HealthCare Predictive Analysis

Cardiovascular Disease

Team Members:

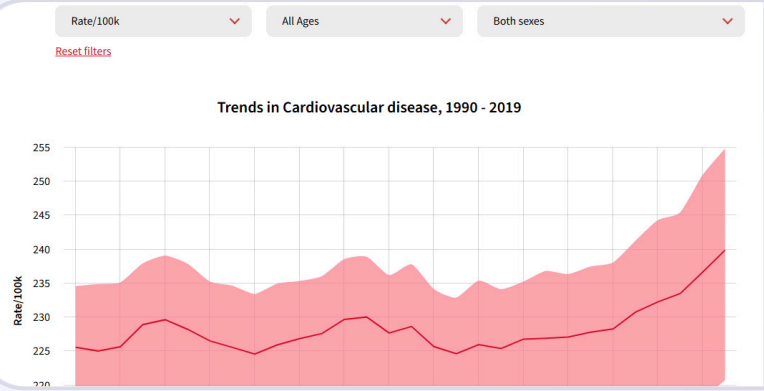
- Amina Zeyad Atta Ibrahim
- Amr Ashour Mohamed Ali
- Elham Mokhtar Ahmed Mohammed
- Enas Samir Abbas Mohamed
- Esraa Mohamed Sayed Mohamed

Training Company: CLS

Group Number: CLS__CAI2_AIS4_G4

Instructor: Mahmoud El-Sayed

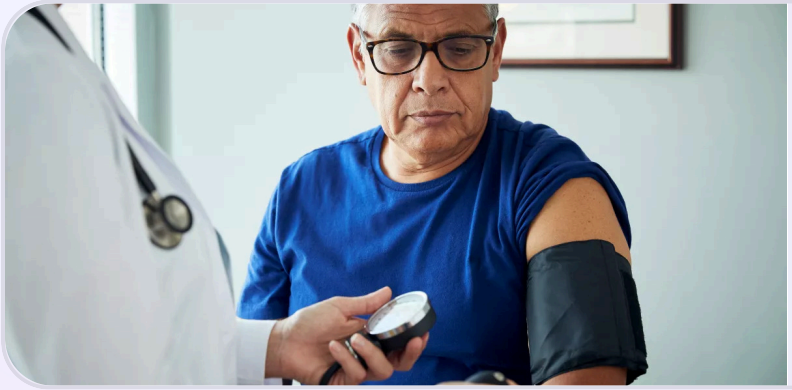
Cardiovascular Statistical Reference



World Heart Observatory

Global Cardiovascular Trends | Insights from the World Heart Obser...

In 2019, nearly 400 million years of healthy life (quantified as disability-adjusted life years or DALYs) were lost due to cardiovascular disease (CVD). This marks ...



Institute for Health Metrics and Evaluation

New report tracks latest trends in global cardiovascular health

Cardiovascular disease (CVD) remains the leading cause of death across the globe, according to a new "almanac"-style special issue of the Journal of the...

Data Science Lifecycle

1. Business Problem Understanding

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for an estimated 18 million deaths each year (WHO).

2. Data Collection

[Medical Examination Dataset Analysis]

3. Data Cleaning and Preparation

Checking the nulls, duplicates, outliers, ...etc.

4. Exploratory Data Analysis

Data visualization and analyzing data

5. Feature Engineering

Extracting new features; BMI, PP,..etc.

6. Machine Learning

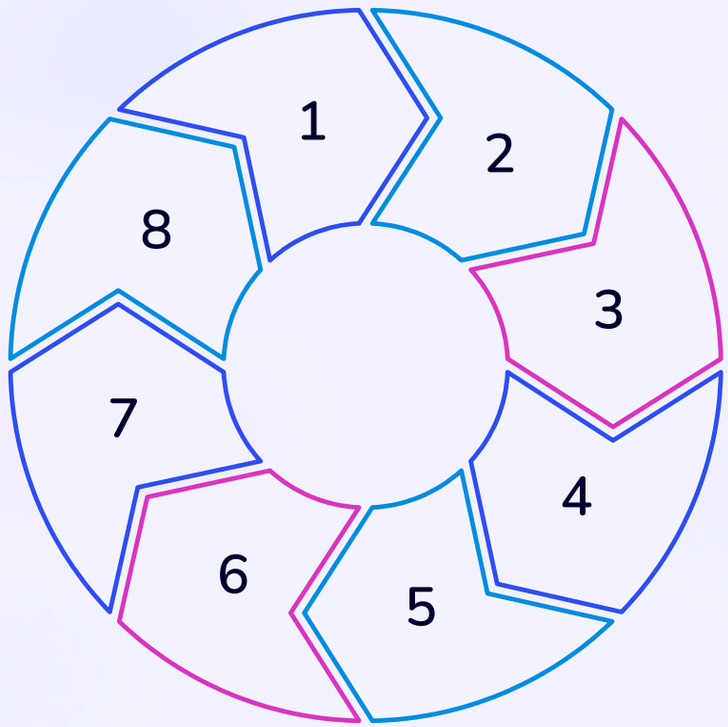
Choosing suitable model that fits our data

