Cardiovascular Risk Assesment

Data Ninja

Meet the Team: Data Ninjas

Ellis Porter

Isha Chaware

Reetu Jakhar

Julia Dettman

Alex Turner

Xiang Li

Introduction

Welcome to our Cardiovascular Risk Assessment Project, an initiative dedicated to unlocking insights from a comprehensive dataset encompassing vital health metrics. Our dataset, named "cardio_train", is a rich compilation of various health-related attributes from numerous individuals, focusing on factors that are critical in understanding and predicting cardiovascular diseases.

Applying the Data

Overview

Our analysis revolves around data points like age, gender, physical measurements (height and weight), blood pressure readings (systolic and diastolic), cholesterol levels, glucose levels, and lifestyle factors such as smoking, alcohol consumption, and physical activity. Most significantly, it includes a critical indicator of the presence or absence of cardiovascular disease, offering a valuable opportunity to explore correlations and patterns that may assist in predicting cardiovascular health risks.

Goals

Identify key factors contributing to cardiovascular diseases.

Understand the interplay between lifestyle choices and cardiovascular health.

Develop predictive models to assess the risk of cardiovascular diseases.

Create a model to predict hospital readmission using machine learning model on a healthcare dataset with these input features:

•Objective: factual information;

•Examination: results of medical examination;

•Subjective: information given by the patient.

Importance

Cardiovascular diseases are among the leading causes of morbidity and mortality worldwide. By analyzing this dataset, we aim to contribute to the broader understanding of these diseases and potentially aid in the development of strategies for prevention and early detection. This could help reduce costly readmissions.

