Classification of Smoker Status using Bio-Signals

Johnroe Paulo Cañamaque

Data Science Student

November 24, 2023

Agenda

01. Data Presentation

Business Case

Data specifics and definition

02. Data Analysis

Methodology

Data Insights

03. Model Design

Model Training and Evaluation

Best Model

Business Case

Smoking is a well-known cause of a variety of health problems and is a major contributor to preventable diseases and deaths globally.

To improve the effectiveness of smoking cessation, let us use ML to predict better. Our aim is to create a model that can predict an individual's smoking status using bio-signals.

Data Presentation

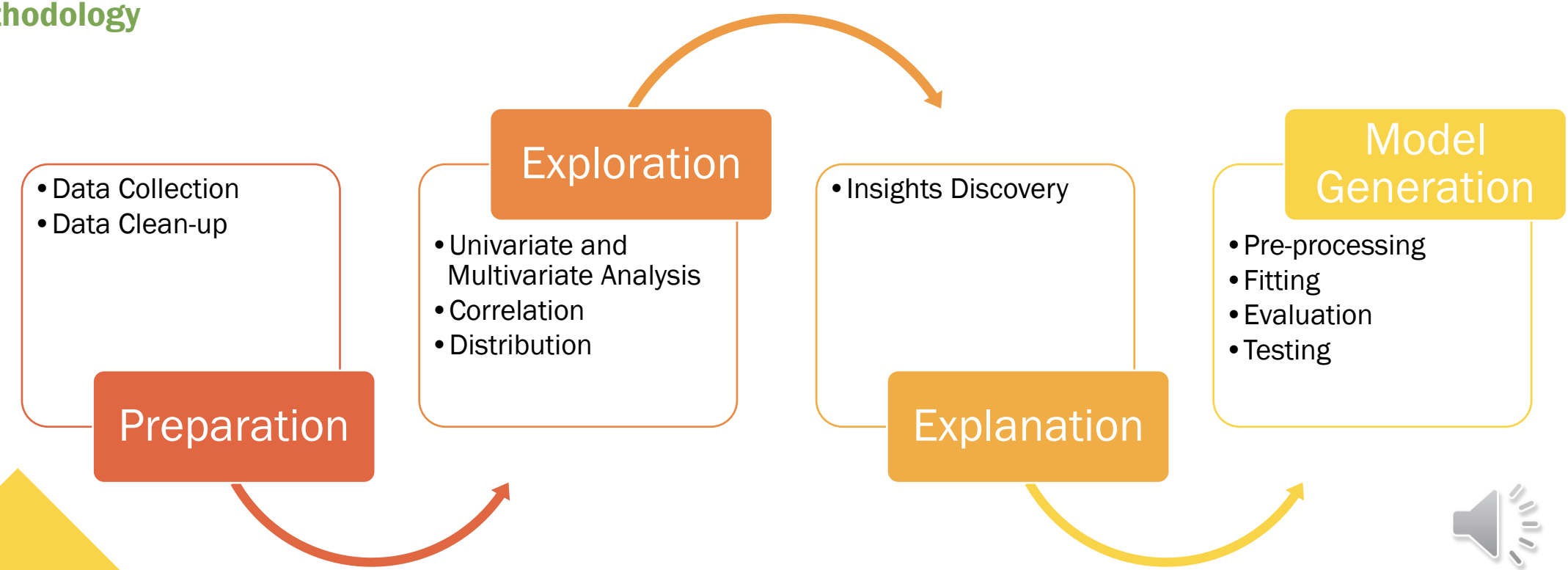
Data Description

- This data contains a sample of bio-signals gathered from individuals who are either smokers or non-smokers.
- Data used has 38,984 rows and 23 columns
- Accessible through here

Attribute	Description
age	Age of the individual
height(cm)	Height in centimeters
weight(kg)	Weight in kilograms
waist(cm)	Waist circumference in centimeters
eyesight(left)	Left eye eyesight measurement
eyesight(right)	Right eye eyesight measurement
hearing(left)	Left ear hearing assessment
hearing(right)	Right ear hearing assessment
systolic	Systolic blood pressure
relaxation	Diastolic blood pressure (relaxation)
fasting blood sugar	Fasting blood sugar level
Cholesterol	Total cholesterol level
triglyceride	Triglyceride level in the blood
HDL	HDL cholesterol level
LDL	LDL cholesterol level
hemoglobin	Hemoglobin level in the blood
Urine protein	Presence of urine protein
serum creatinine	Serum creatinine level
AST	AST (glutamic oxaloacetic transaminase) level
ALT	ALT (glutamic pyruvic transaminase) level
Gtp	γ-GTP level
dental caries	Presence of dental caries
smoking	Smoking status