7. Model Evaluation

Applying different metrics; precision, recall,...etc

8. Deployment and Monitoring

Using Streamlit & MLflow



Data Collection and Exploration

Dataset Overview

The dataset consists of 70,000 records with 13 factors including age, gender, height, weight, blood pressure readings, cholesterol levels, glucose levels, lifestyle factors, and cardiovascular disease status.

Initial Analysis

Initial checks revealed no duplicates or null values in the dataset. Age was converted from days to years for better interpretation, and outliers were identified in height, weight, and blood pressure readings.

Data Cleaning

Outliers were handled by establishing medical thresholds: height (100–200 cm), weight (45–190 kg), systolic pressure (50–250 mmHg), and diastolic pressure (40–150 mmHg).

Additional Handling

Dynamic Imputation Strategy for Diastolic Pressure Outliers

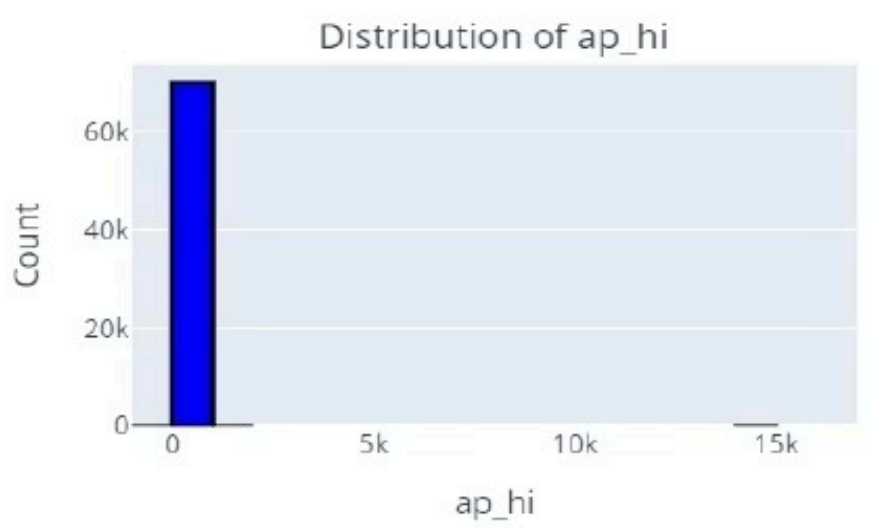
- Thresholds of **ap_lo < 40 mmHg** or **ap_lo > 150 mmHg**.
- Assumed these cases have **cardio = 1.**
- **Subset of 837 records** matched this criteria.
- Calculated the **mean diastolic pressure** of this subset.
- Imputed outlier values with this mean to ensure data consistency.

