Predicting Heart Disease using Machine Learning

ORCA E 2500/4500 - Final Project

By:
Chayan Rellan (cr3194@columbia.edu)
Eric Palermo (ep2979@columbia.edu)
Fariha Imam (fi2183@columbia.edu)

Introduction

- Machine Learning - widely used in healthcare:
  - To forecast a rare disease
  - To detect an unusual disease

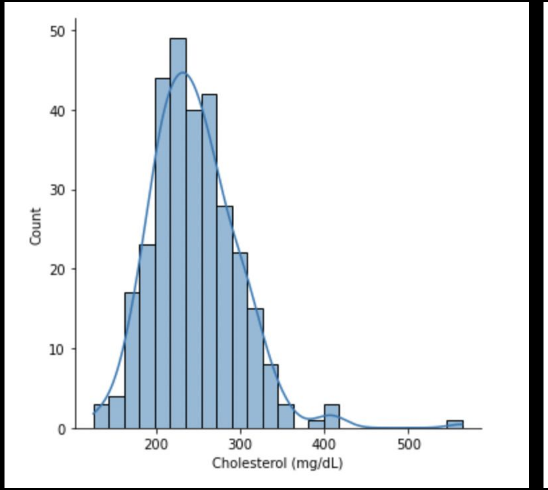
- 17.5 million deaths due to heart disease/strokes every year.

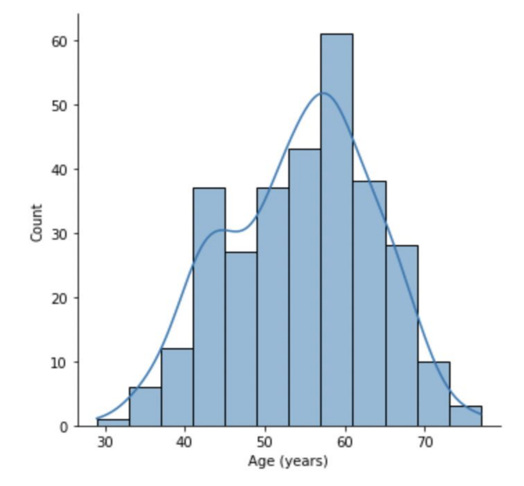
- Aim : To predict if the patient has a heart disease or not

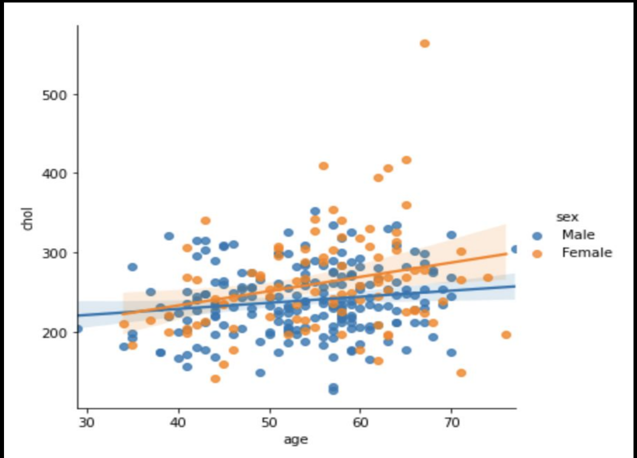
- Dataset used - Heart Disease UCI (303,14)

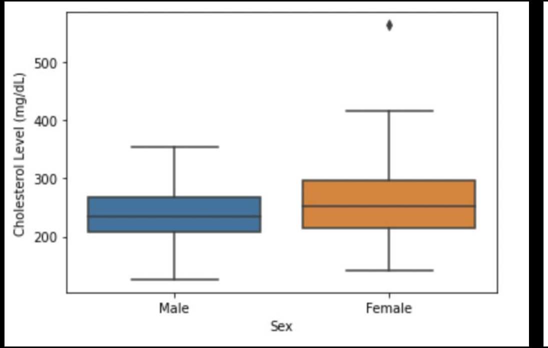
- Some common features - age, sex, resting blood pressure, cholesterol level, fasting blood sugar levels, maximum heart rate