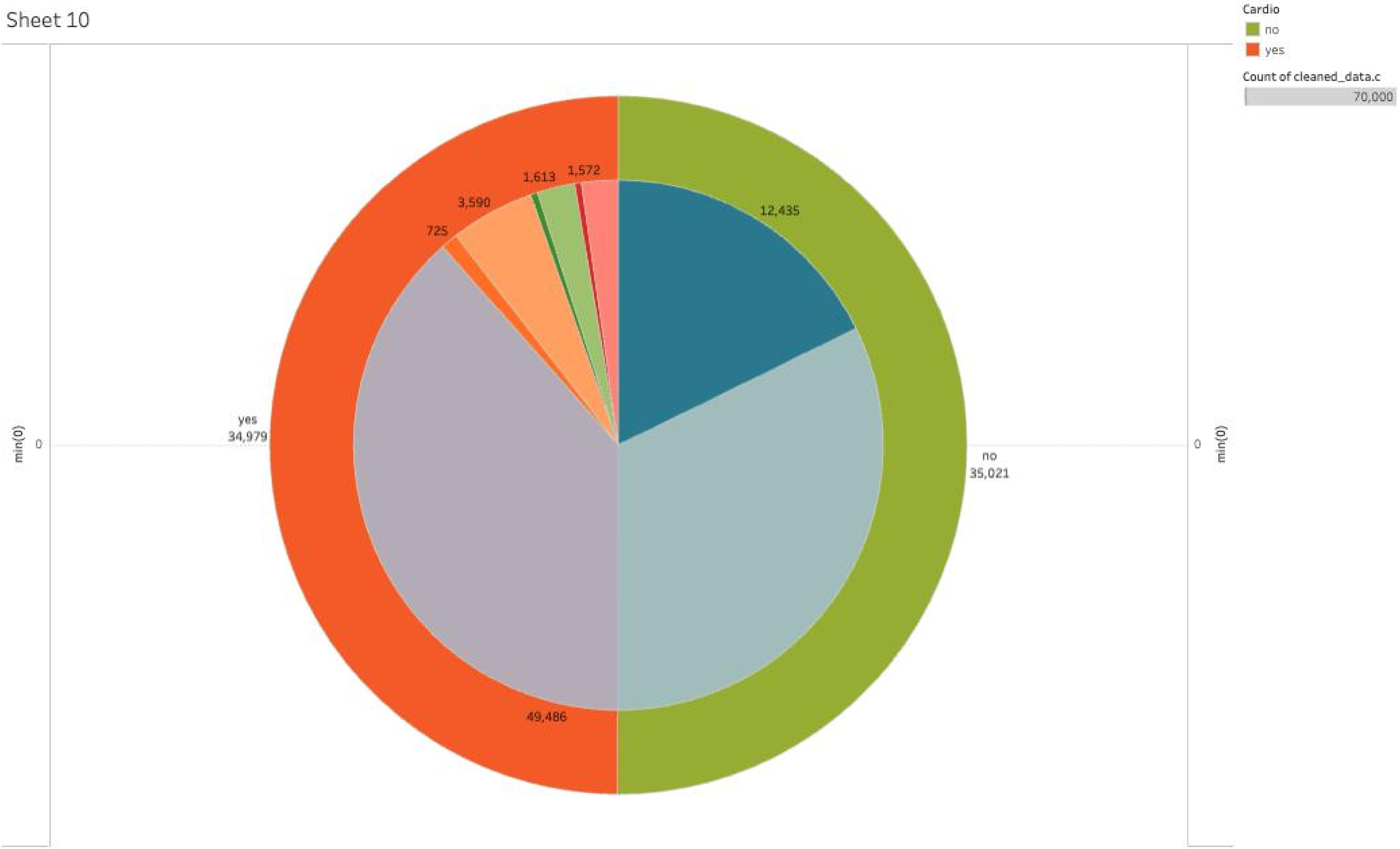
Data Set & Data Cleaning

https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset?resource=download

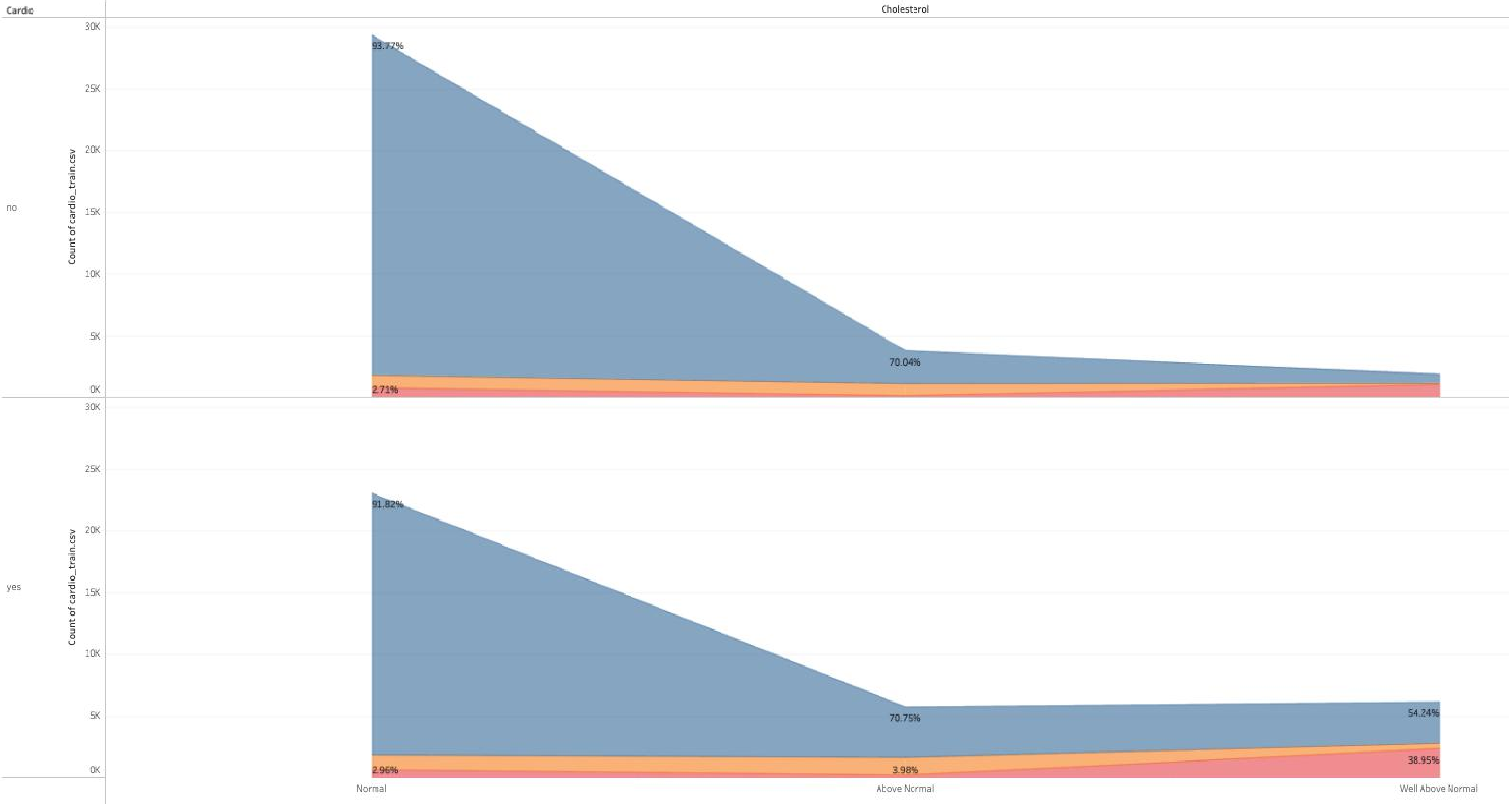
- Dropped Null Values
- Added a new column "age_years"
- Converted age in days to age in years
- Dropped duplicates

