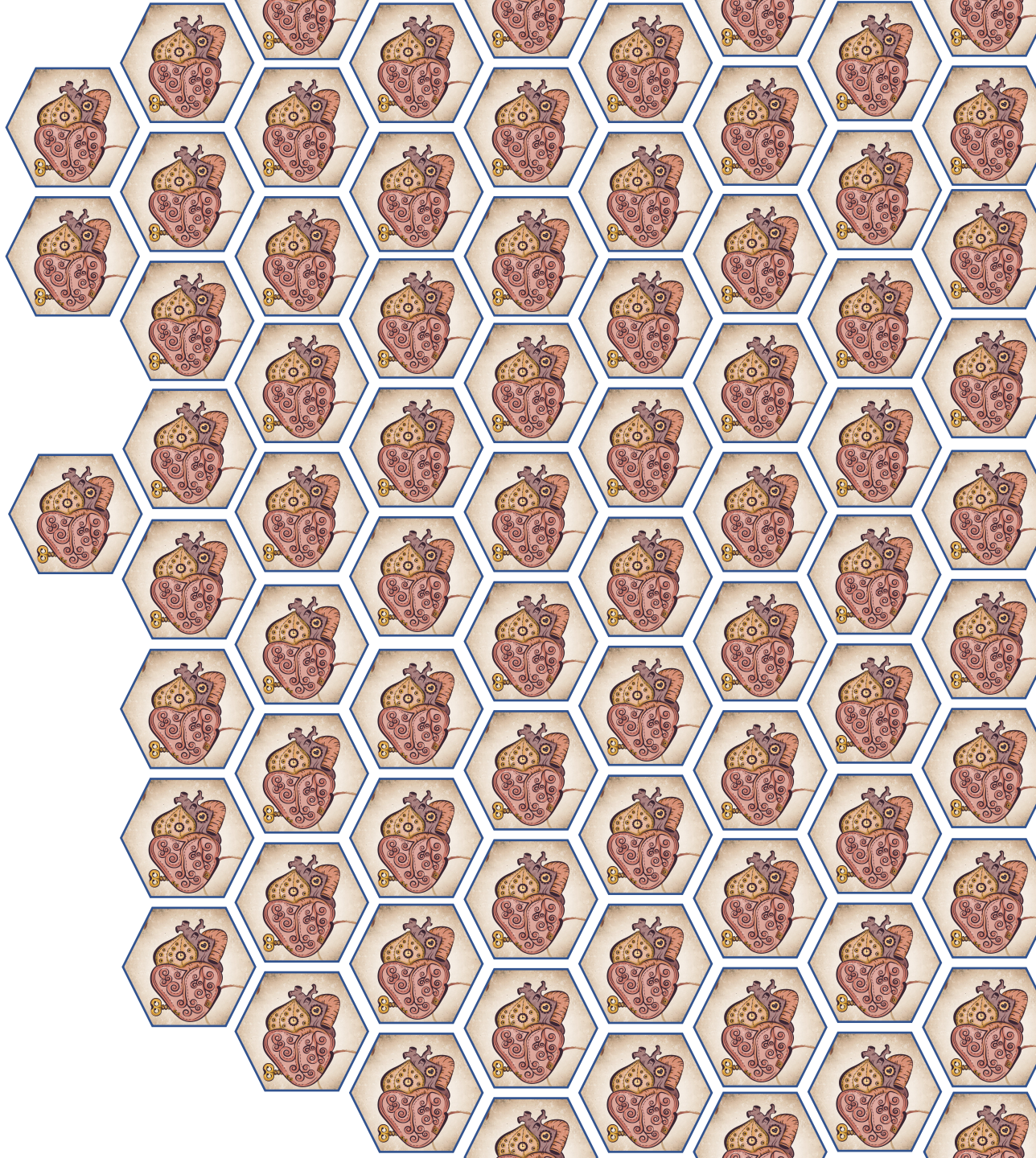
Risk Factors For Cardiovascular Heart Diseases