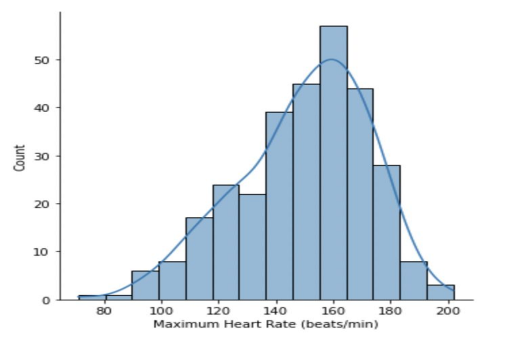
Data Visualization

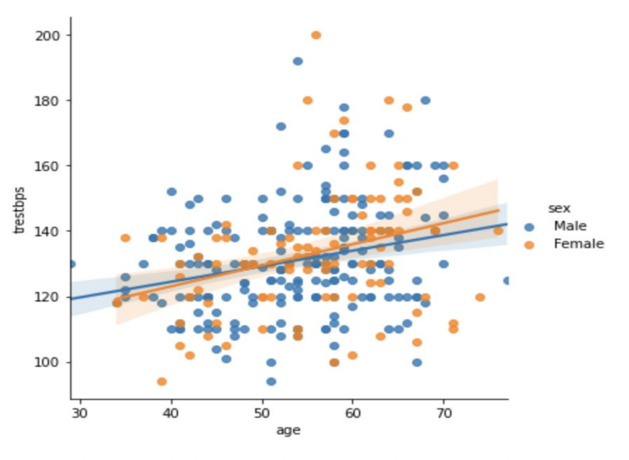






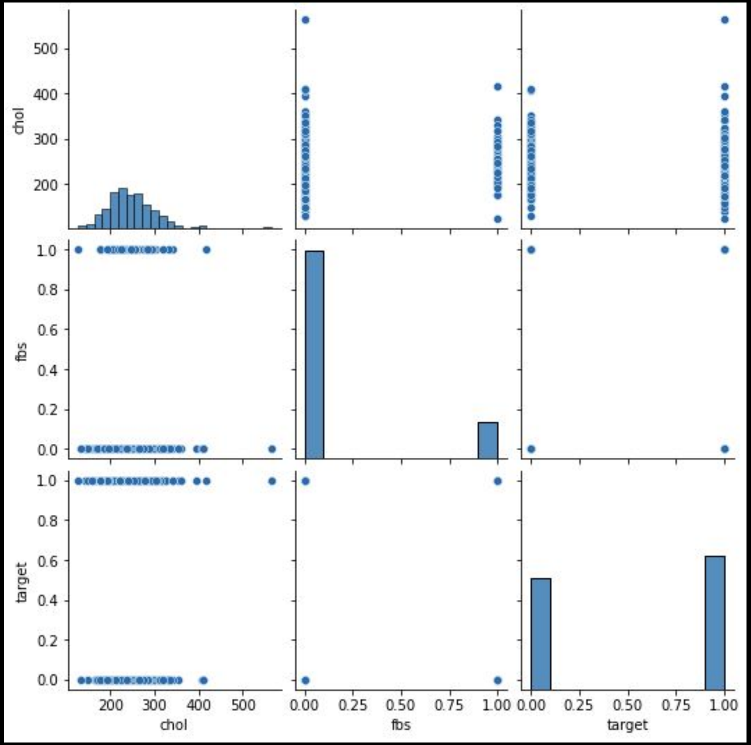






Data Preprocessing and Feature Engineering

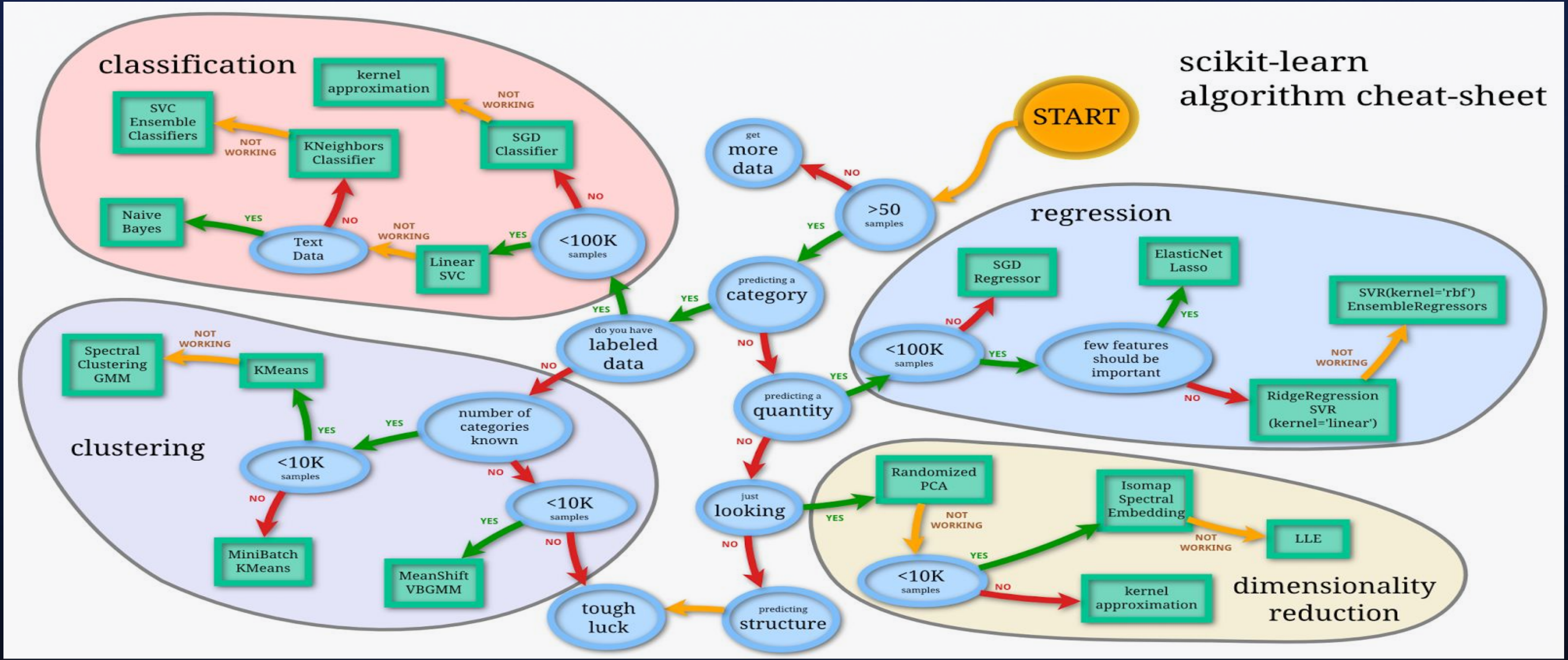
- Correlation between predictor variables and target variable measured with Pearson's correlation coefficient *r*
- No significant correlation between predictor variables
- *chol* and *fbs* had little correlation with the target variable (*r* = -0.09, *r* = -0.02)