Visualizations

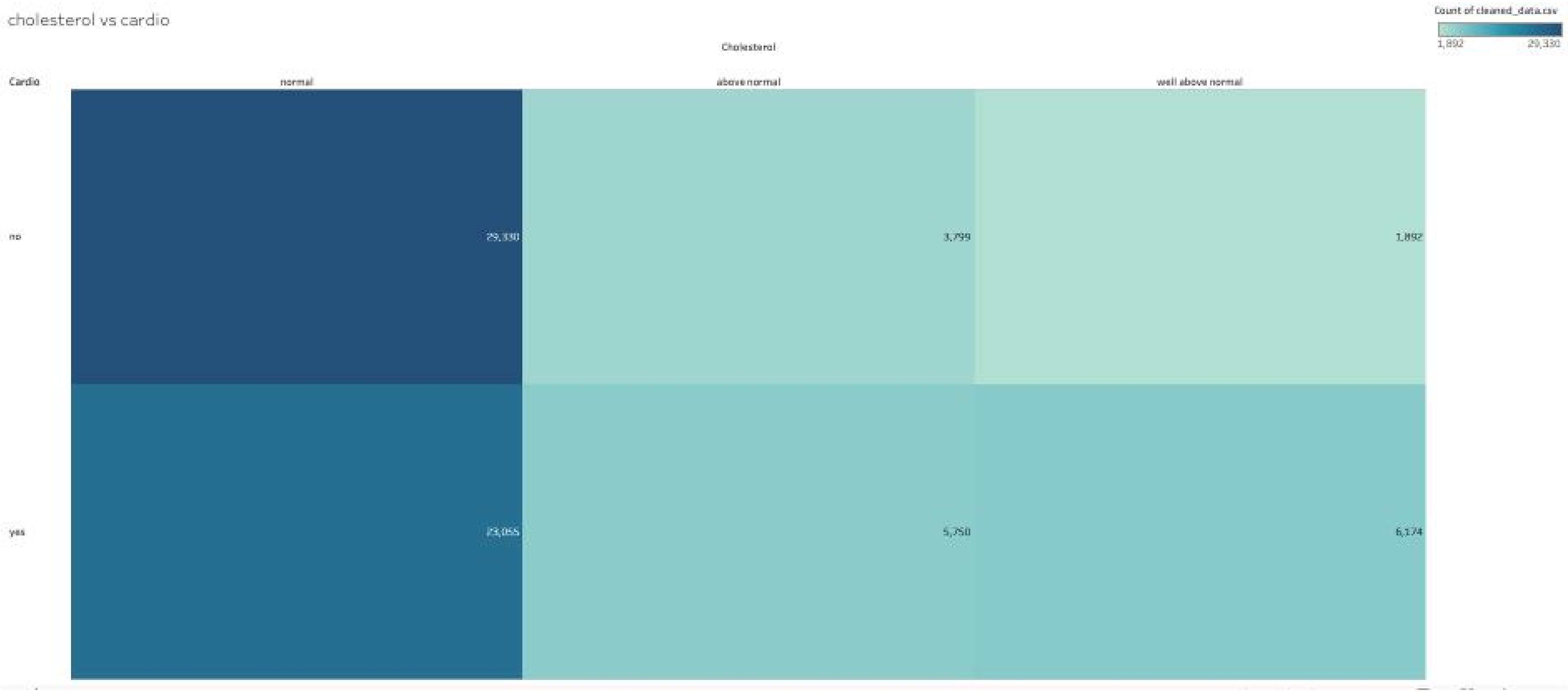
Data Distribution



Cholesterol/Gluc vs Cardio

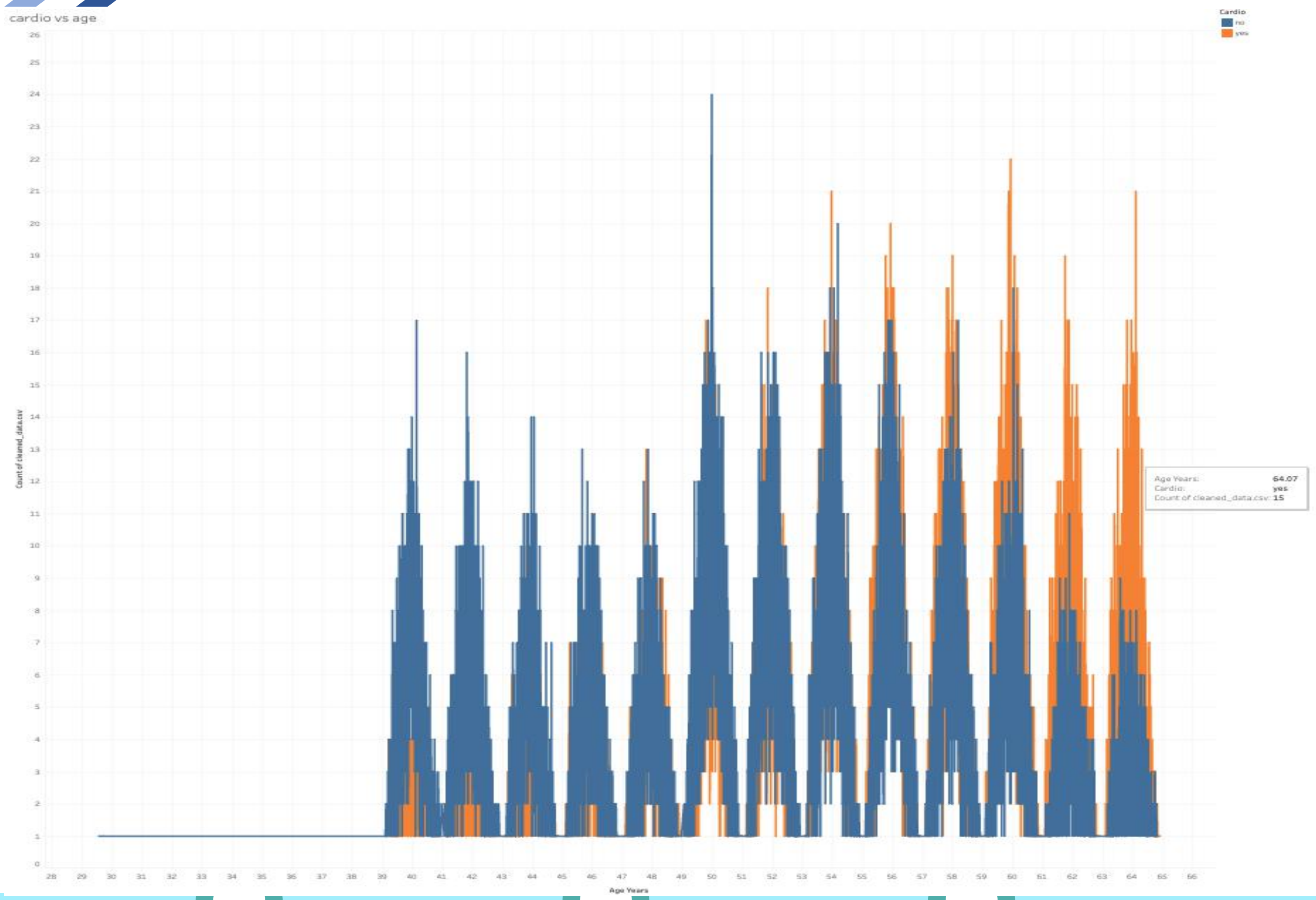


Cholesterol vs Cardio



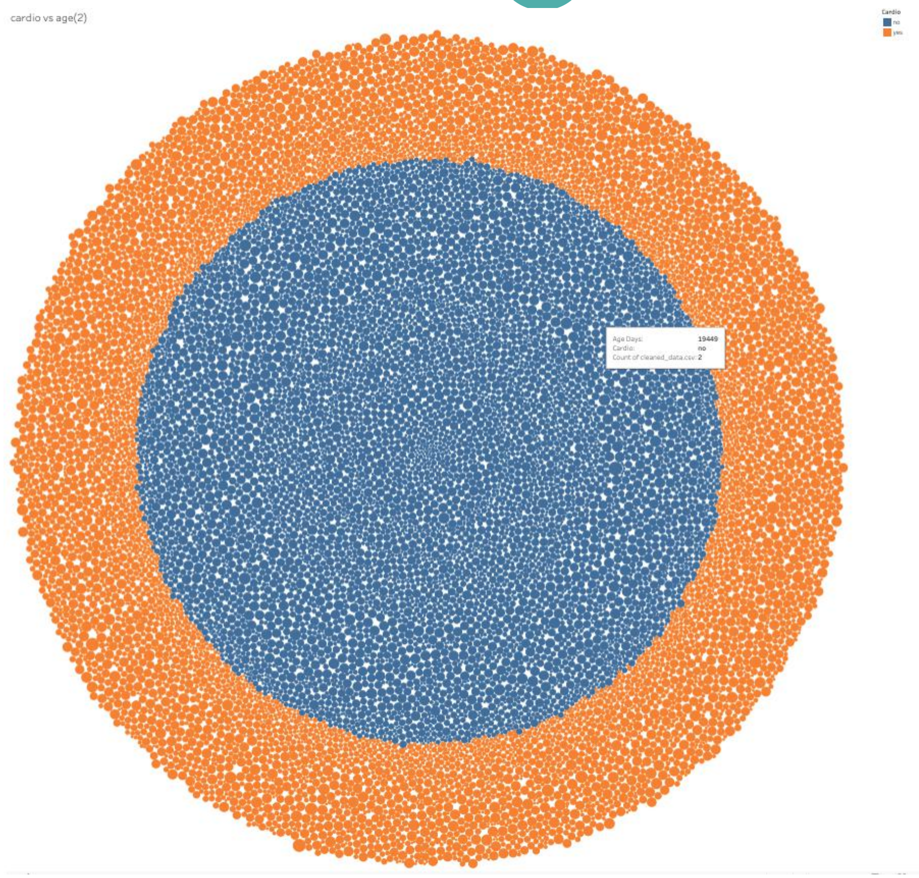
	Cholesterol		