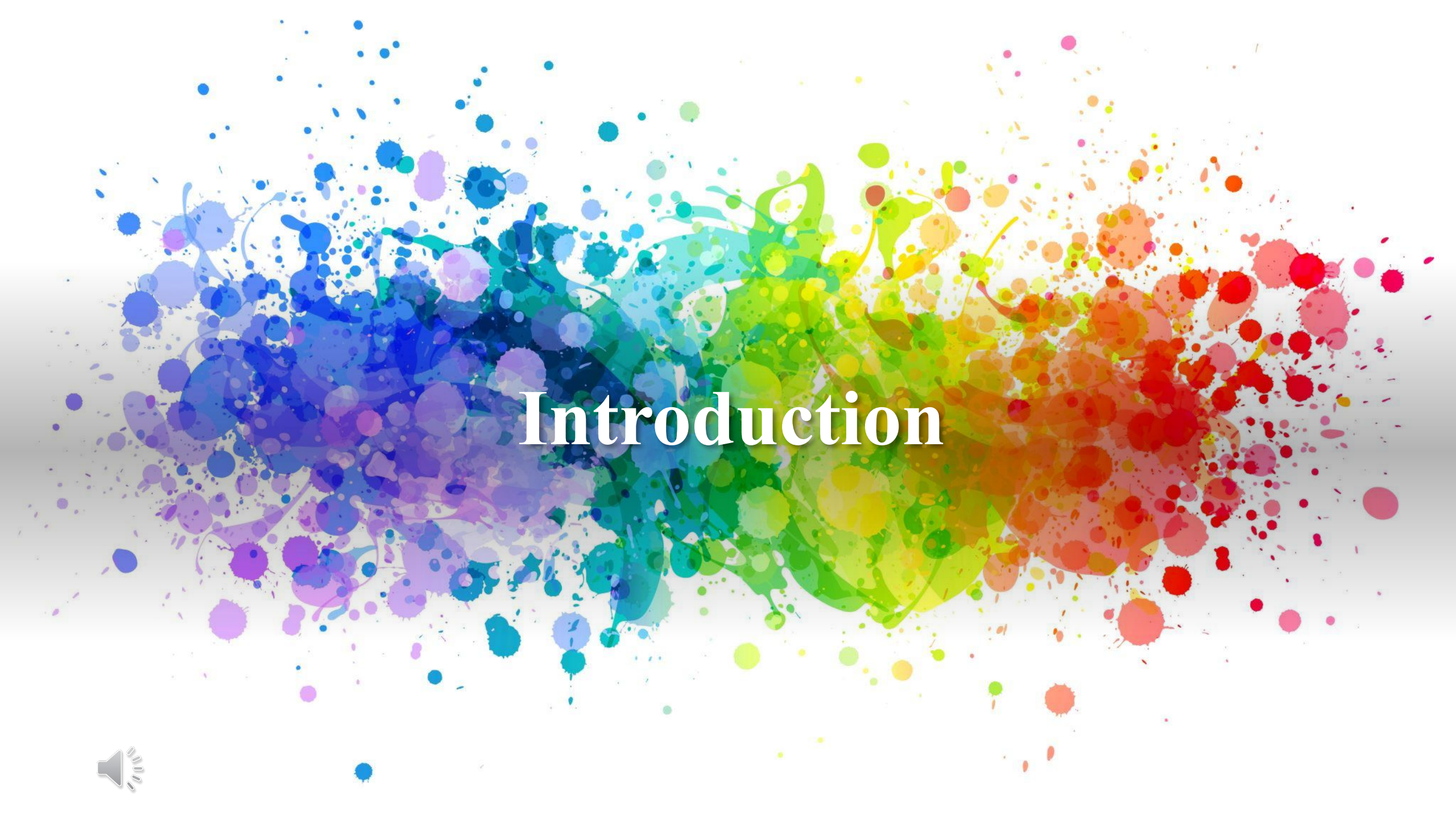
Jorge Méndez-Benegassi Márquez
Augusto Díaz-Leante Ferreiro
Álvaro Barja Galles
Fernando Heredero López

04/30/2023



Index

- Introduction
- Timeline & Methodology
- Data Preprocessing
- Data Exploration
- Data Mining
- Conclusions

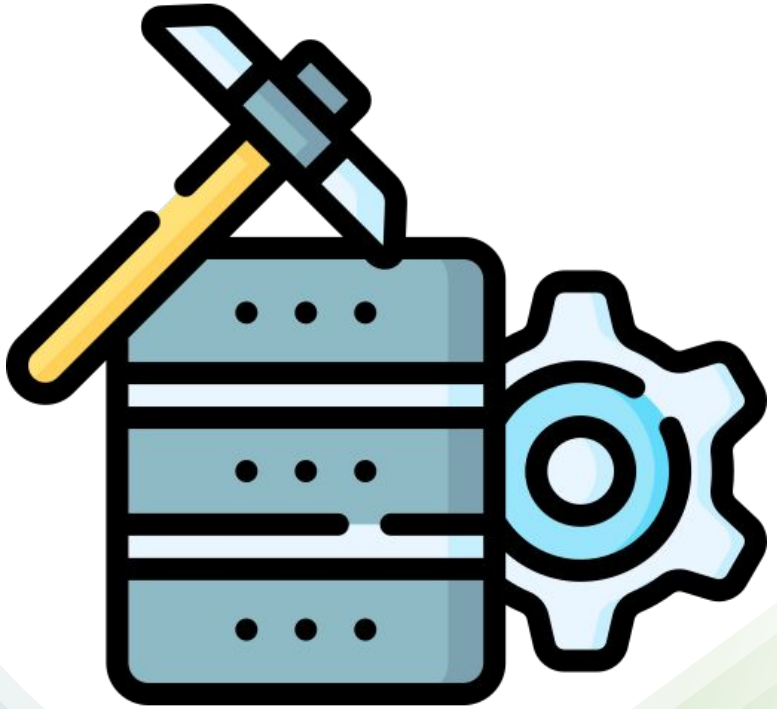

Introduction

Introduction

- Data mining helps improve medical diagnosis and treatment plans by identifying patterns and trends in patient data across different medical fields.