Data Preprocessing and Feature Engineering

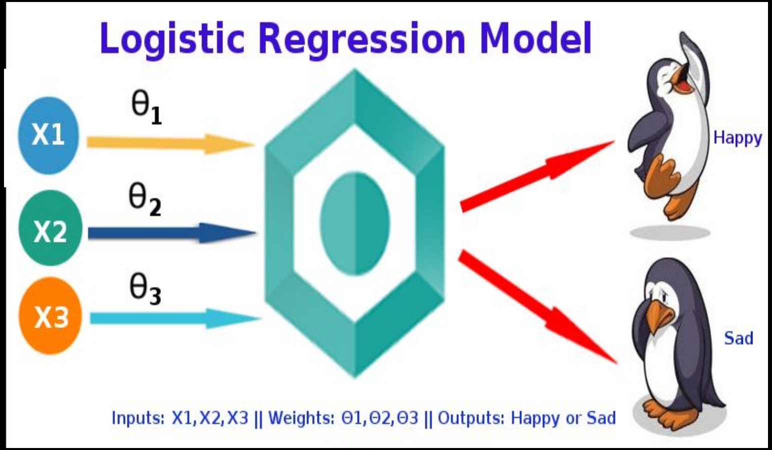
- Data set split into test and training sets
- Predictor variables subdivided into categorical and continuous:
  - Categorical: *sex*, *thal*, *exang*
  - Continuous: *age*, *ca*, *cp*, *oldpeak*, *restecg*, *slope*, *thalach*, *trestbps*
- Pipeline for categorical variables:
  - OneHotEncoder
  - PolynomialFeatures
- Pipeline for continuous variables:
  - PolynomialFeatures
  - StandardScaler
- ColumnTransformer to combine both pipelines

