohan Sarnad

- *Data exploration and Preprocessing*
- *Modelling*
- *Hyperparameter tuning for XG Boost*
- *Literature survey*
- *Github*
- *IEEE report*

THANK

YOU