- Use data mining to analyze patient data so we can identify risk factors for cardiovascular disease.

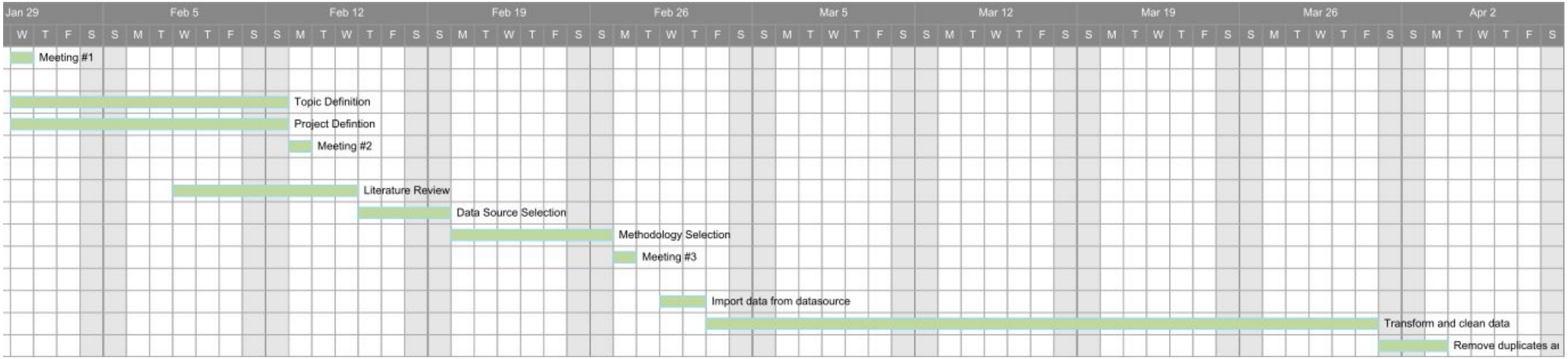
- "Data analytics will be the key to unlocking the potential of personalized medicine, allowing clinicians to make more precise diagnoses, tailor treatments to individual patients, and ultimately improve health outcomes.“ Dr. Eric Topol