Proposed Models



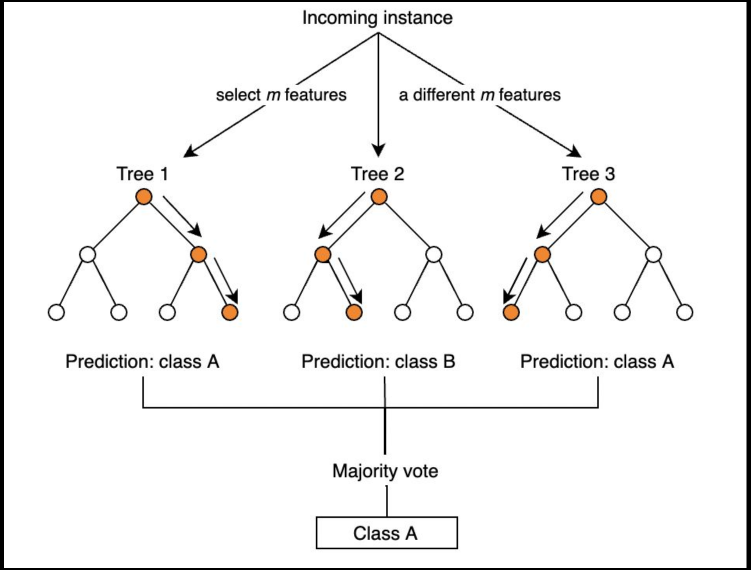
Logistic Regression

- Binary classification model
- Explains relationship between one dependent binary variable and multiple independent variables
- Cross Validation performed using
  - GridSearchCV
  - 5 fold split
- Hyper-parameters optimized:
  - Column transformer polynomial degree


Logistic Regression Model
X1
X2
X3
θ1
θ2
θ3
Happy
Sad
Inputs: X1,X2,X3 || Weights: Θ1,Θ2,Θ3 || Outputs: Happy or Sad

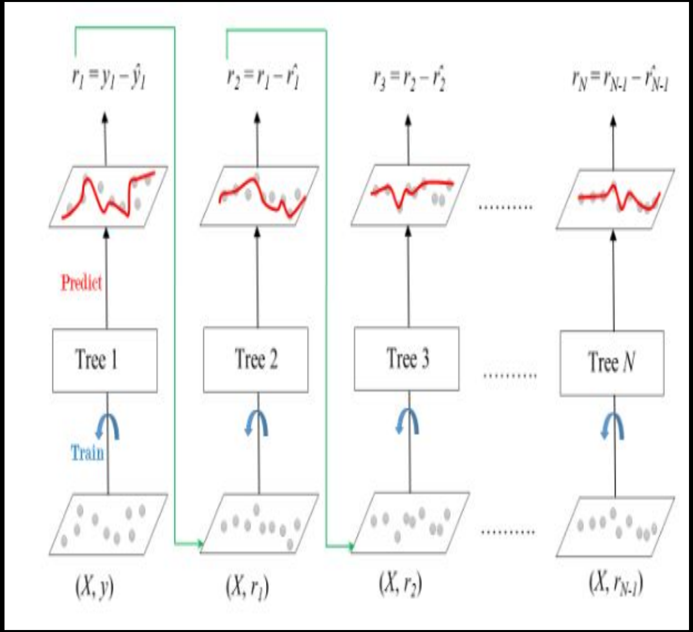
Random Forest Classifier

- Builds several decision trees and averages together their results
- Allows for feature selection
- Cross Validation performed using
  - GridSearchCV
  - 5 fold split
- Hyper-parameters optimized:
  - Column transformer polynomial degree
  - Max depth of tree
  - Minimum sample split



Gradient Boosting Classifier

- Improves model predictions by training the predictor using the errors of the previous model
- Each successive model tries to correct the deficiencies of the models before it
- Cross Validation performed using
  - GridSearchCV
  - 5 fold split
- Hyper-parameters optimized:
  - Column transformer polynomial degree
  - Minimum sample split
  - Learning rate
  - Number of estimators



Results and Conclusions

	Logistic Regression	Random Forest Classifier	Gradient Boosting Classifier
Brier Score	0.0991	0.1081	0.1231
Hyperparameters	● C = 0.1 ● Polynomial Degree = 2	● Max Depth = 4 ● Min Samples Split = 4 ● Min Samples Leaf = 5 ● Polynomial Degree = 2	● Learning Rate = 0.05 ● Min Samples Split = 8 ● # of Estimators = 50 ● Polynomial Degree = 1

Increasing model complexity does not necessarily guarantee better accuracy or better prediction.

