Prediction of Heart Disease

- Mankirat Singh Bhamra MXB220061
- Harikrishna Dev HXD220000
- Krishnan Venkatesan KXV220007
- Medha Priyanga Saravanan MXS220057

Introduction

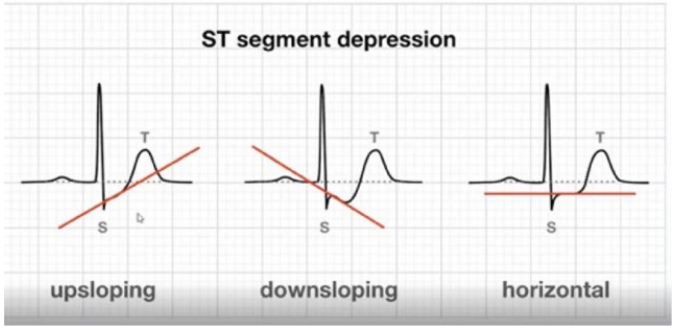
- According to WHO, Cardiovascular diseases are the leading cause of death globally.
- It is estimated that 17.9 million people (32%) deaths were caused by CVDs. Of these deaths, 85% were heart attacks and strokes.
- Detecting cardiovascular disease early is important to begin management with counseling and medication.
- Our objective is to identify high-risk patients for CVD based on their current health-related information obtained from yearly checkups.
- The prediction will assist medical professionals and self-assessment tools to evaluate patients' course of medication and health plan choices.

Data introduction

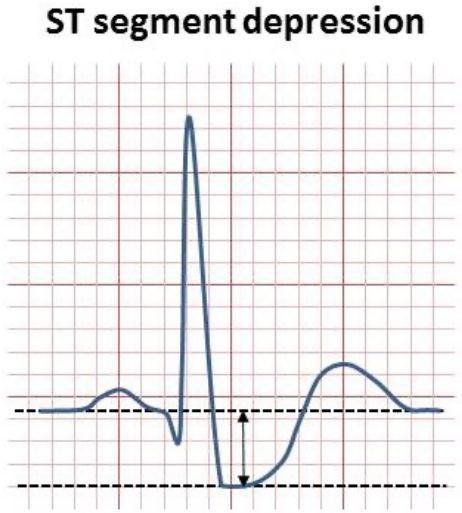
- UCI machine learning repository is the original source, and we are using the cleaned-up dataset from Kaggle.
- The data consists of a random set of patients from 30 May 1989 to 2 Dec 1996 and it has their medical information.
- The data can be accessed from the Kaggle link here.

Model variables

- **Age**: age of the patient [years]
- **Sex**: sex of the patient [M: Male, F: Female]
- **ChestPainType**: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- **RestingBP**: resting blood pressure [mm Hg]
- **Cholesterol**: serum cholesterol [mm/dl]
- **FastingBS**: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- **RestingECG**: relaxing electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- **MaxHR**: maximum heart rate achieved [Numeric value between 60 and 202]
- **ExerciseAngina**: exercise-induced angina [Y: Yes, N: No]
- **Oldpeak**: oldpeak = ST [Numeric value measured in depression]
- **ST_Slope**: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]