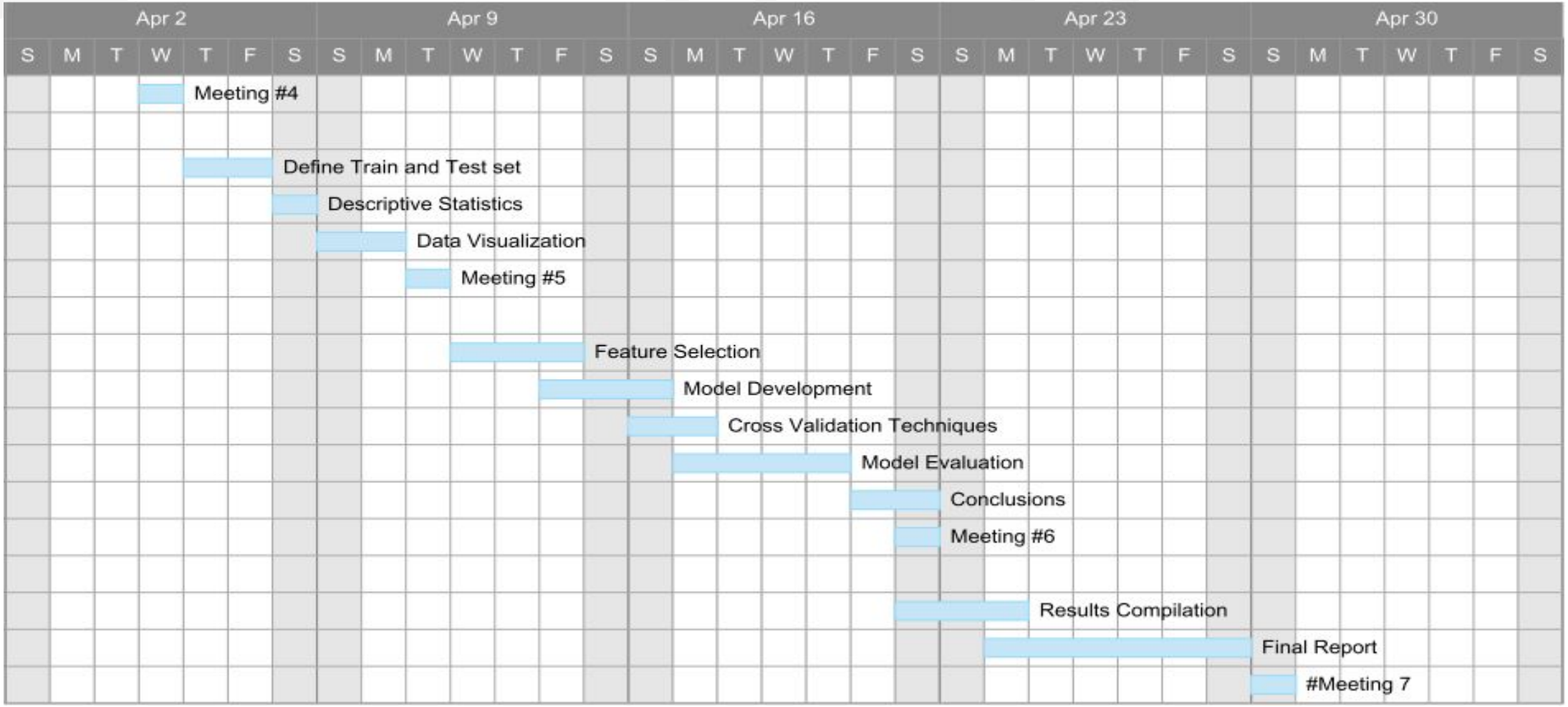

Timeline & Methodology

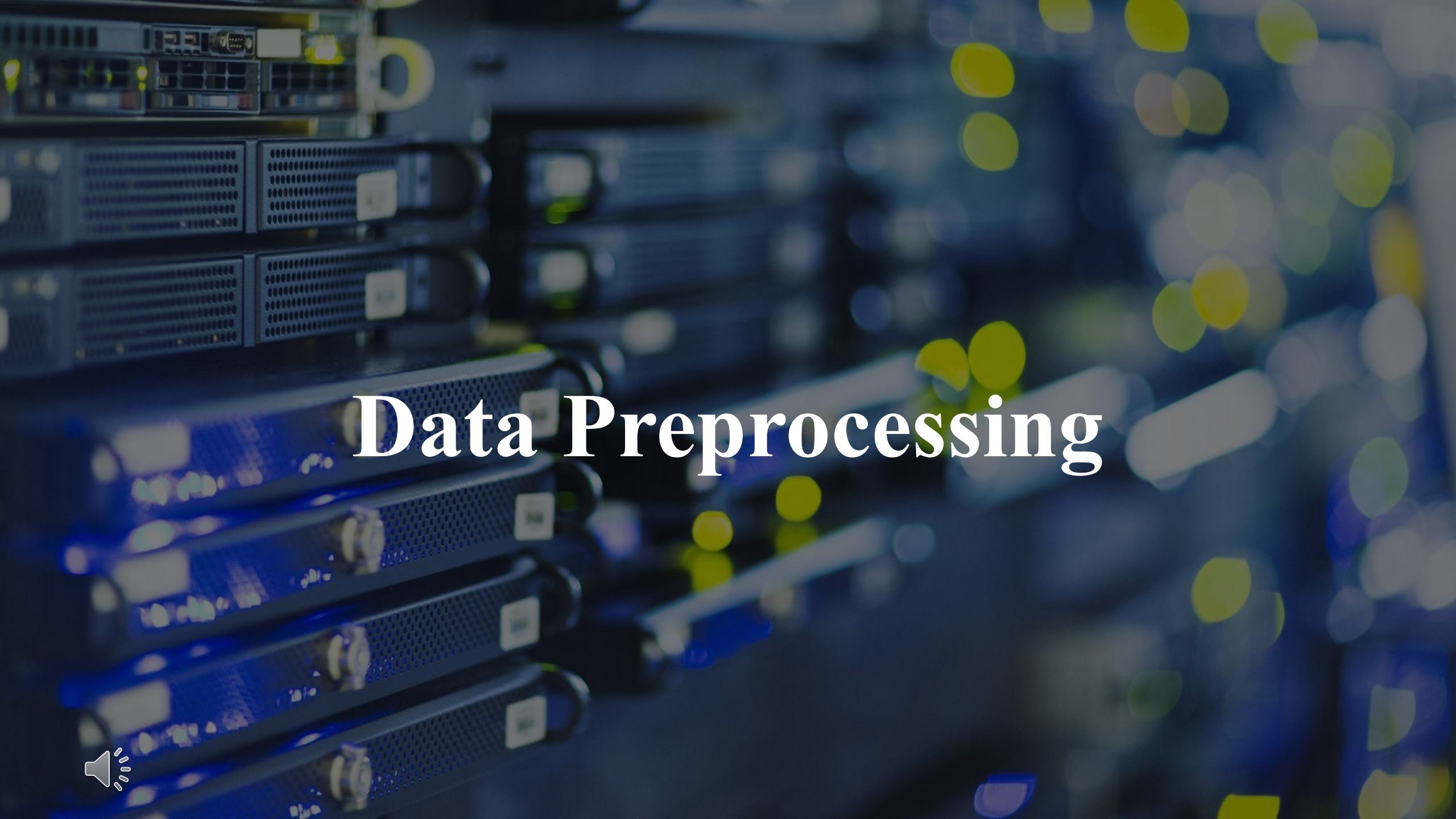
Project Timeline & Methodology

- Scrum Methodologies
- Gant diagram



Project Timeline & Methodology




Data Preprocessing

Data Preprocessing - Dataset

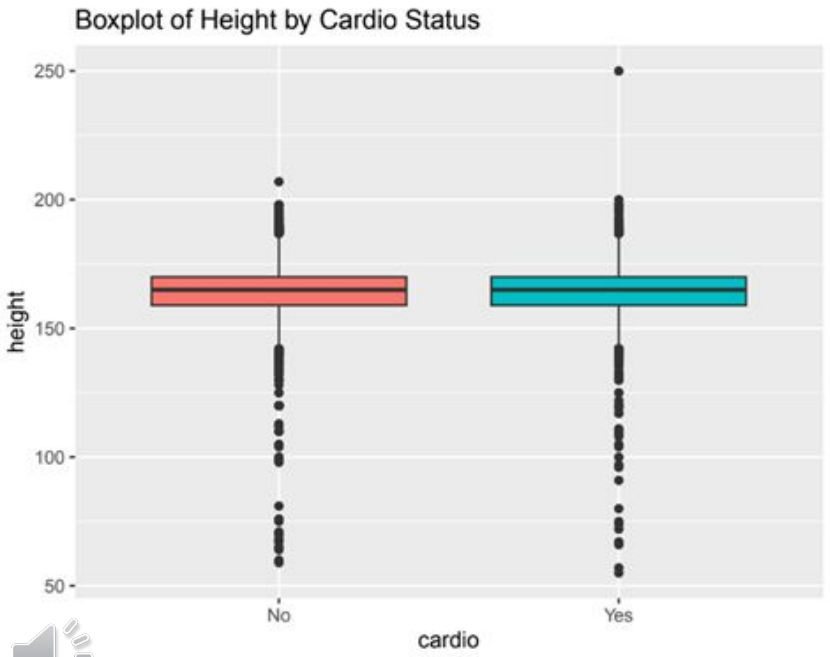
Field name	Type	Description
Age	Integer	Age of the individual
Gender	Binary	Gender of the participant
Height	Integer	Height in cm
Weight	Integer	Weight in Kg