Data Analysis

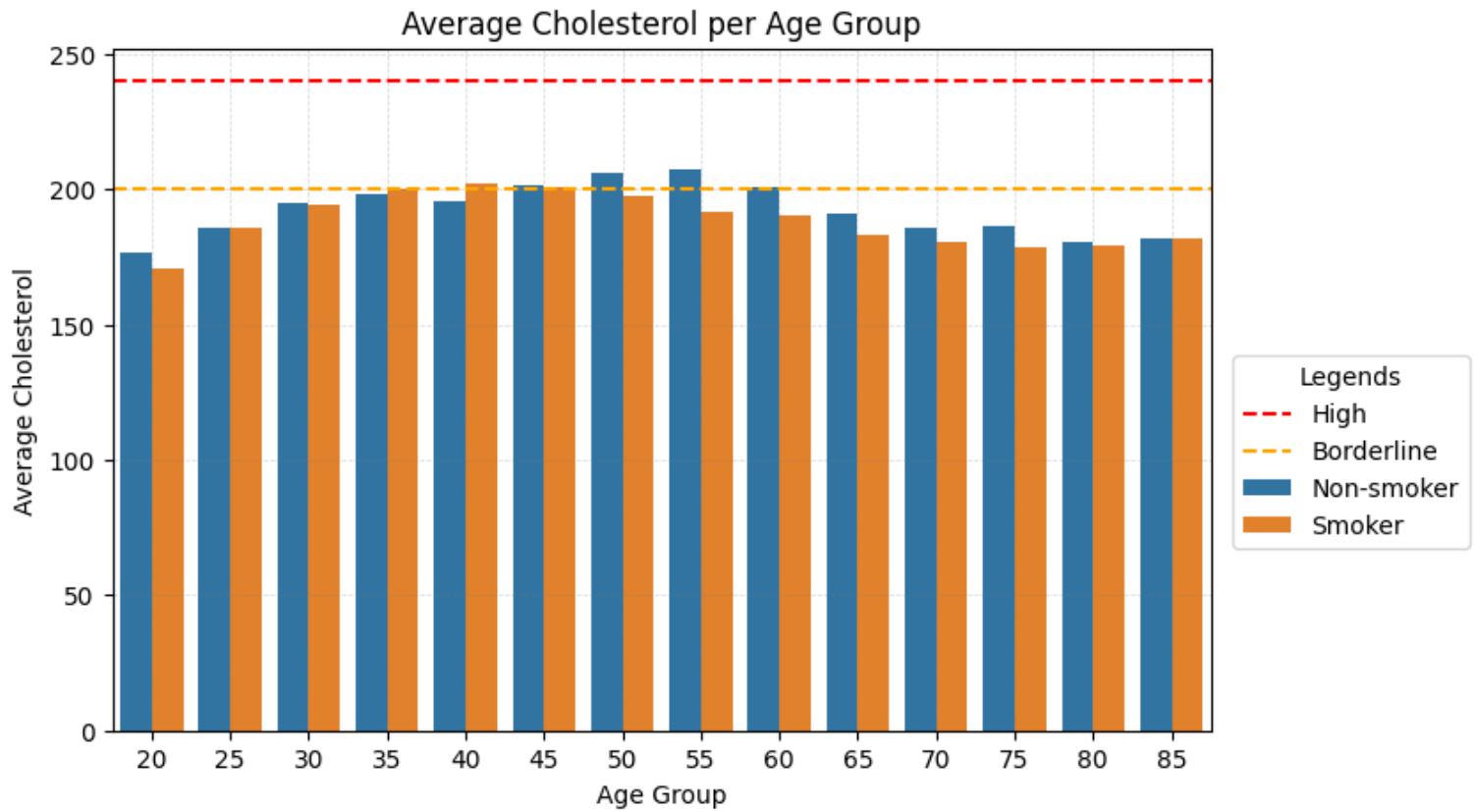
Methodology



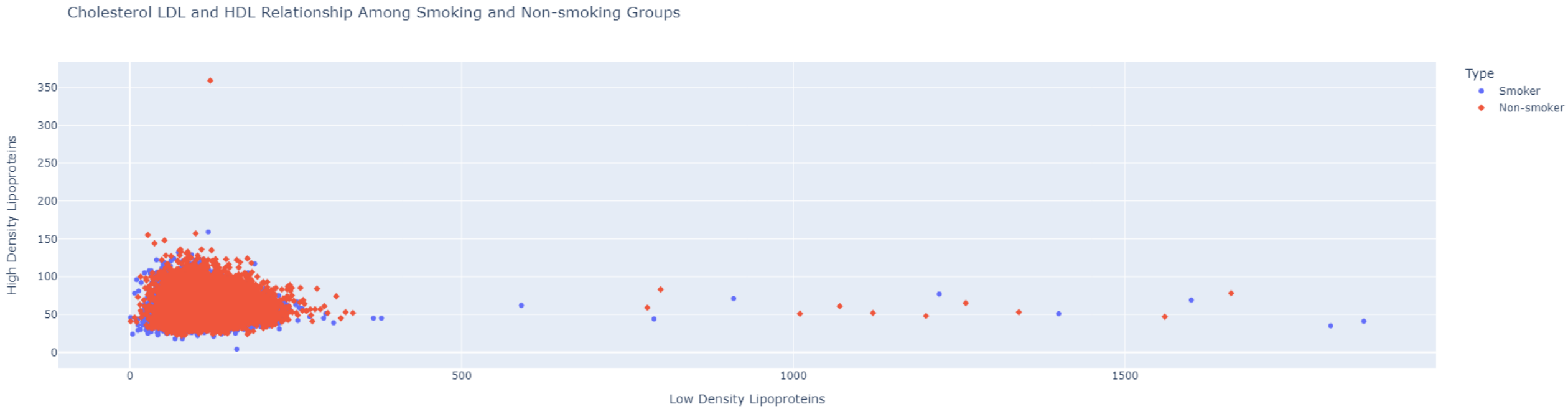
Data Analysis

Insights Discovered

- Age groups from 35 to 60 are having an average of 200 cholesterol level.
- Majority of the age groups have higher cholesterol levels from non-smokers than the smokers
- There is no indicative correlation of cholesterol to smoking habits of a person based on this insight



