Kalitsios Georgios

Pantelidou Konstantina

Tsechelidou Konstantina

Final Project

Advanced Topics in Machine Learning

Professor : Tsoumakas Grigorios

Master on Artificial Intelligence

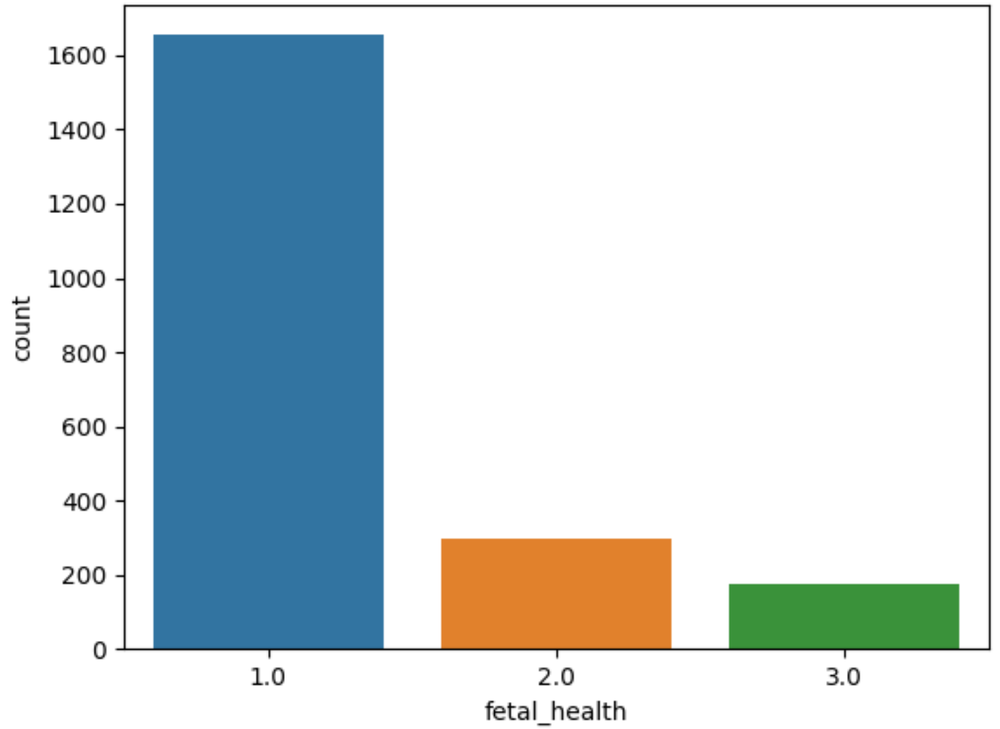
June 2021

Main Topics

1. Class Imbalance
2. Cost Sensitive Learning
3. Machine Learning Explainability

Class Imbalance

- Fetal Health Dataset taken from Kaggle
- 2126 records of features extracted from CTG exams
- 3 classes:
  - Normal (1655)
  - Suspect (295)
  - Pathological (176)