Cardio	normal	above normal	well above normal
no	29,330	3,799	1,892
yes	23,055	5,750	6,174

Age vs Cardio



cardio vs age

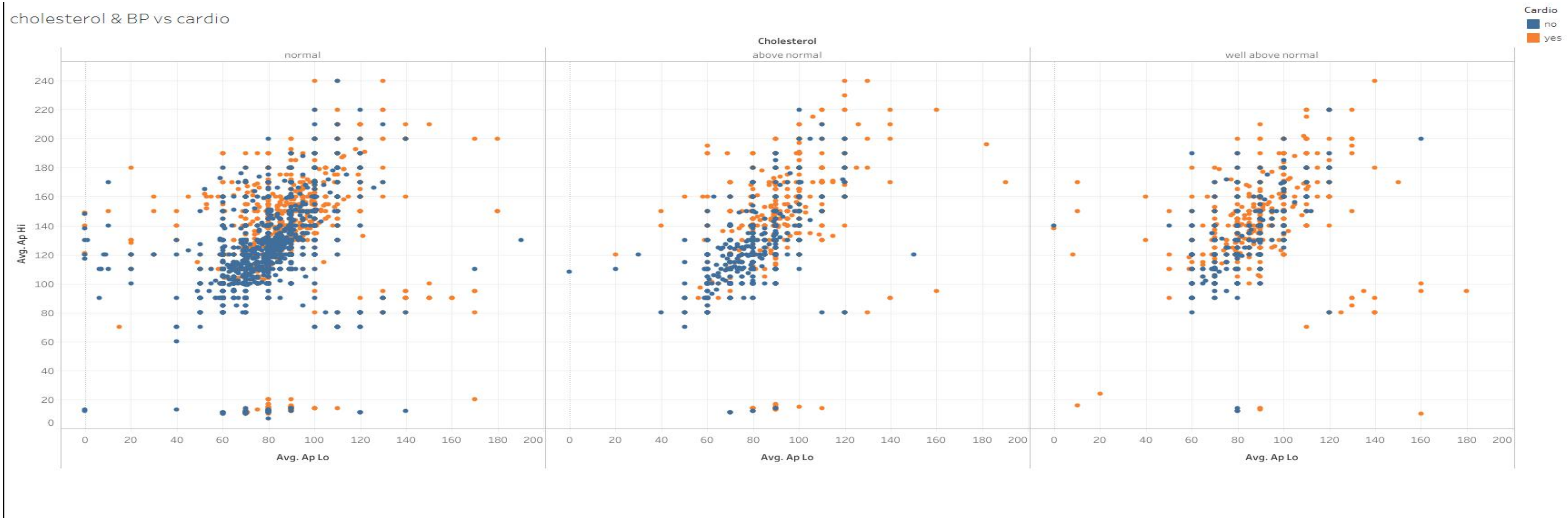
Age Years: 64.07
Cardio: yes
Count of cleaned_data.csv: 15