Data Analysis



Model Design

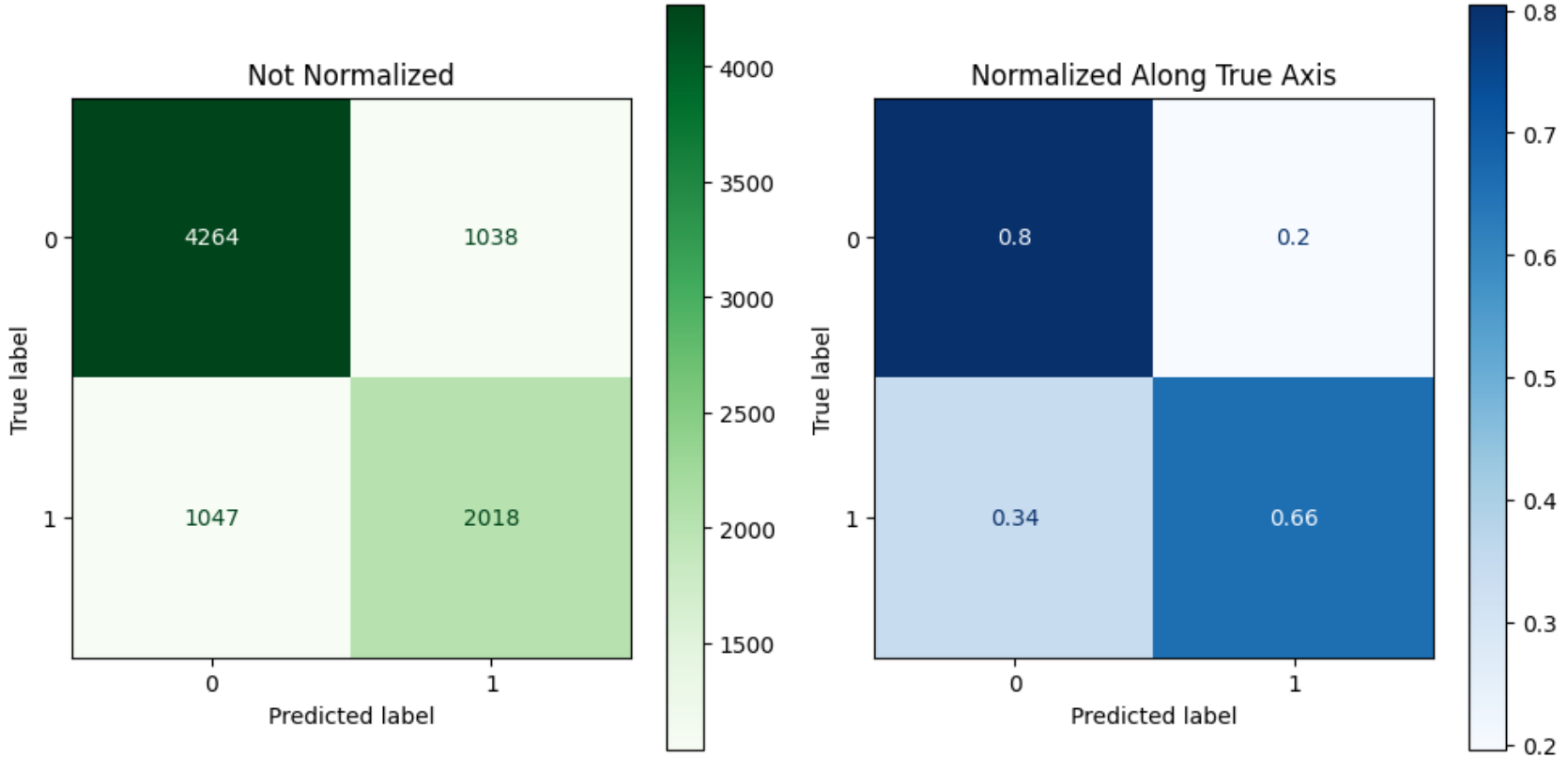
Comparison of implemented models*

	KNeighbors	GradientBoosting	LightGBM	XGBoost
Accuracy @ Train (in %)	100	78	78.7	76.5
Accuracy @ Test (in %)	72.9	74.8	74.4	75.1
F1 Score @ Train (in %)	100	76.3	77.2	74.7