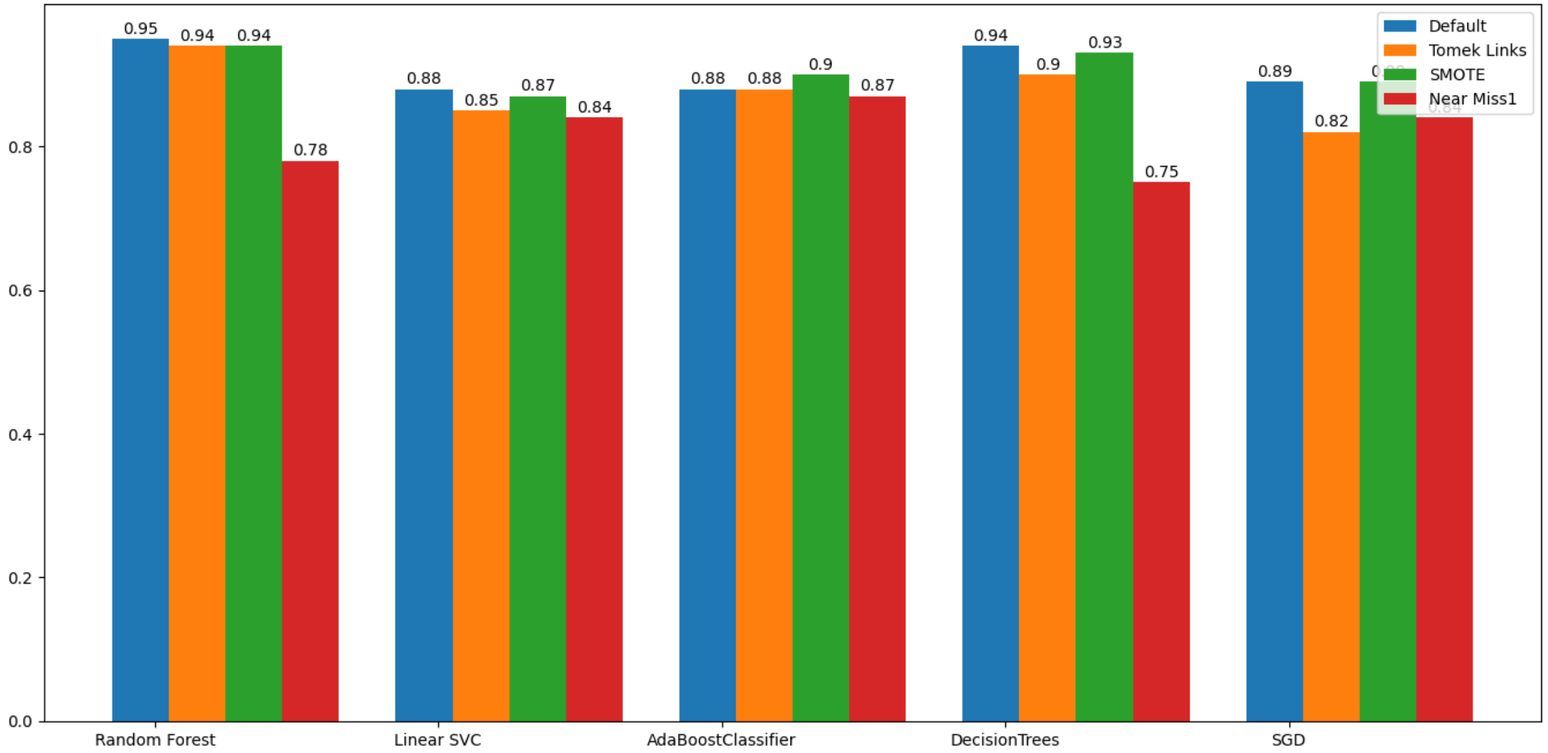
Class Imbalance

- Three main techniques
  - SMOTE
  - Tomek Links
  - NearMiss

- Five algorithms
  - DecisionTree
  - Random Forest
  - LinearSVC
  - AdaBoostClassifier
  - Stochastic Gradient Descent