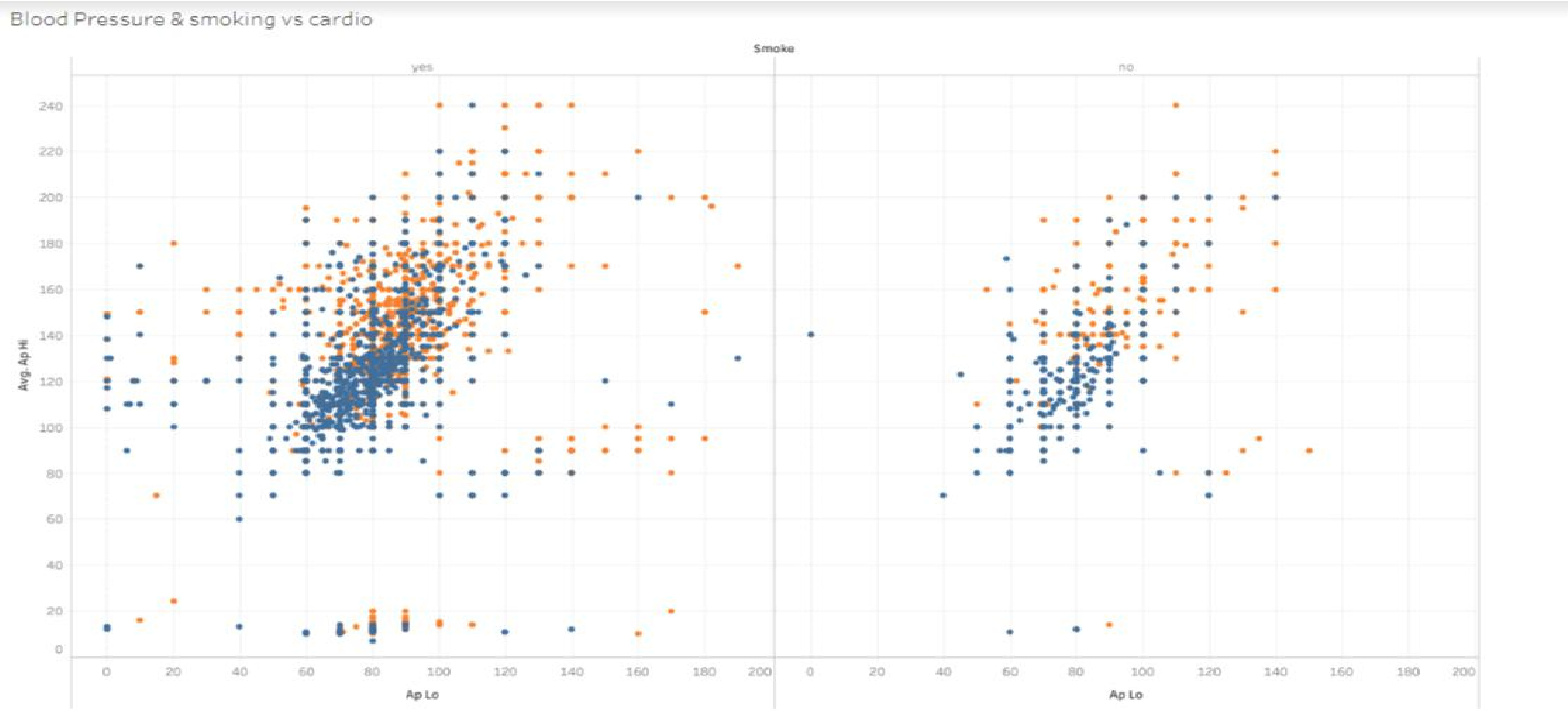
cardio vs age(2)