F1 Score @ Test (in %)	70.2	72.9	72.5	73.1

* Tuned model measuring on ROC AUC

Model Design

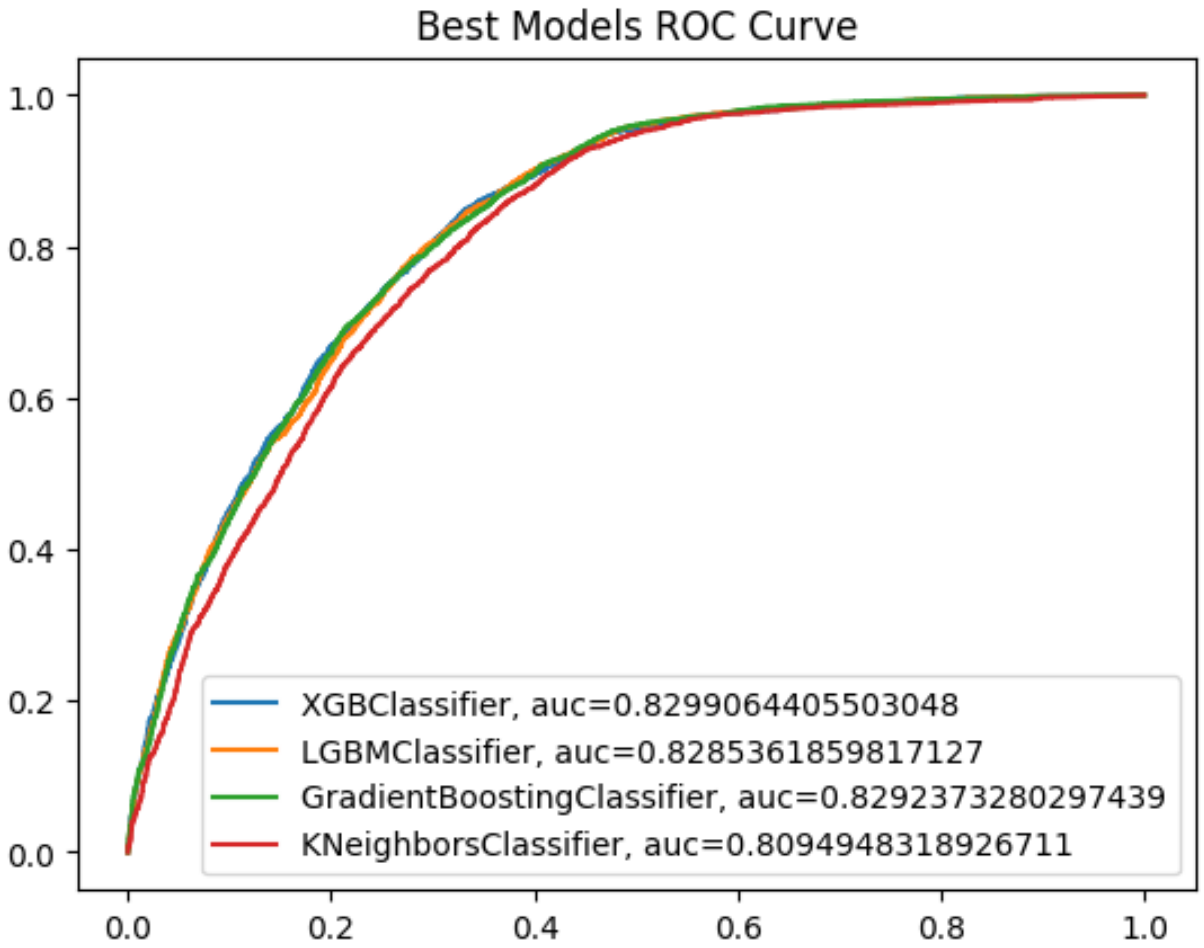
Confusion Matrix of the Best Model



* tuned by max_depth only

Model Design

ROC AUC Curves for all the Best Models



Thank you