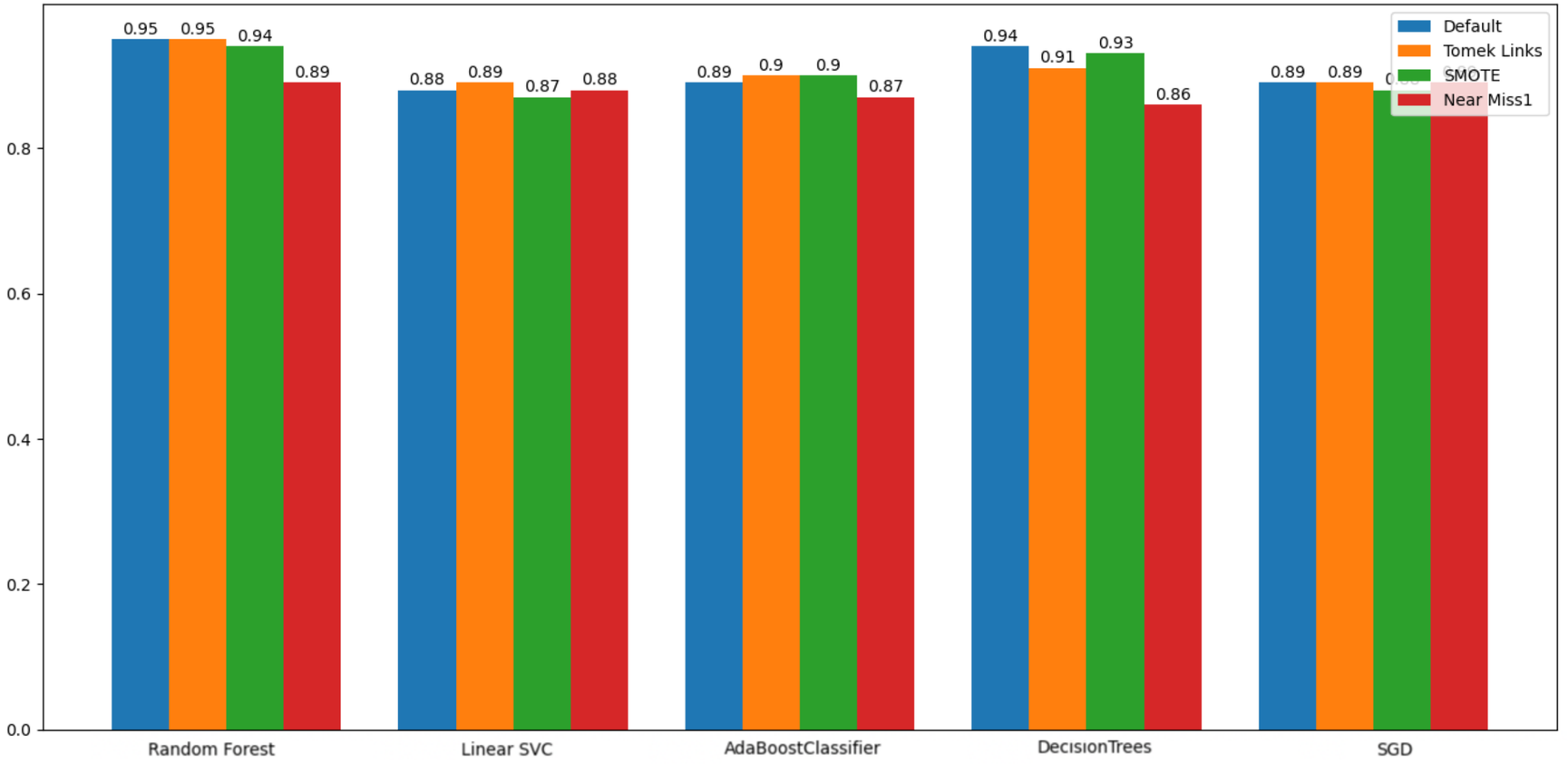
Class Imbalance

- F1 Score



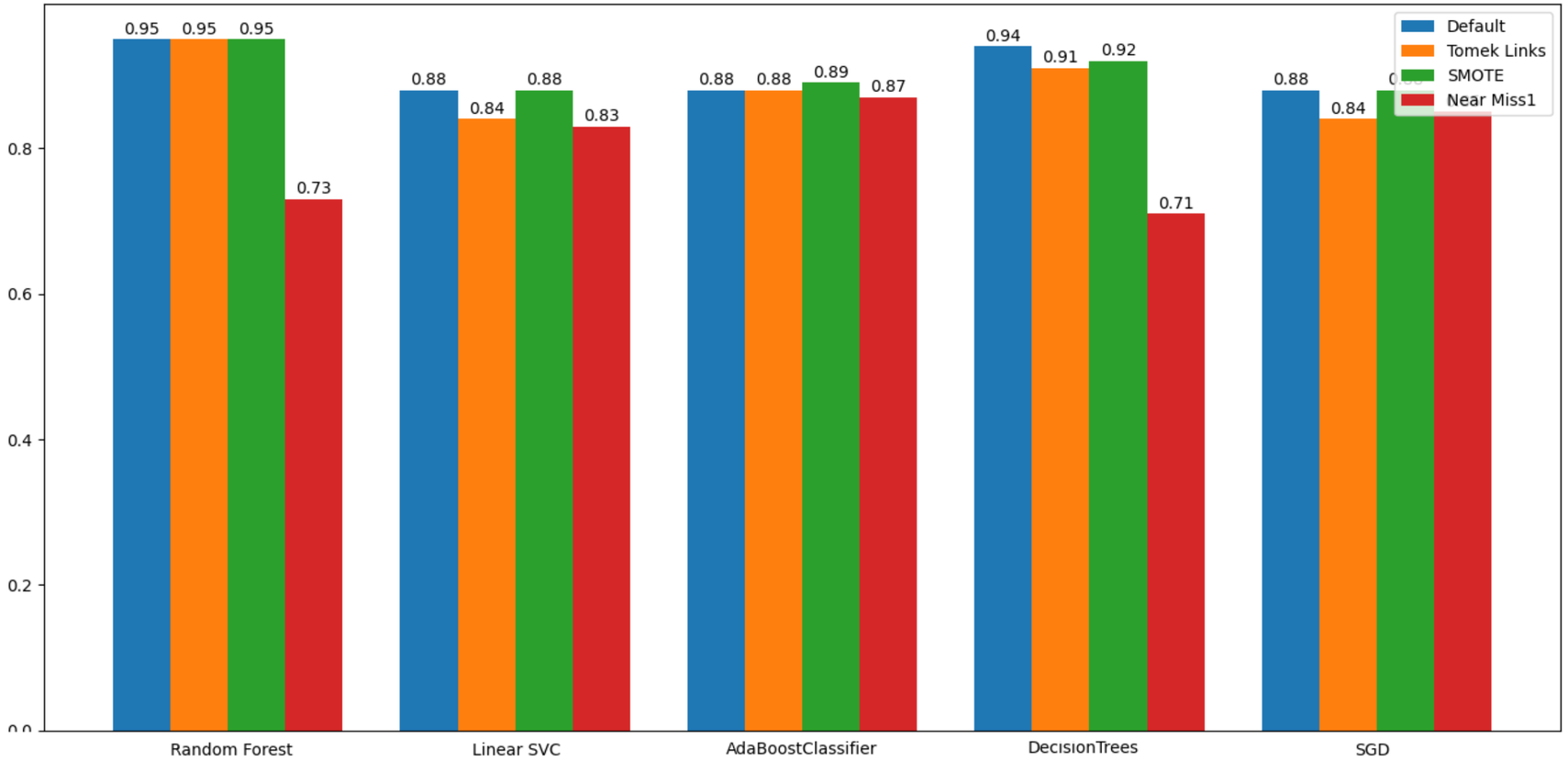
Class Imbalance

- Precision Score



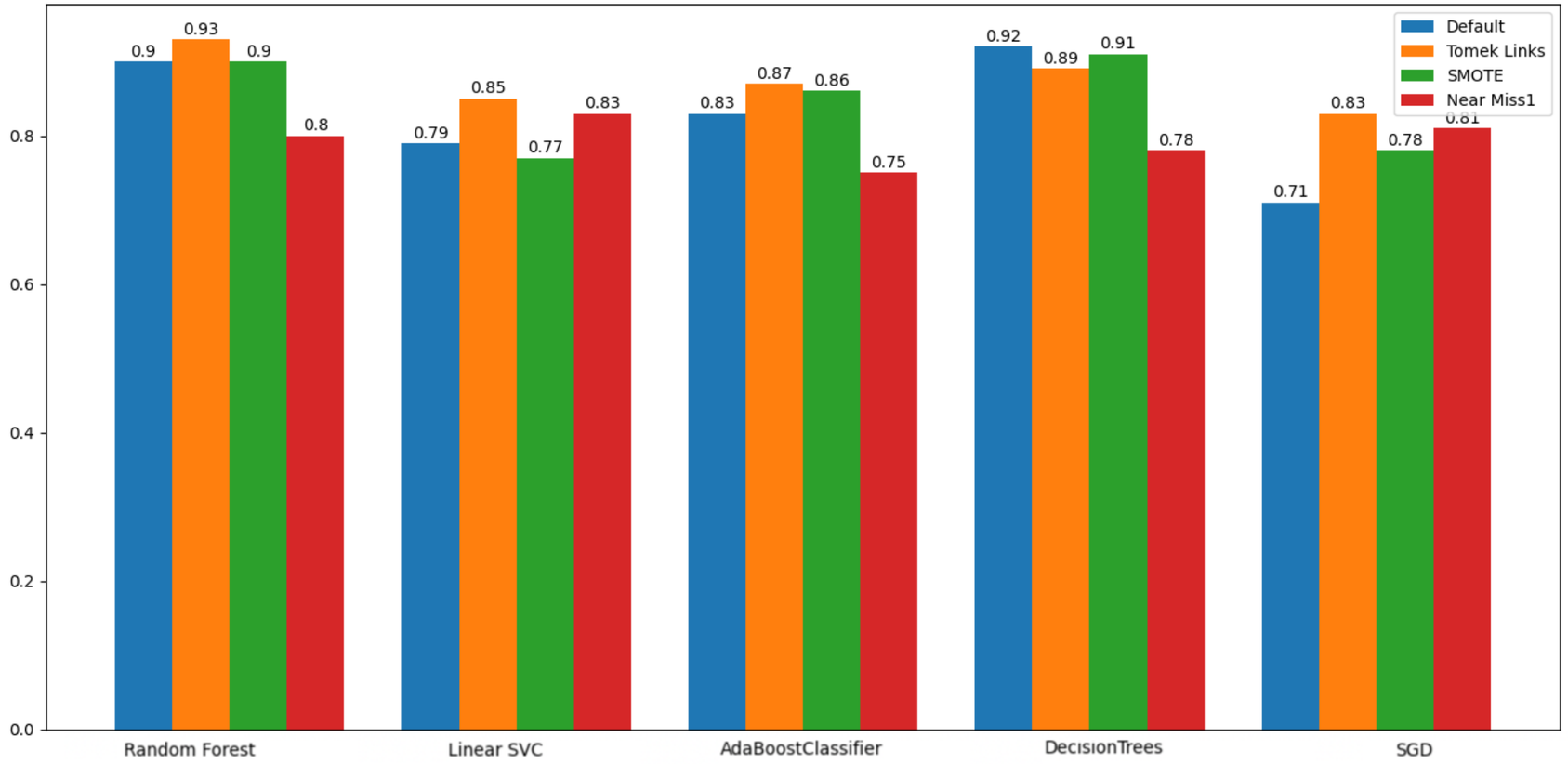
Class Imbalance

- Recall Score



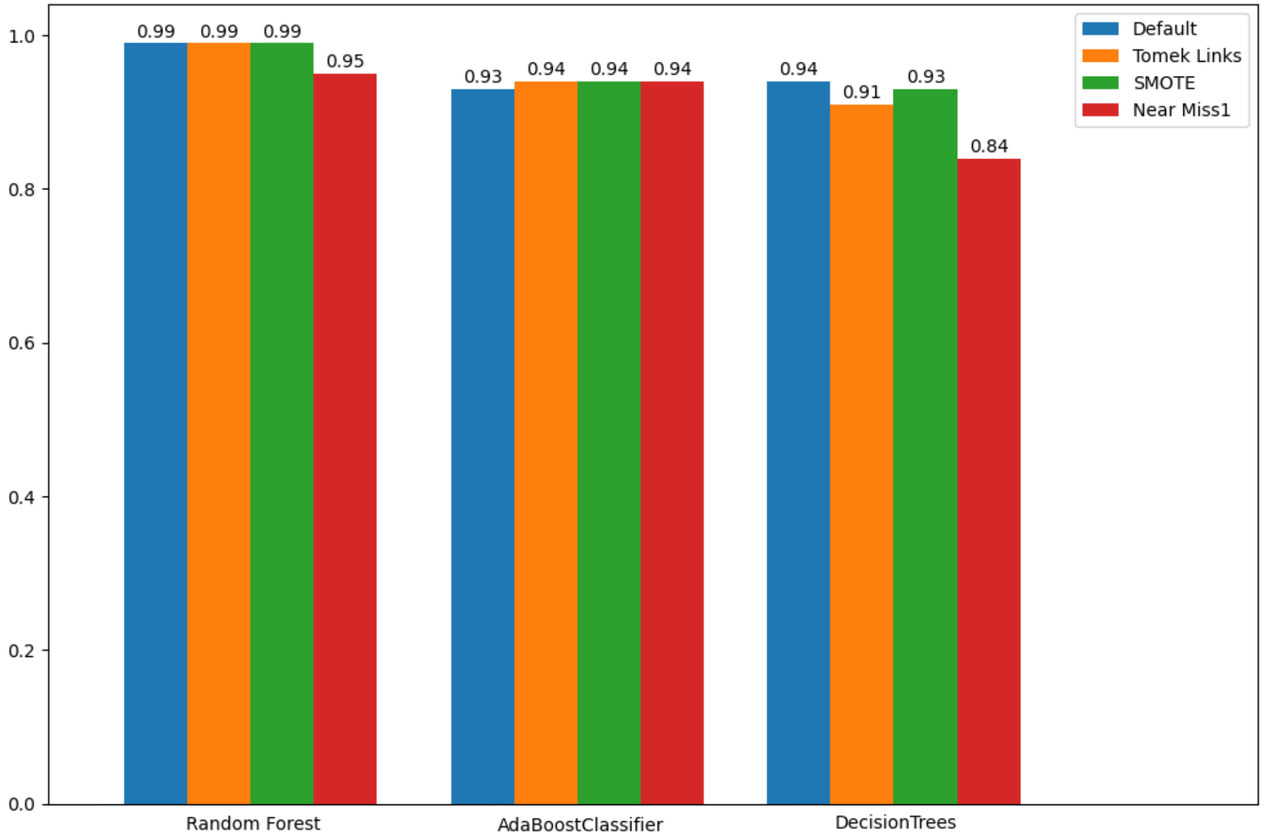
Class Imbalance

- Balanced Accuracy Score



Class Imbalance

- ROC AUC Score



Cost-Sensitive Learning

1. Three-class Classification Problem: multiclass cost-matrix

		Ground Truth		
		Normal	**Suspect**	**Pathological**
Prediction	**Normal**	0	4	5
	Suspect	1	0	1
	Pathological	1	1	0

2. Binary Classification Problem

		Ground Truth	
		Normal	**Pathological**
Prediction	**Normal**	0	5
	Pathological	1	0