Ap_hi	Integer	Systolic blood pressure
Ap_lo	Integer	Diastolic blood pressure
Cholesterol	Categorical (3 levels)	Cholesterol level
Glucose	Categorical (3 levels)	Glucose level
Smoke	Binary	Smoking status
Alcohol	Binary	Alcohol consumption status
Active	Binary	Physical activity level
Cardio	Binary	Presence or absence of cardiovascular disease

Data Preprocessing

Blood Pressure Features

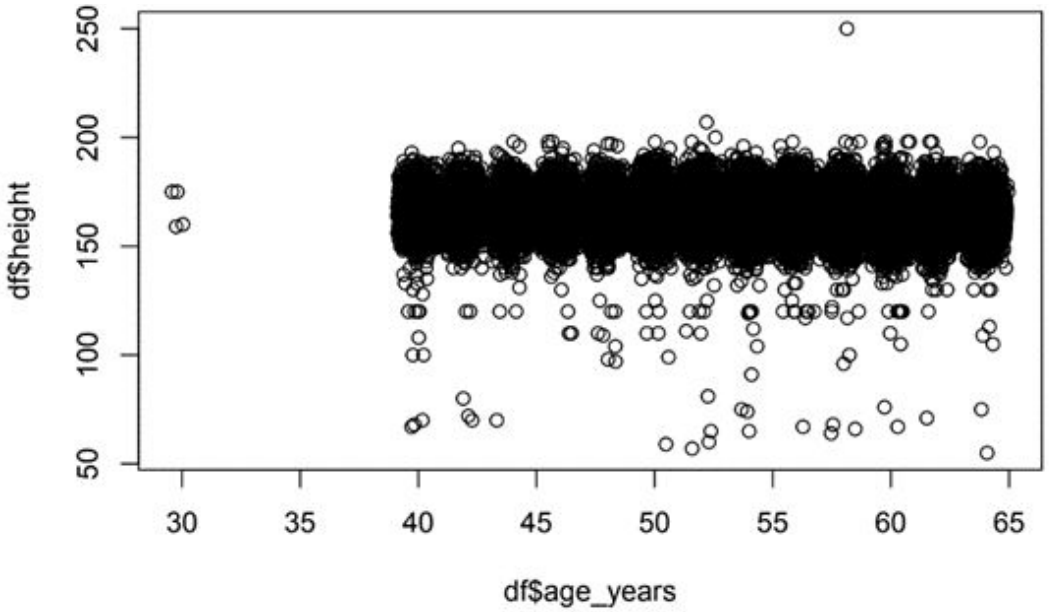
Systolic outliers → Values greater than 180 mmHg