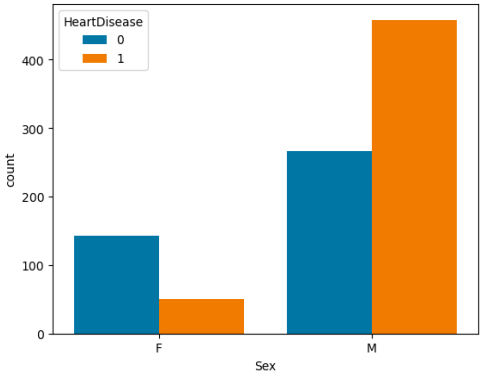
OldPeak

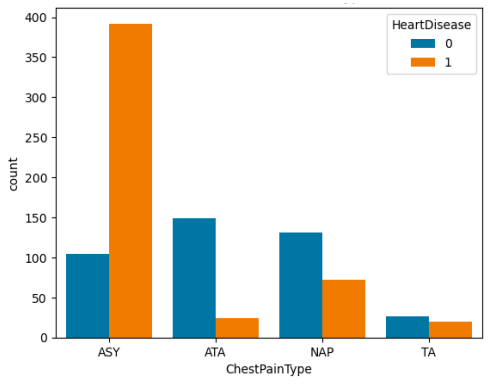


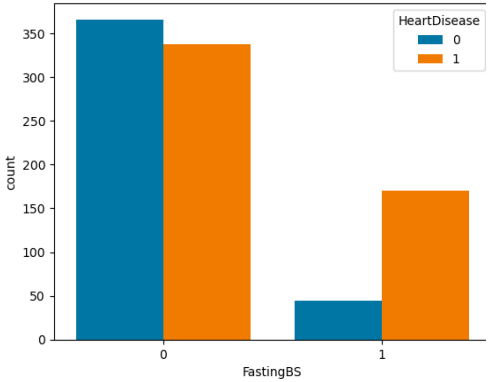
Exploratory Data analysis

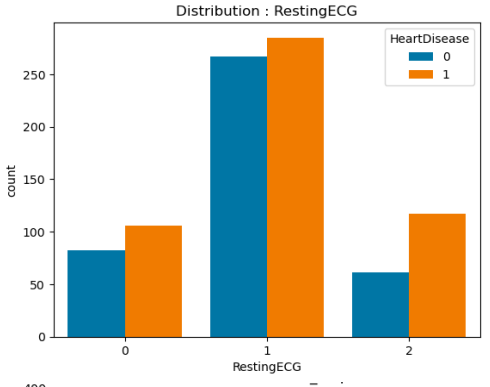
- **Sex**: Male patients > Female patients
- **ChestPainType**: ASY > NAP > ATA > TA
- **FastingBS**: (FBS < 120 mg/dl) > (FBS > 120 mg/dl)
- **RestingECG**: Normal > ST > LVH
- **ExerciseAngina**: Angina > No Angina
- **ST_Slope**: Flat > Up > Down

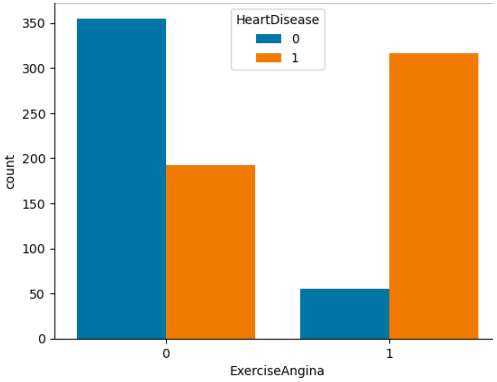
Variable	Chi-Squared Score
ExerciseAngina	139.775283
ChestPainType	127.478652
ST_Slope	67.425731
FastingBS	16.015879
Sex	15.600925
RestingECG	0.031521