Cost-Sensitive Learning

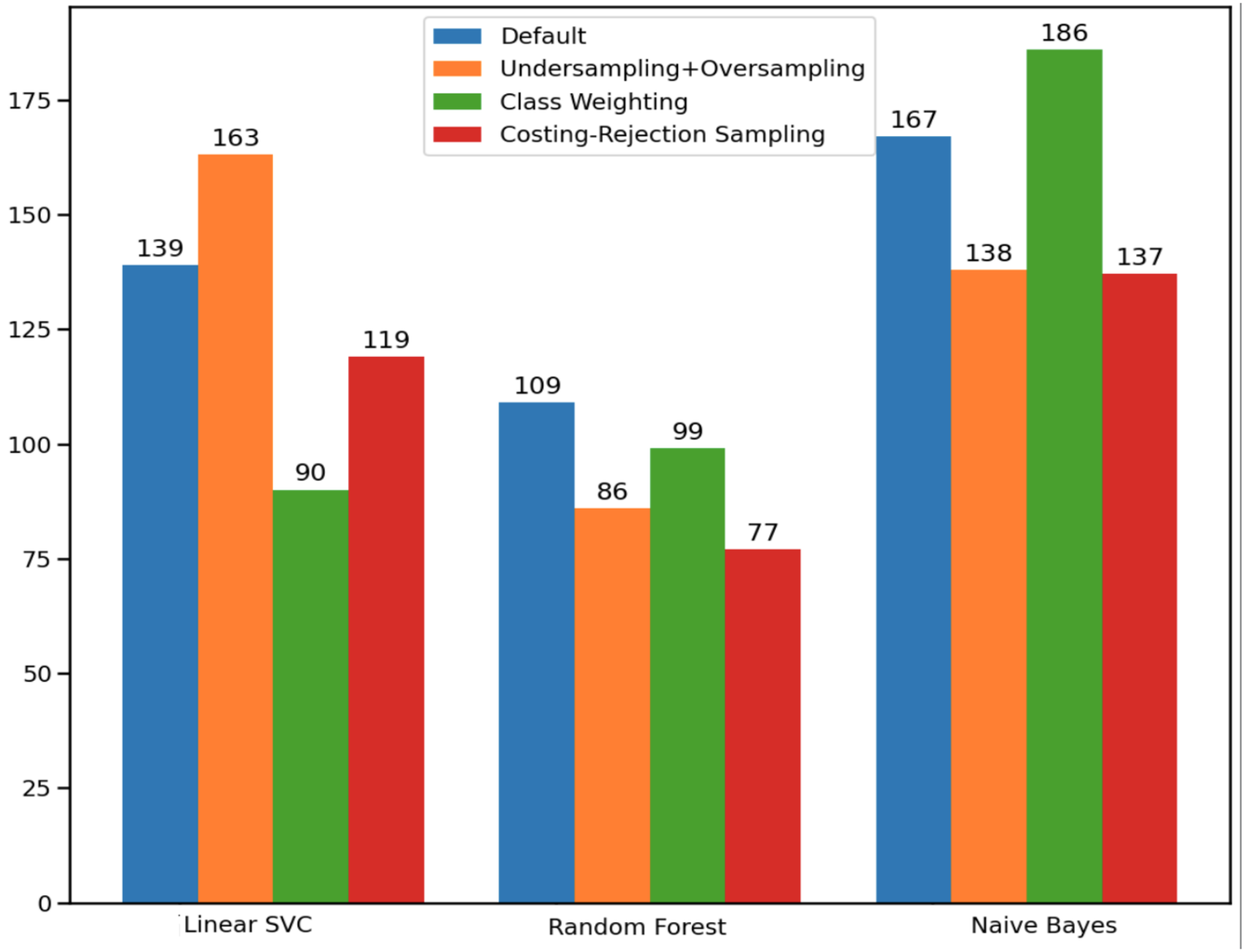
Training Set: Normal: 1241, Suspect: 221, Pathological: 132

1. Rebalancing: rebalance the classes according to their misclassification costs, for class j the sum of all $C(i,j)$
   a) **Under-Sampling** : Random Under-Sampler, [200, 221, 132]
   b) **Over-Sampling :** Random Over-Sampler, [1241, 1000, 1200]
   c) **Combination :** [200, 1000, 1200]

2. Class Weighting : weigh each example according to its misclassification cost
   class_weights : {1:2, 2:5, 3:6}

3. **Costing** (Ensemble Method) : Multiple runs (20) of **Rejection Sampling** combining with **Hard Voting**
   a) z = 5
   b) c = [2, 5, 6]