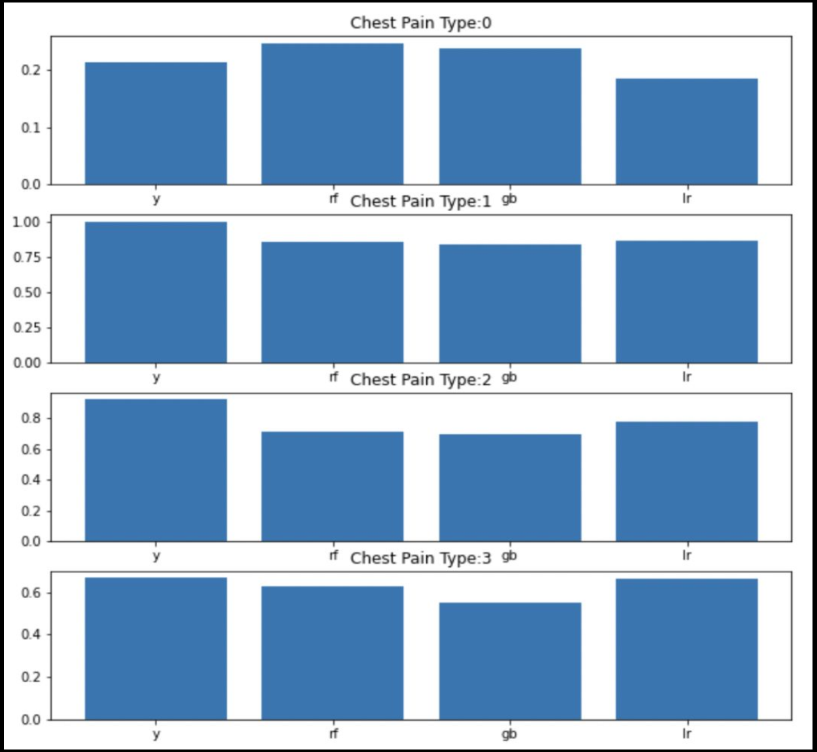
Results and Conclusions

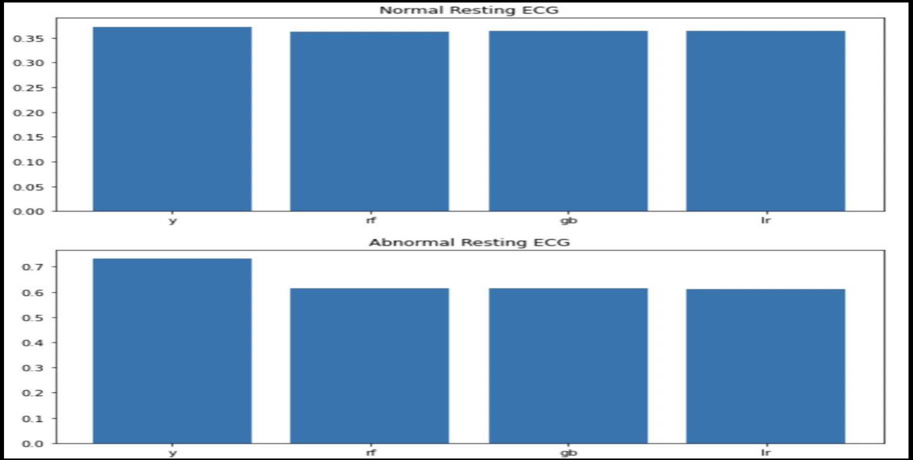
Classification Report :

Class 0 (Have a Heart Disease)	**Precision**	**Recall**	**F1 - Score**
Logistic Regression	0.90	**0.90**	**0.90**
Random Forest Classifier	0.89	0.86	0.88
Gradient Boosting Classifier	0.78	0.86	0.82

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad F1 - Score = \frac{2 \cdot (Precision) \cdot (Recall)}{(Precision + Recall)}$$

Results and Conclusions





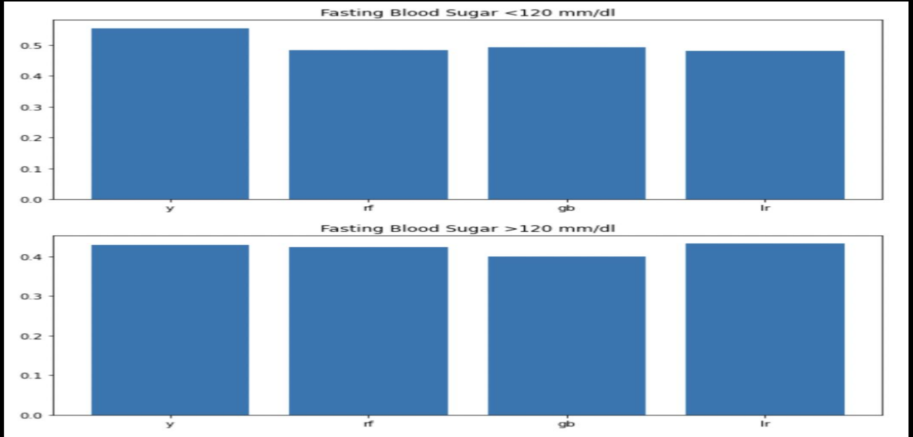


Figure 3: Fasting Blood Sugar

Thank You !

PHOTO: DARCY PETERKA