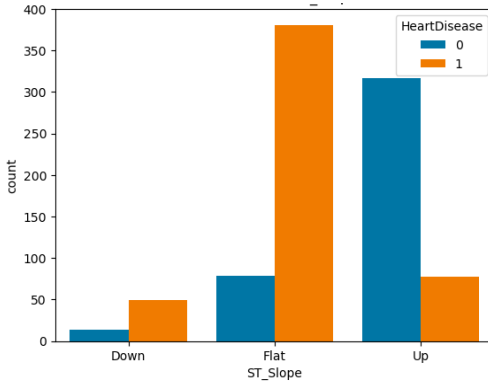




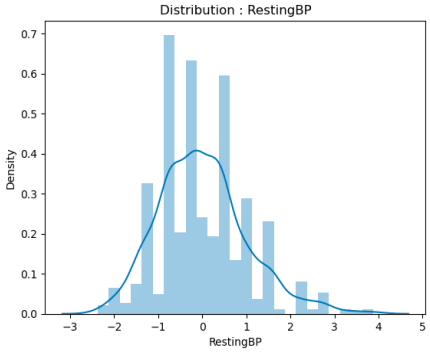


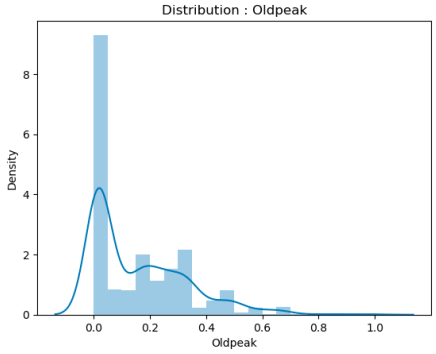


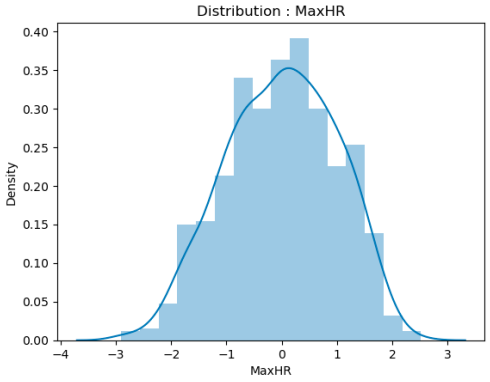


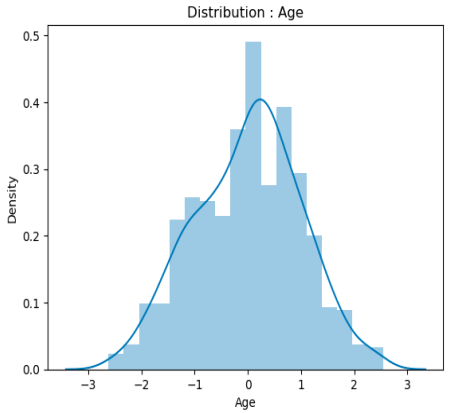


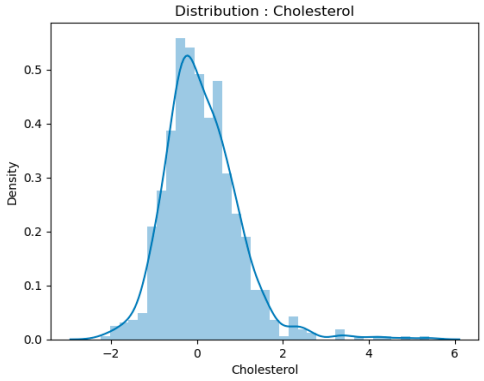
Feature Engineering

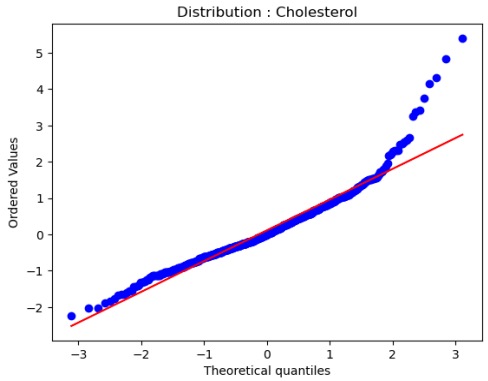












- **Age**: 50+
- **RestingBP**: 95 - 170
- **Cholesterol**: 160 - 340
- **MaxHR**: 70 - 180
- **Oldpeak** : 0 - 4

Variables	ANOVA Score
Oldpeak	242.364279
MaxHR	123.425078
Age	72.839116
RestingBP	23.020309
Cholesterol	8.113851

Based on the distribution plots, we used these scalers:

Variables	Scaler
Oldpeak	Minmax
Cholesterol	Robust
Other Variables	Standard

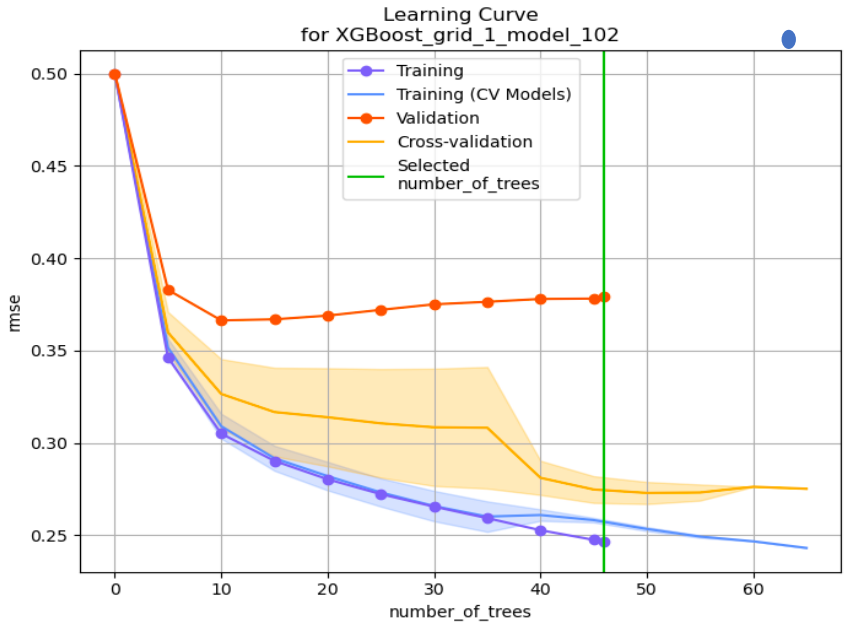
Modeling approach

- Utilized the h2o ML platform to run multiple models concurrently and hyperparameterize them simultaneously, and integrated AutoML functionality with the Python environment.
- Facilitated the selection of the best model and enhanced ease of use.
- Identified the XGBoost classifier as the best model.