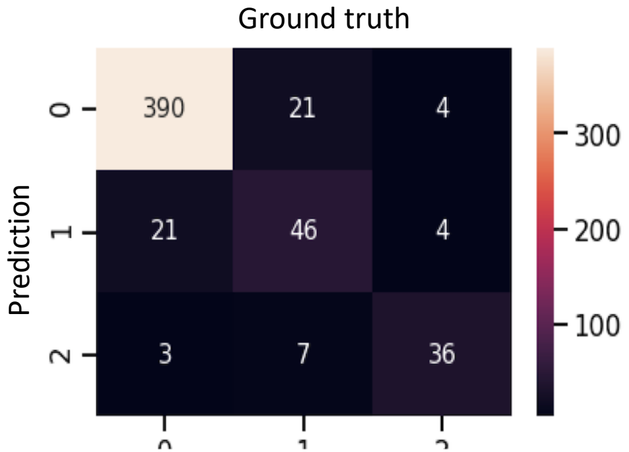
Cost-Sensitive Learning



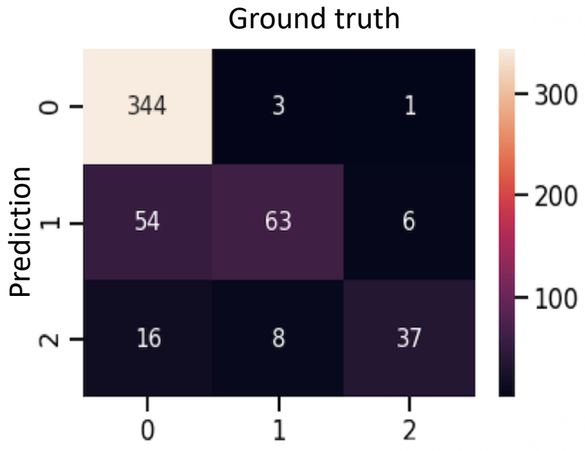
Total Cost of different training models

Cost-Sensitive Learning

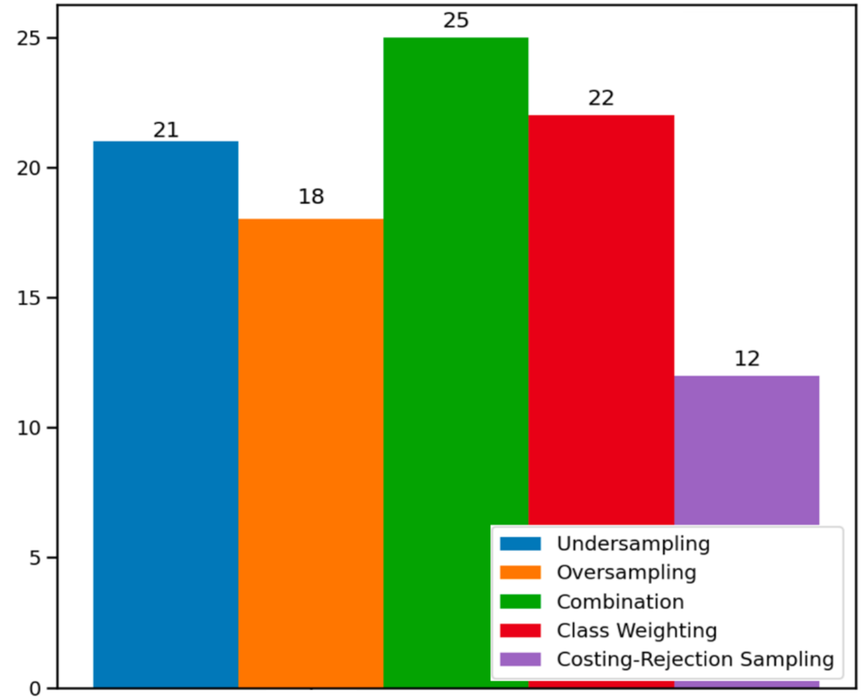
Only after applying the Cost-Sensitive techniques, many examples of costly classes (pathological and suspect) are not misclassified as normal and this is what we actually tried to achieve.



Default – Linear SVC



Over-Sampling – Linear SVC



How many examples of costly classes are classified correctly, after applying Linear SVC with Cost-Sensitive techniques

Machine Learning Explainability

Extract human-understandable insights from any model.

1. Permutation Importance

- What features does a model think are important ?
- Which features might have a greater impact on the model predictions than the others ?

2. Partial Dependence Plots

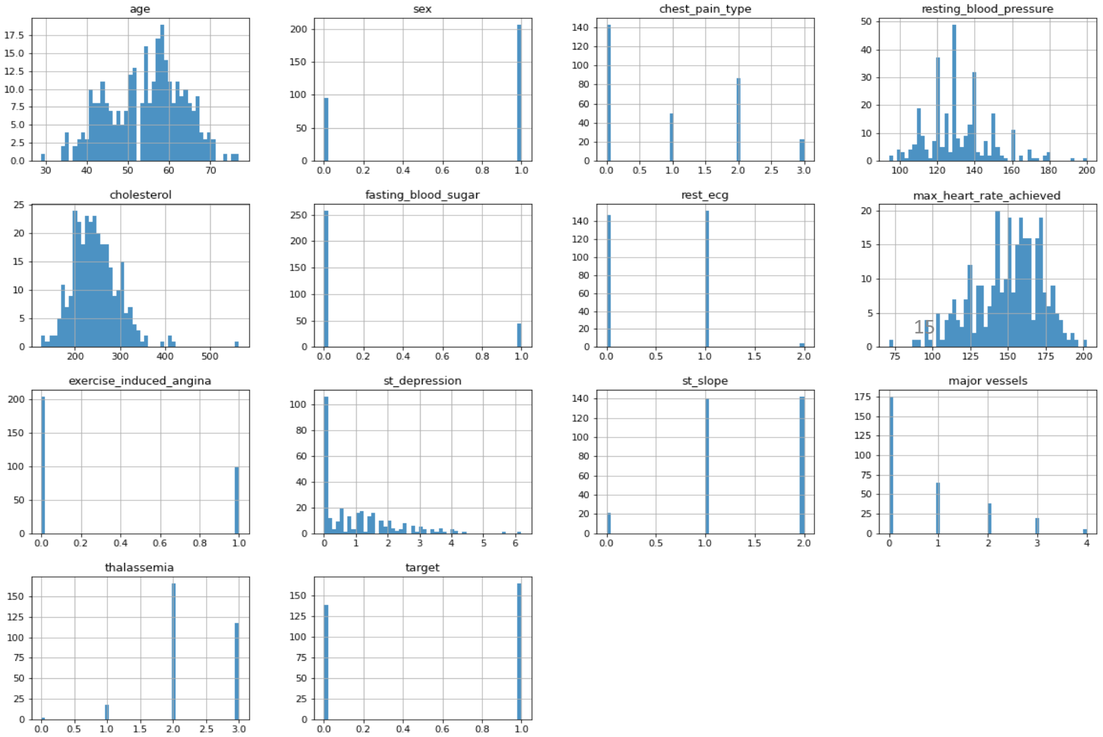
- How does each feature affect your predictions ?