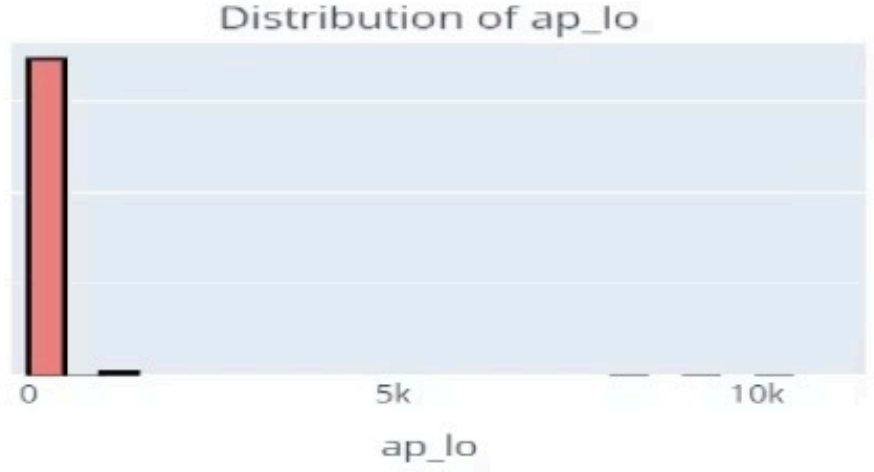
Handling Inconsistent and Rare Records

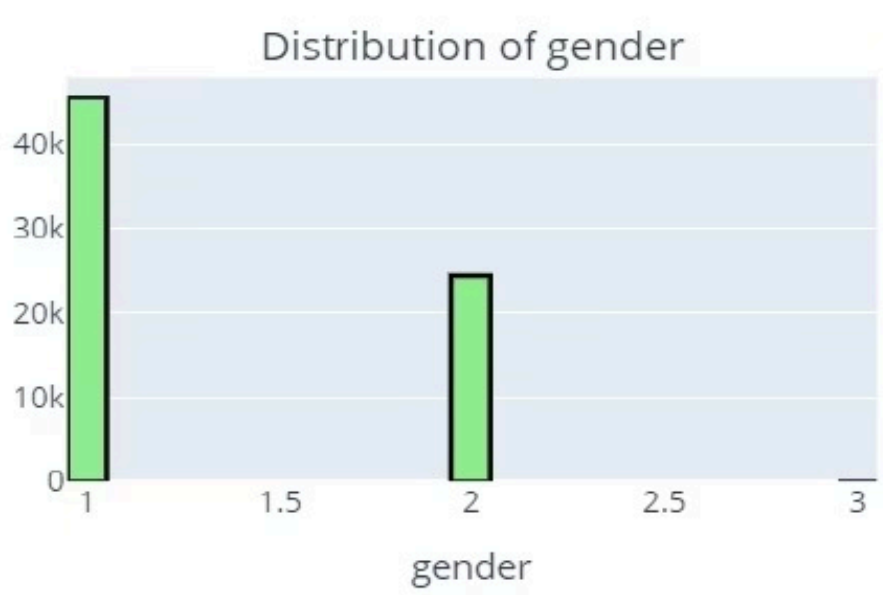
- **Removed 88 records** where **(ap_lo) > (ap_hi)**.
- **Removed 11 records** with **gender = 3**