Model used: XGBoost Classifier model

- After running multiple binomial classification models (such as Decision Trees, etc.), we were able to conclude that the XGBoost Classifier is the best model.

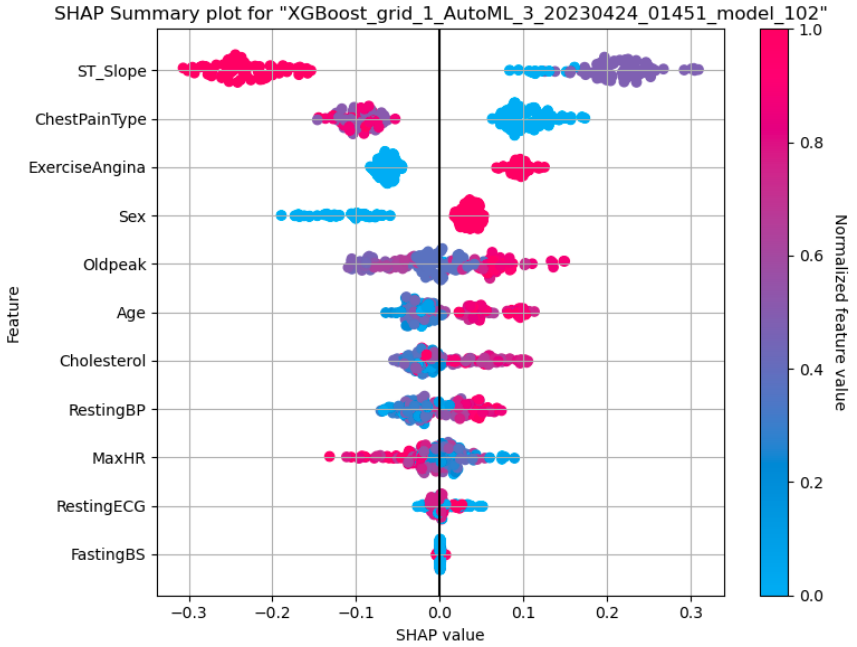
- The Model has 46 trees which were determined by 5-fold cross-validation.

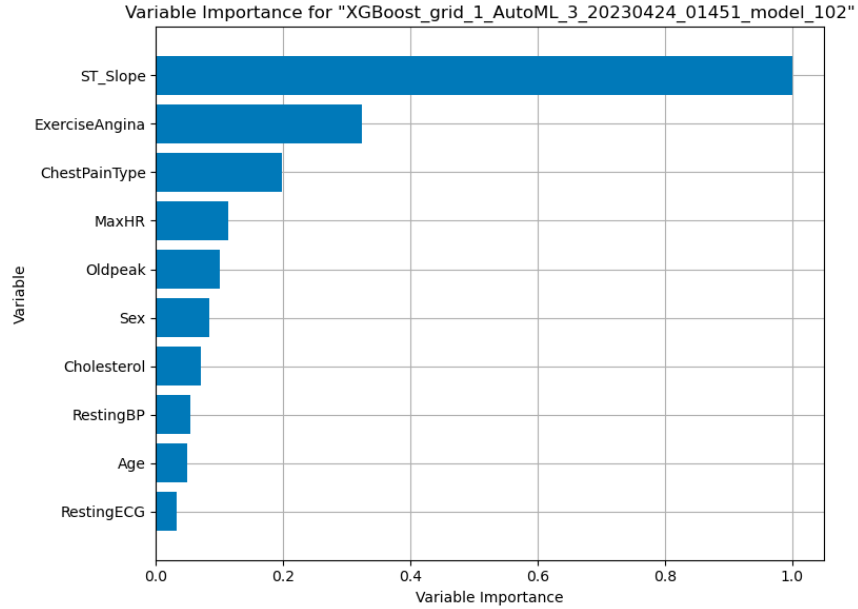


Model Results

- ST Slope is the most influential variable based on the variable importance graph.
- ST Slope has a wide range of SHAP values, indicating its significance in the model.

MSE: 0.1438
RMSE: 0.379
MAE: 0.2686
RMSLE: 0.2642
Mean Residual Deviance: 0.14380
Accuracy: 89.61%
Cross-validation Score: 92.20%





Future project implementation scope:

- Incorporating time component of features can monitor patient improvement and identify seasonal effects.
- Evaluation of treatment effectiveness and lifestyle changes.
- Applying AI to analyze large health datasets to identify correlations between health factors.
- Developing personalized treatment plans based on individual patient health data and medical history.
- Implementing telemedicine platforms that enable doctors to remotely monitor patient progress and communicate with patients in real-time, improving access to care and reducing costs.