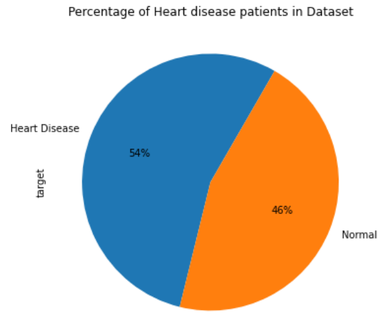
3. SHAP Values

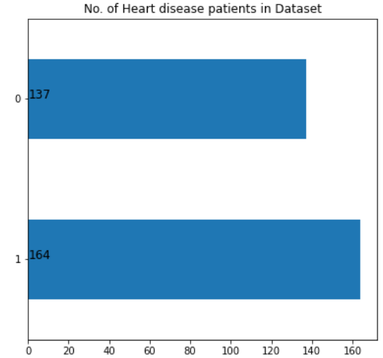
- Understanding individual predictions.

Machine Learning Explainability

- Heart Disease UCI Dataset
- This dataset consists of 13 features and a target variable.
- Dataset features







Machine Learning Explainability

1. Permutation Importance

Weight	Feature
0.0426 ± 0.0161	chest_pain_type_typical angina
0.0393 ± 0.0161	thalassemia_reversable defect
0.0262 ± 0.0334	rest_ecg_ST-T wave abnormality
0.0230 ± 0.0445	major vessels
0.0230 ± 0.0262	thalassemia_fixed defect
0.0164 ± 0.0207	chest_pain_type_atypical angina
0.0164 ± 0.0207	rest_ecg_normal
0.0033 ± 0.0482	st_depression
0.0033 ± 0.0131	exercise_induced_angina_no
0.0033 ± 0.0131	thalassemia_normal
0.0000 ± 0.0293	cholesterol
0.0000 ± 0.0359	sex_female
0.0000 ± 0.0207	age
0 ± 0.0000	rest_ecg_left ventricular hypertrophy
0 ± 0.0000	fasting_blood_sugar_greater than 120mg/ml
0 ± 0.0000	fasting_blood_sugar_lower than 120mg/ml
-0.0033 ± 0.0245	sex_male
-0.0033 ± 0.0131	st_slope_downsloping
-0.0033 ± 0.0131	st_slope_upsloping
-0.0066 ± 0.0161	chest_pain_type_non-anginal pain
-0.0066 ± 0.0161	chest_pain_type_asymptomatic
-0.0098 ± 0.0161	st_slope_flat
-0.0098 ± 0.0262	exercise_induced_angina_yes
-0.0230 ± 0.0334	max_heart_rate_achieved
	… 1 more …

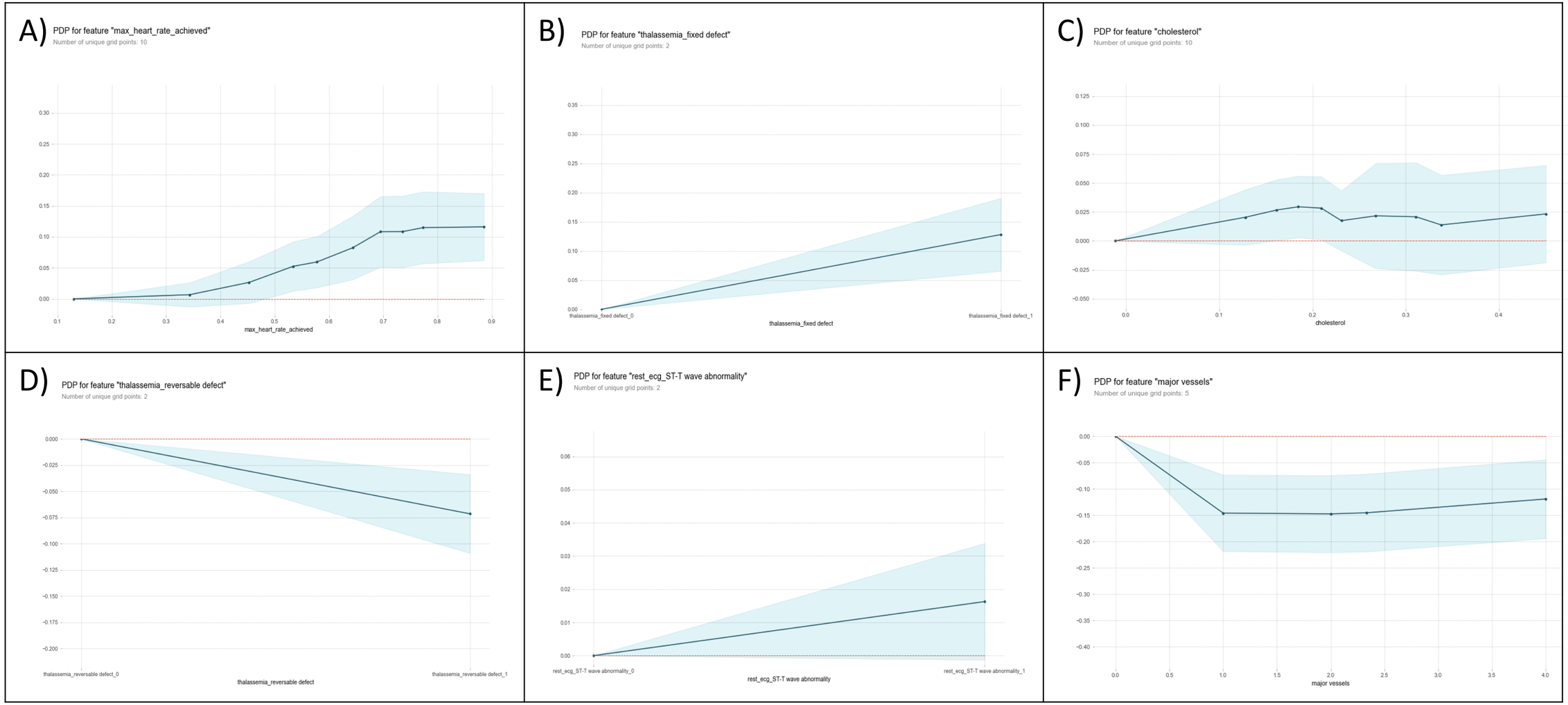
Permutation importance is calculated after a model has been fitted!