Diastolic outliers → Values greater than 120 mmHg



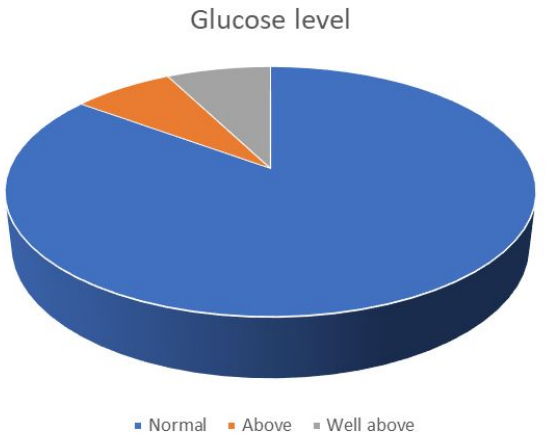
Height Feature

Outlier Elimination →

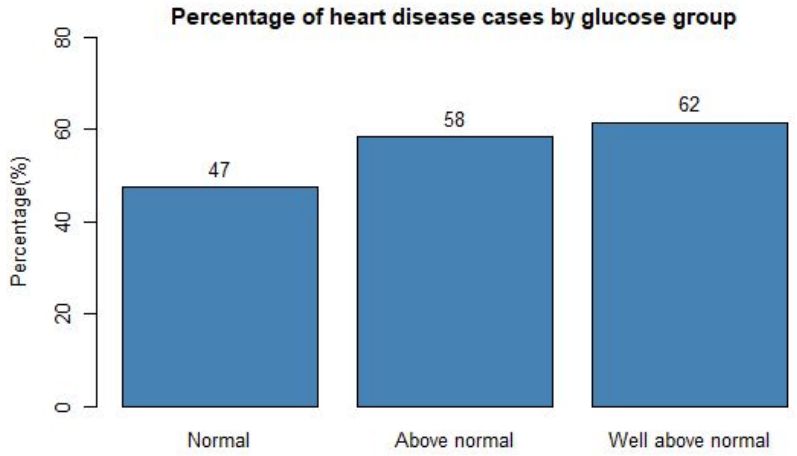


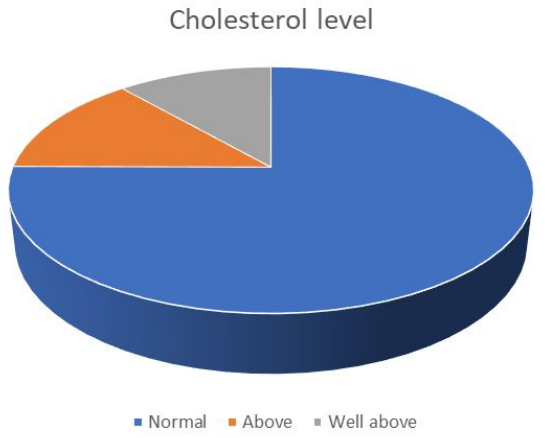
Data Exploration

Data Exploration

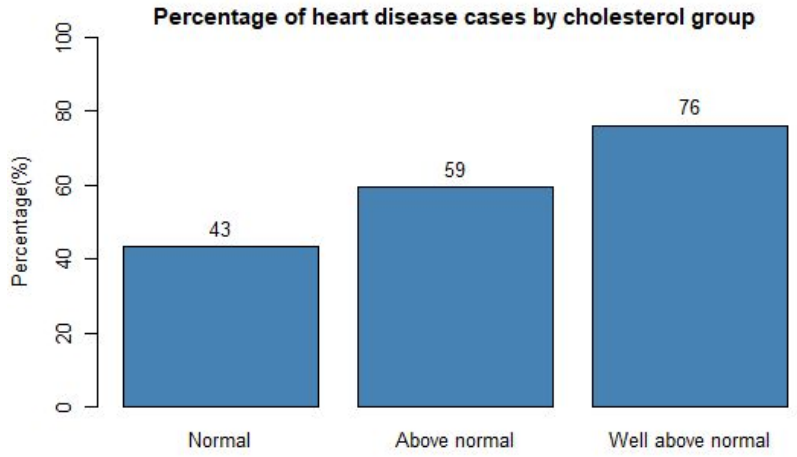


Considering Cardio Cases



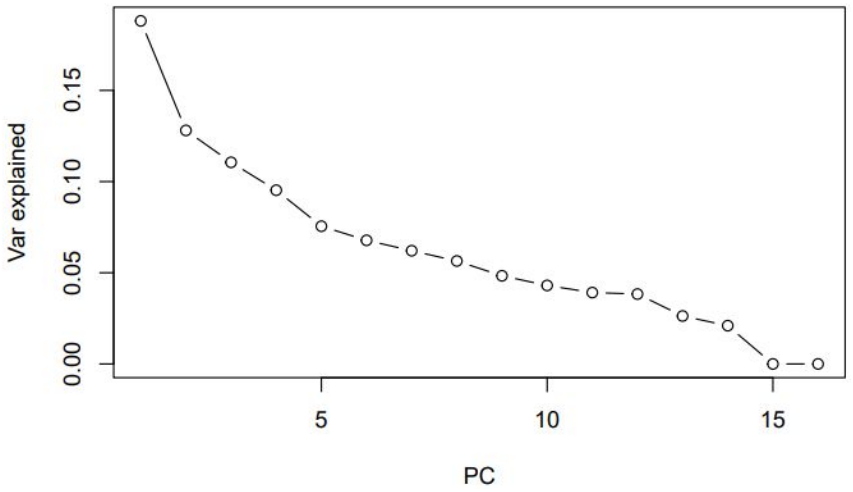


Considering Cardio Cases



Data Mining

Data Mining – Feature Selection

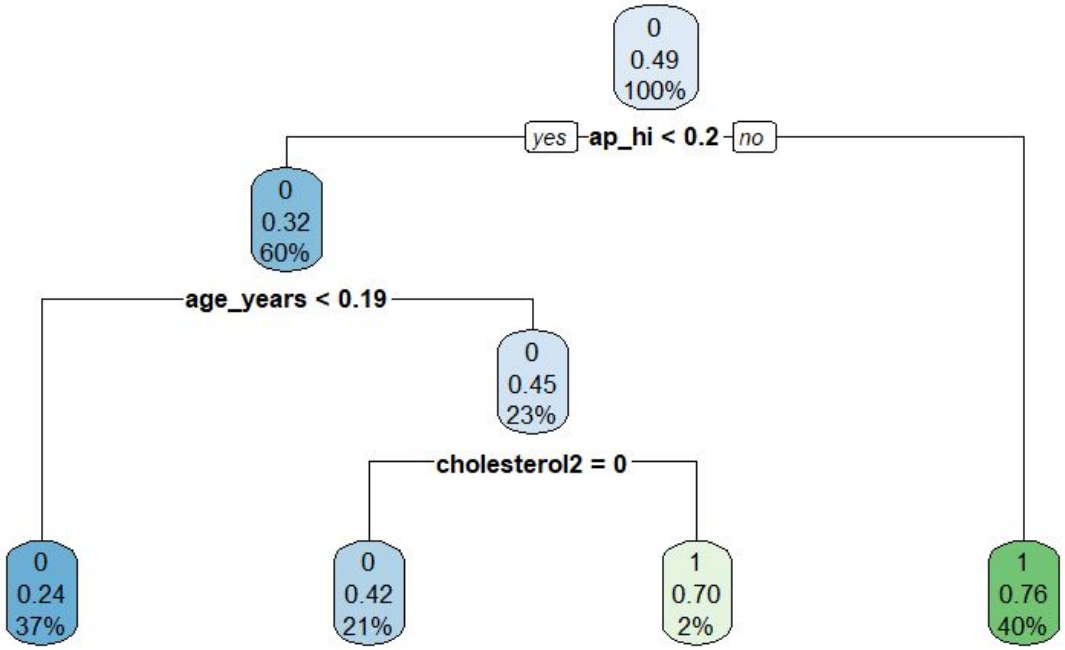


	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.6534870	0.0075450	86.612	< 2e-16	***
gender	-0.0009582	0.0043451	-0.221	0.825	
height	-0.0085488	0.0020783	-4.113	3.90e-05	***
weight	0.0342596	0.0018710	18.311	< 2e-16	***
ap_hi	0.1362601	0.0022887	59.535	< 2e-16	***
ap_lo	0.0477322	0.0022569	21.150	< 2e-16	***
cholesterol0	-0.2089635	0.0062606	-33.378	< 2e-16	***
cholesterol1	-0.1238404	0.0074709	-16.576	< 2e-16	***
cholesterol2	NA	NA	NA	NA	
gluc0	0.0583522	0.0073123	7.980	1.49e-15	***
gluc1	0.0607421	0.0095079	6.389	1.68e-10	***
gluc2	NA	NA	NA	NA	
smoke	-0.0259750	0.0066651	-3.897	9.74e-05	***
alco	-0.0402084	0.0080269	-5.009	5.48e-07	***
active	-0.0448806	0.0042515	-10.556	< 2e-16	***
age_years	0.0726252	0.0017476	41.556	< 2e-16	***

- Principal Component Analysis (PCA).
- Correlation Analysis
- Significant Feature Analysis