Age Days: 19449
Cardio: no
Count of cleaned_data.csv: 2

Cholesterol/BP vs Cardio



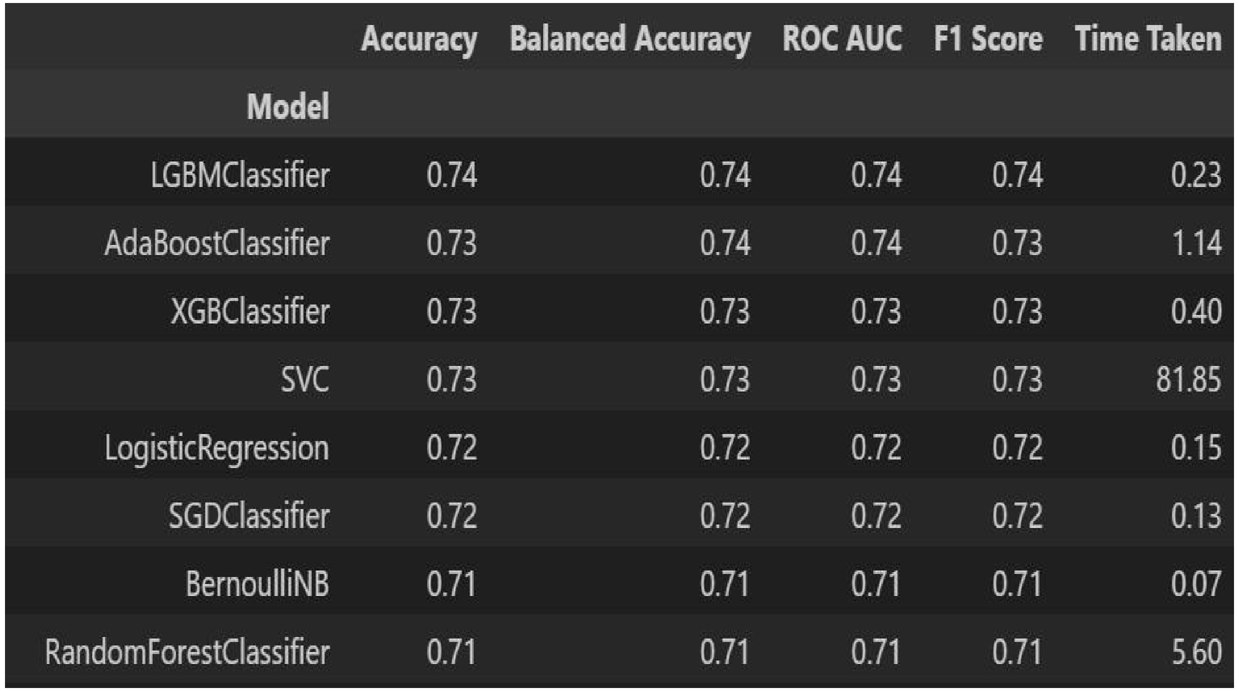
Smoking/BP vs Cardio


Blood Pressure & smoking vs cardio

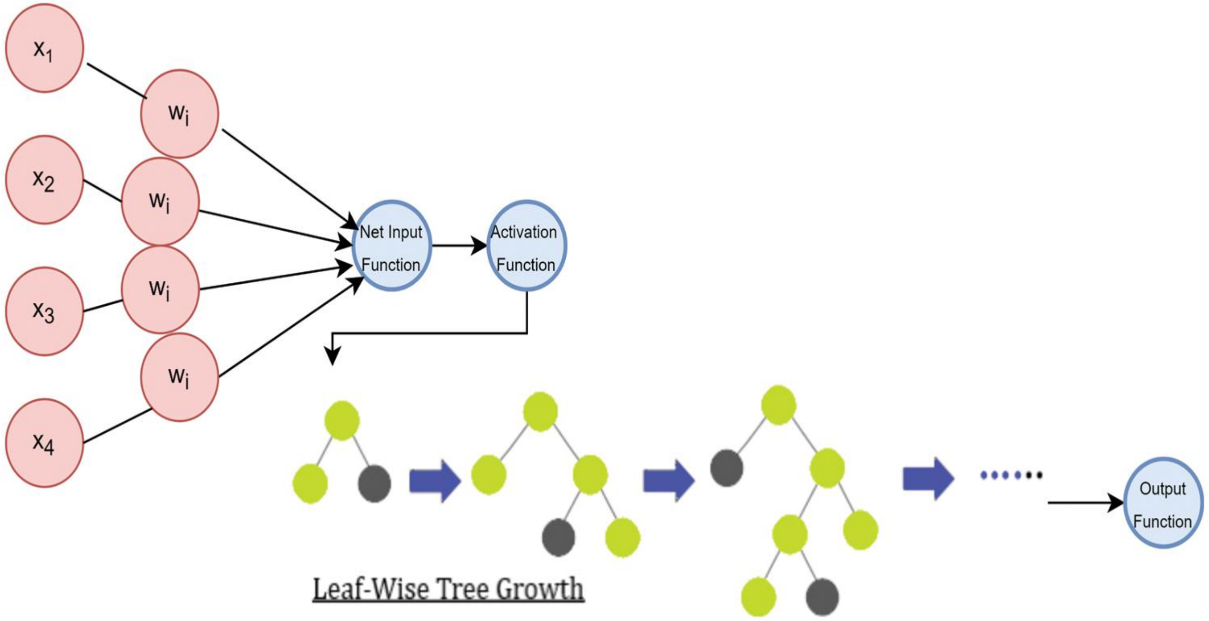
Machine Learning

Model Selection

- Ethics regarding data collection errors and navigating outliers
- Chose to keep dataset as is and use decision tree model to navigate outliers
- Used LazyClassifier to compare models
- Top 3 performers were decision tree based models



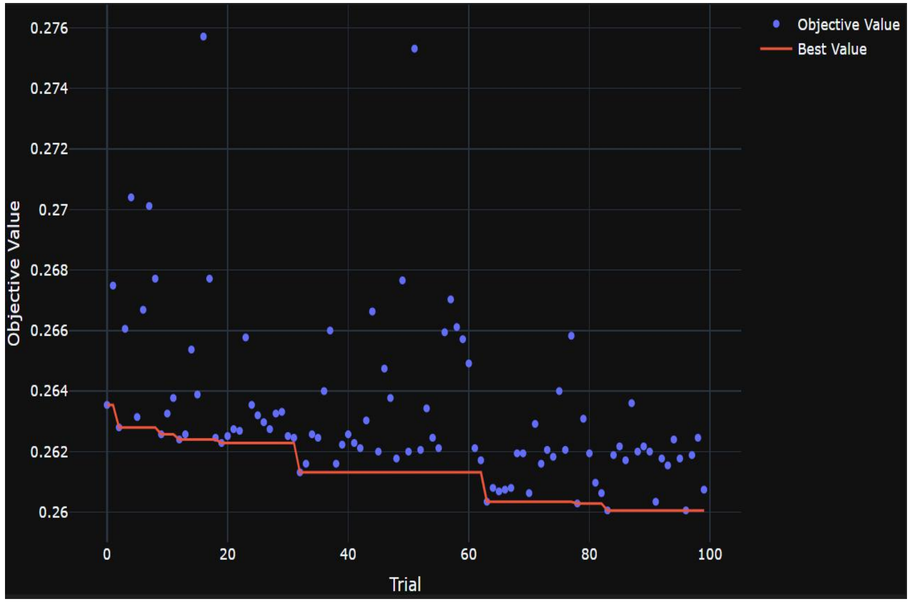
Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
LGBMClassifier	0.74	0.74	0.74	0.74	0.23
AdaBoostClassifier	0.73	0.74	0.74	0.73	1.14
XGBClassifier	0.73	0.73	0.73	0.73	0.40
SVC	0.73	0.73	0.73	0.73	81.85
LogisticRegression	0.72	0.72	0.72	0.72	0.15
SGDClassifier	0.72	0.72	0.72	0.72	0.13
BernoulliNB	0.71	0.71	0.71	0.71	0.07
RandomForestClassifier	0.71	0.71	0.71	0.71	5.60