Here top 5 important features :

1. **chest_pain_type_typical angina**
2. **thalassemia_reversable defect**
3. **rest_ecg_ST-T wave abnormality**
4. **major vessels**
5. **thalassemia_fixed defect**

Machine Learning Explainability

2. Partial Dependence Plots (1D)



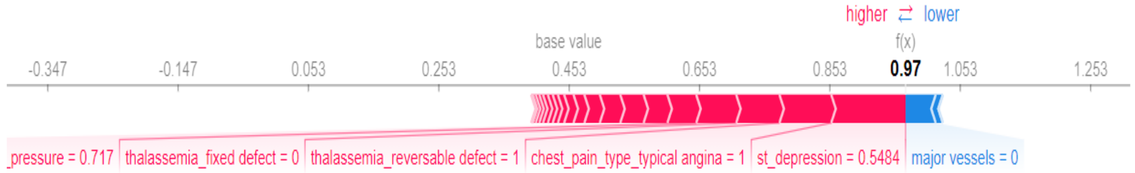
Machine Learning Explainability

2. Partial Dependence Plots (2D)

Machine Learning Explainability

3. SHAP Values

20th record of test set :

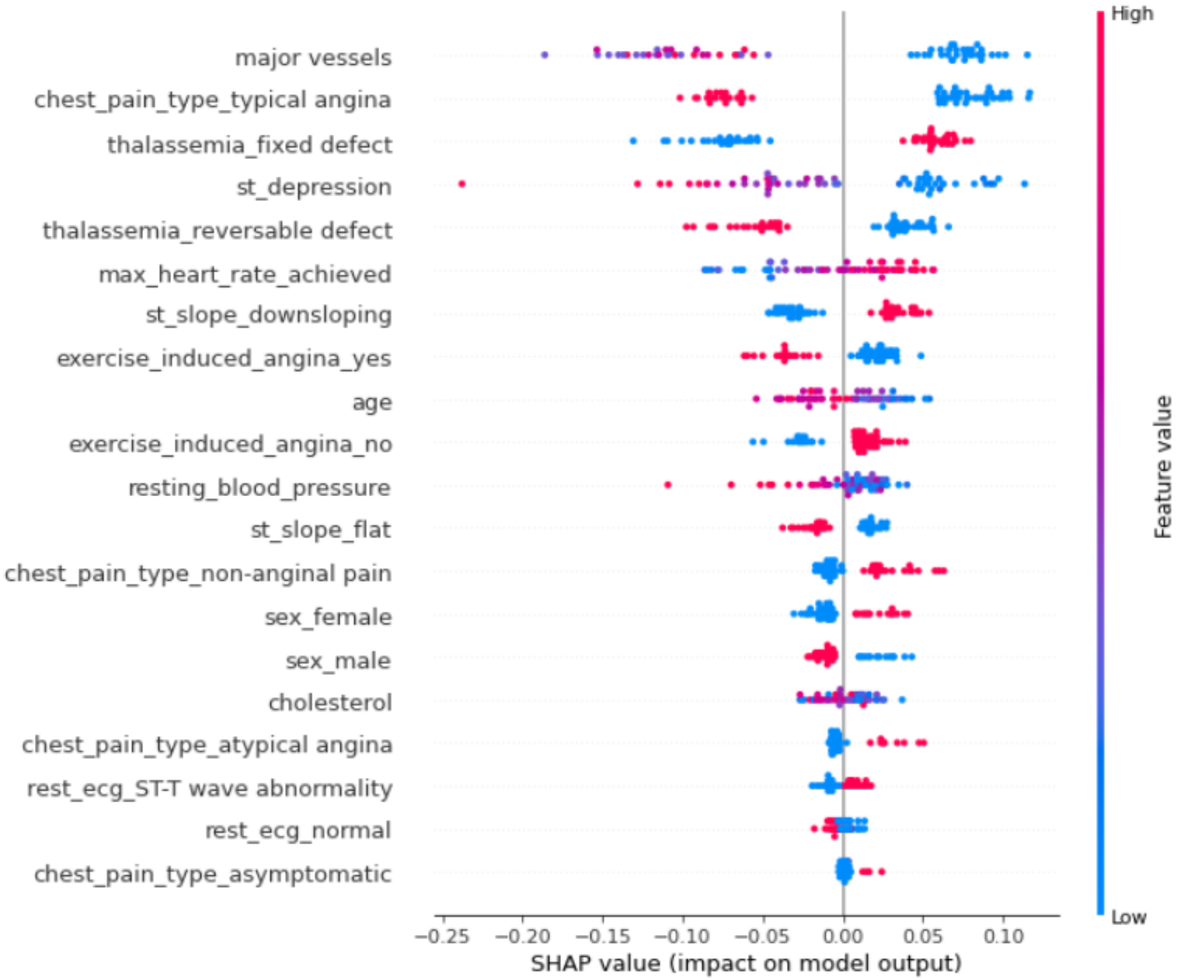


- Shap values show how much a given feature changed our prediction (compared to if we made that prediction at some baseline value of that feature).

- Features pushing the prediction higher are shown in red, those pushing the prediction lower are in blue.

- The base_value here is 0.453 while our predicted value is 0.97.
- st_depression=0.5484 has the biggest impact on increasing the prediction, while
- major vessels=0 the feature has the biggest effect in decreasing the prediction.

SHAP Summary SHAP Summary Plot Plot



Any questions?

Thank you for your time!