Data Mining – Model Training

Decision Tree



Accuracy: 0.725

Random Forest

Accuracy: 0.737

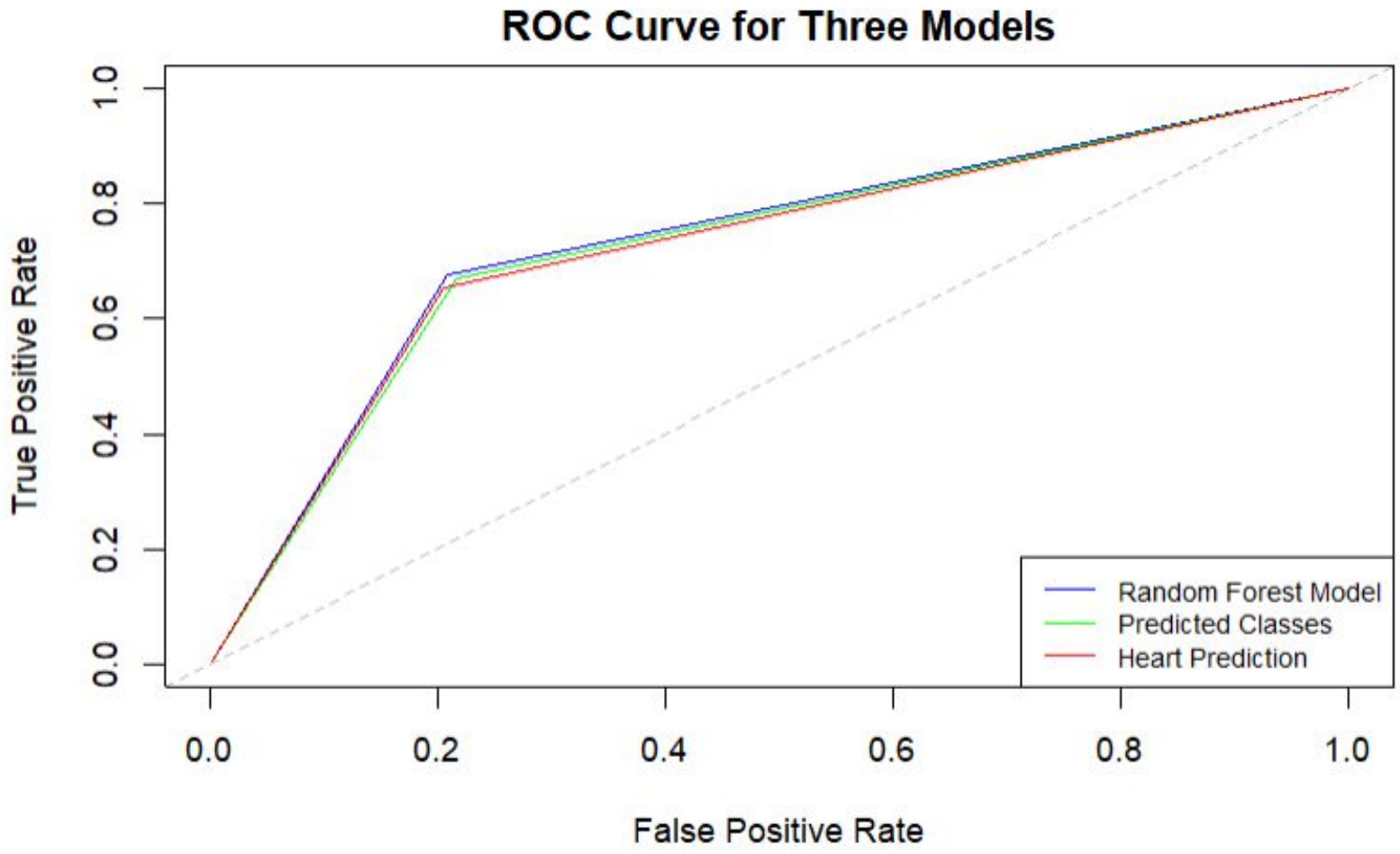
Logistic Regression

Accuracy: 0.729

Data Mining – Model Evaluation

	Binary Decision Tree	Random Forest	Logistic Regression
Accuracy	0.725	0.737	0.729
Sensitivity	0.795	0.796	0,784
Specificity	0.6523	0, 676	0,671
Precision	0.701	0.716	0.71
Recall	0.795	0.796	0.785
F1 Score	0.745	0.754	0.745

Data Mining – Model Evaluation



Conclusions & Future Work

Conclusions

- High cholesterol and glucose levels increase the risk of heart disease
- Gender does not influence the occurrence of heart disease
- Random forest analysis provides the best results on the dataset used
- Continued investment in data mining and analysis could lead to more personalized and effective interventions for preventing and treating heart disease.

Future Work

- Explore additional features to better understand heart disease risk
- Compare the performance of alternative ensemble methods.
- Implement other machine learning models to improve the accuracy of predicting heart disease.