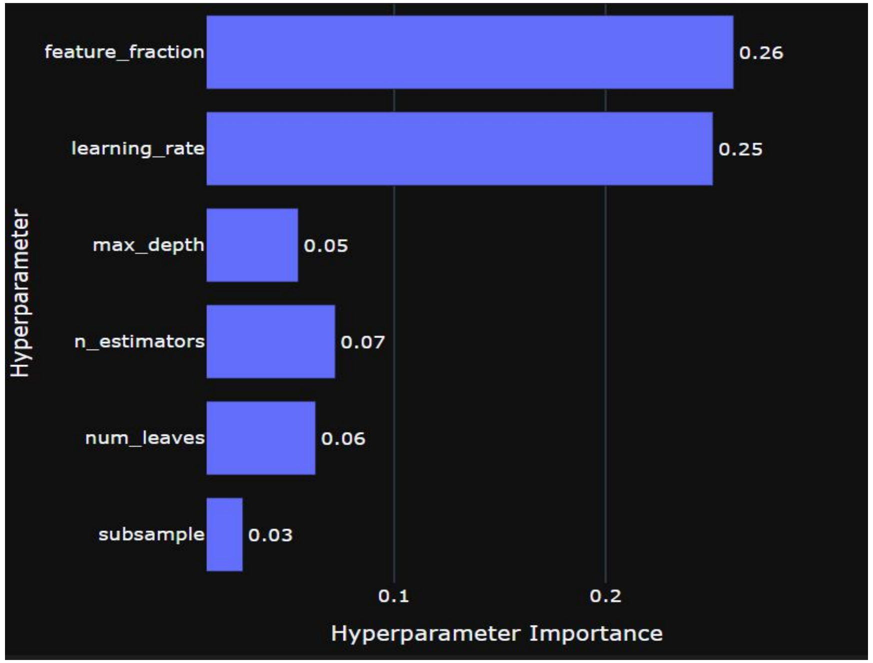


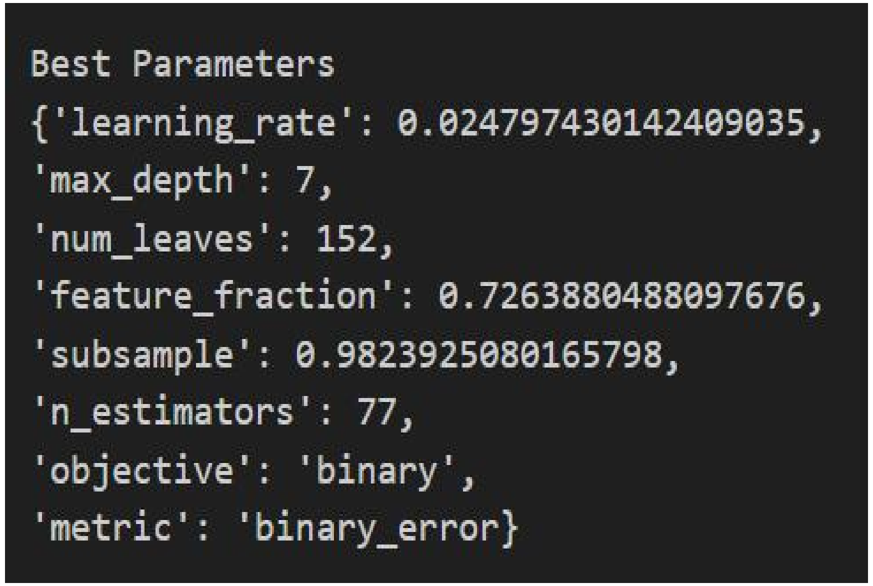
- Light Gradient—Boosted Machine (LGBM) Classifier
- Strengths: efficiency, scalability, accuracy, handling of mixed numerical and categorical variables
- Leaf—wise Growth:
  - prioritizes nodes with most informational gain
  - improves efficiency
  - deeper and narrower trees
- Lightweight implementation for real—time predictions and limited comuptational power
- Weaknesses: overfitting, hyperparameters

Model Optimization

- Hyperparameter tuning with Optuna
- Objective function：minimize binary error
- Low max depth to help with overfitting
- High hyperparameter importance of feature fraction and learning rate



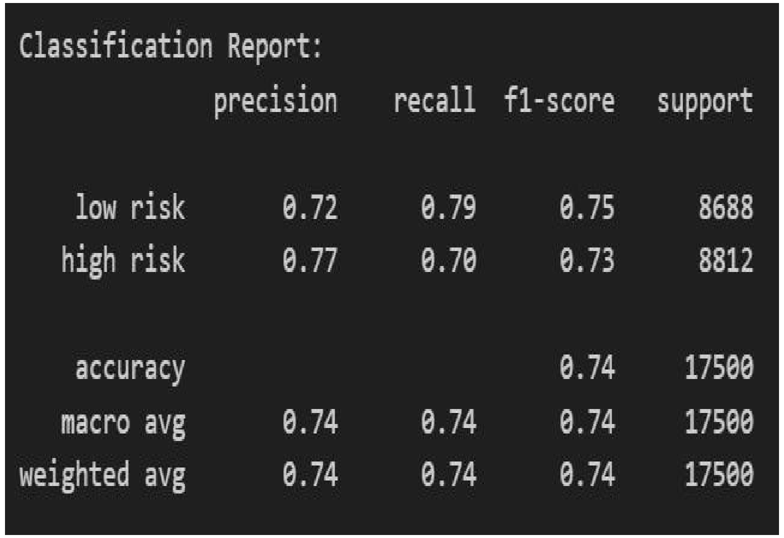




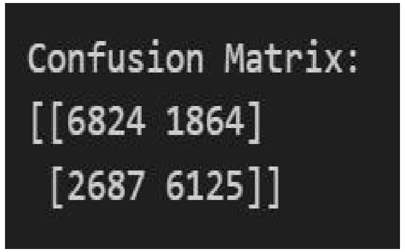
```
Best Parameters
{'learning_rate': 0.024797430142409035,
'max_depth': 7,
'num_leaves': 152,
'feature_fraction': 0.7263880488097676,
'subsample': 0.9823925080165798,
'n_estimators': 77,
'objective': 'binary',
'metric': 'binary_error}
```

Model Performance Analysis

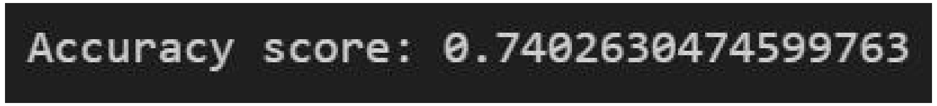
- "Good" accuracy score of 0.74
- Relatively balanced precision, recall and f1 score
- Slightly higher precision of high risk than low risk
- 70% of actual high risk correctly identified



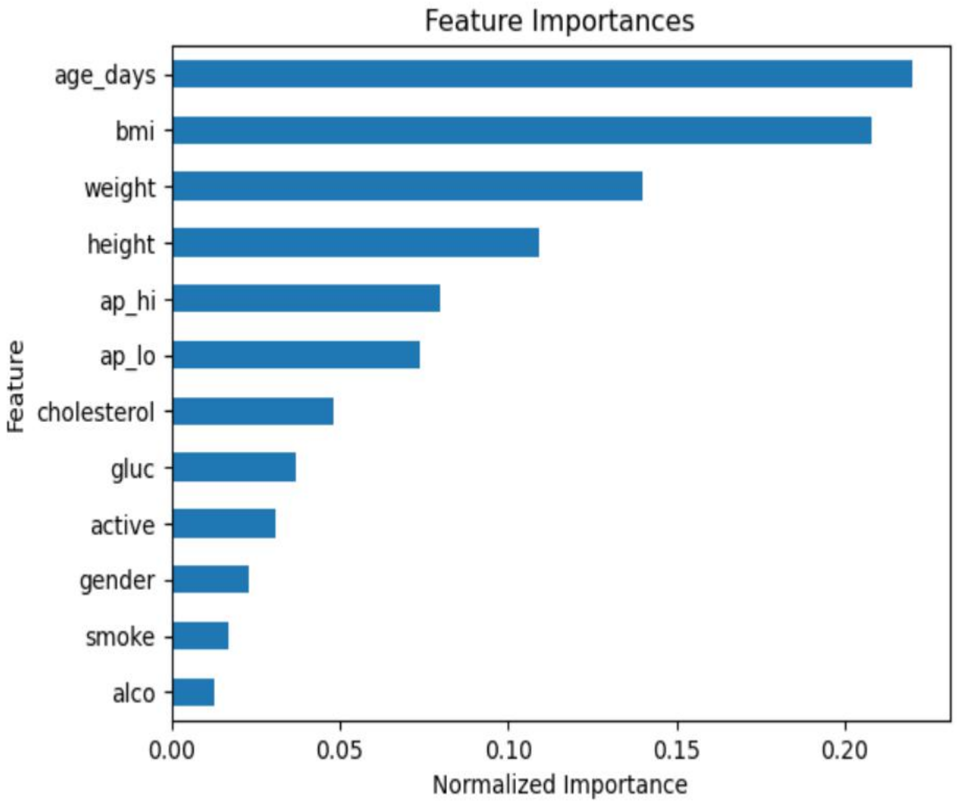
```
Classification Report:
              precision    recall  f1-score   support

   low risk       0.72      0.79      0.75      8688
  high risk       0.77      0.70      0.73      8812

   accuracy                           0.74     17500
  macro avg       0.74      0.74      0.74     17500
weighted avg       0.74      0.74      0.74     17500
```



```
Confusion Matrix:
[[6824 1864]
 [2687 6125]]
```



```
Accuracy score: 0.7402630474599763
```



- Most influential features： Age and BMI
- Systolic and diastolic pressure moderate
- Trends in importance by data types
  - Objective
  - Examination
  - Subjective

Continuing on....

Further questions and possible expansions

Limitations:

The model will only run if all the categories are fulfilled. Therefore, not everyone would present the data in the way we have. For example, having the cholestrol and glucose values be at value 1, 2 or 3 ; most people would simply present their cholestrol at face value. You would have to have exactly the same categories to run the model. The data for our target variable was so balanced, a 50/50 balance is not likely to happen on a real world data set. Our outliers and strange data ranges makes it seem as though this data set was oversampled previously as opposed to raw. Applying this to real life medical scenarios would be ethically wrong.

Future Directions:

With the scope of this project, the data we could find that was free and accessible was what we used. A lot of the limitations could be resolved with a more real world data set. A lot fo the data was hidden behind paywalls or privacy barriers since it is patient data. If we were to do this, we could implement this real time. This could be an extra tool the hospital could use to further patient care and lowe the cost of readmission.

Closing Thoughts:

The journey through the Cardiovascular Disease Analysis Project has been enlightening, offering a deeper understanding of one of the most pressing health issues of our time. Our findings underscore the importance of continuous research and public health efforts in combating cardiovascular diseases.

THANK YOU!