Before Vs After Handling

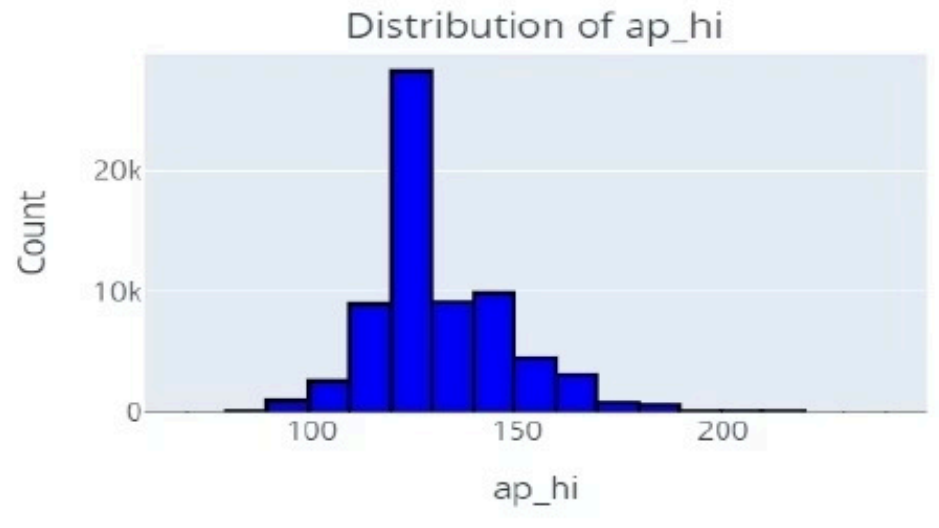
Before

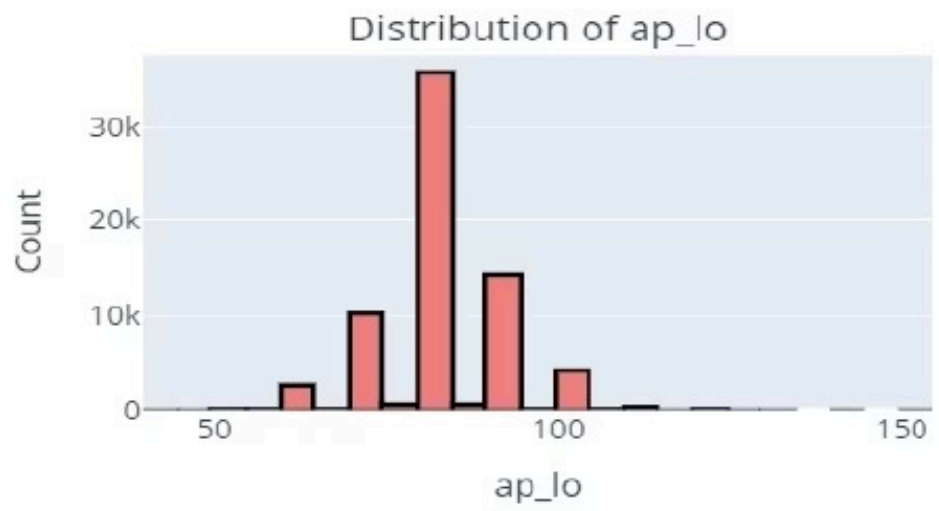


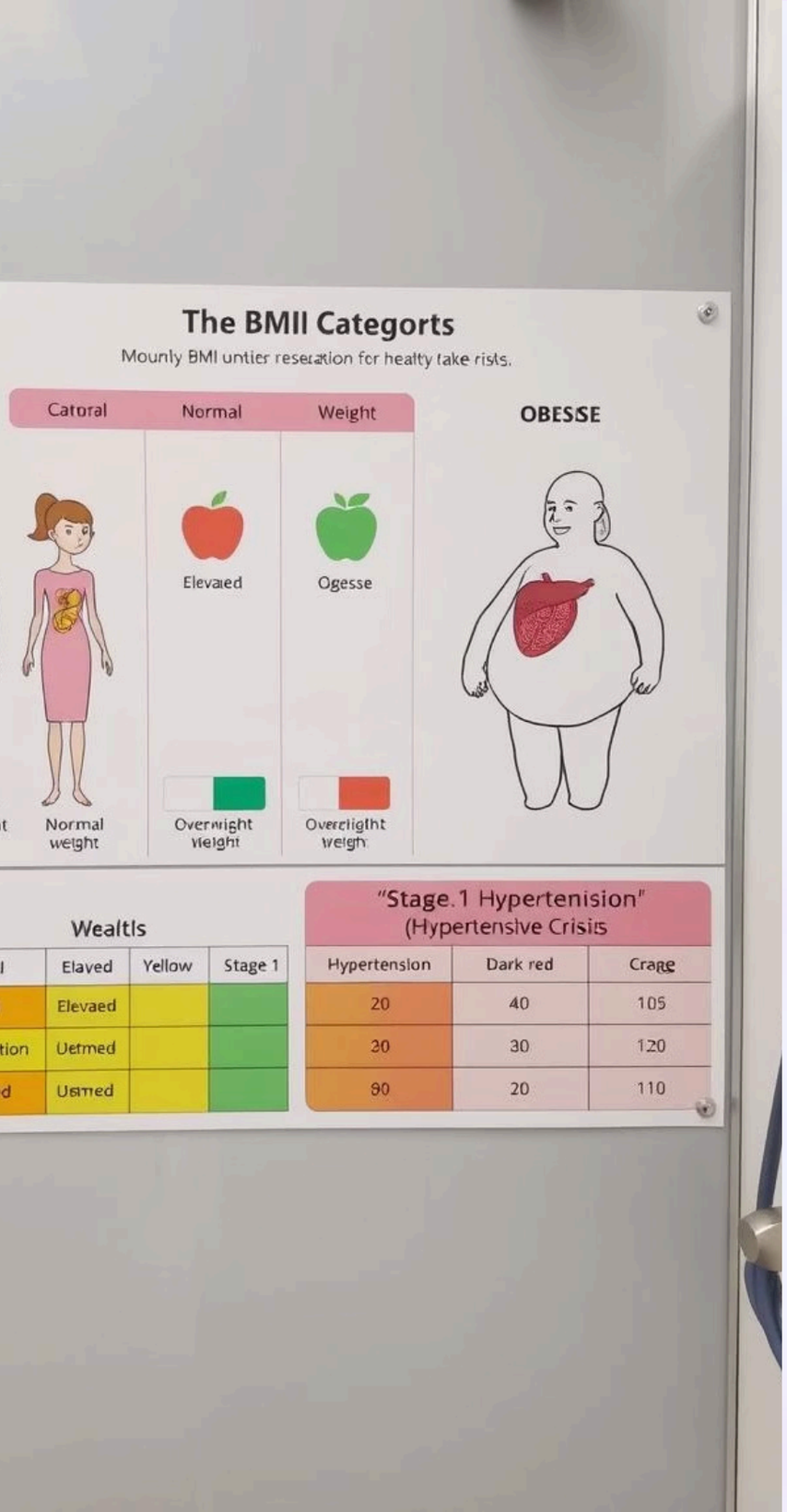




After







Feature Engineering

Body Mass Index (BMI)

BMI was calculated using weight and height, then categorized into: Underweight (<18.5), Normal weight (18.5-24.9), Overweight (25-29.9), and Obesity (≥30).

Hypertension Feature

A binary hypertension flag was created, set to 1 if systolic pressure >130mmHg or diastolic pressure >90mmHg, indicating stage 1 hypertension or higher.

Pulse Pressure & MAP

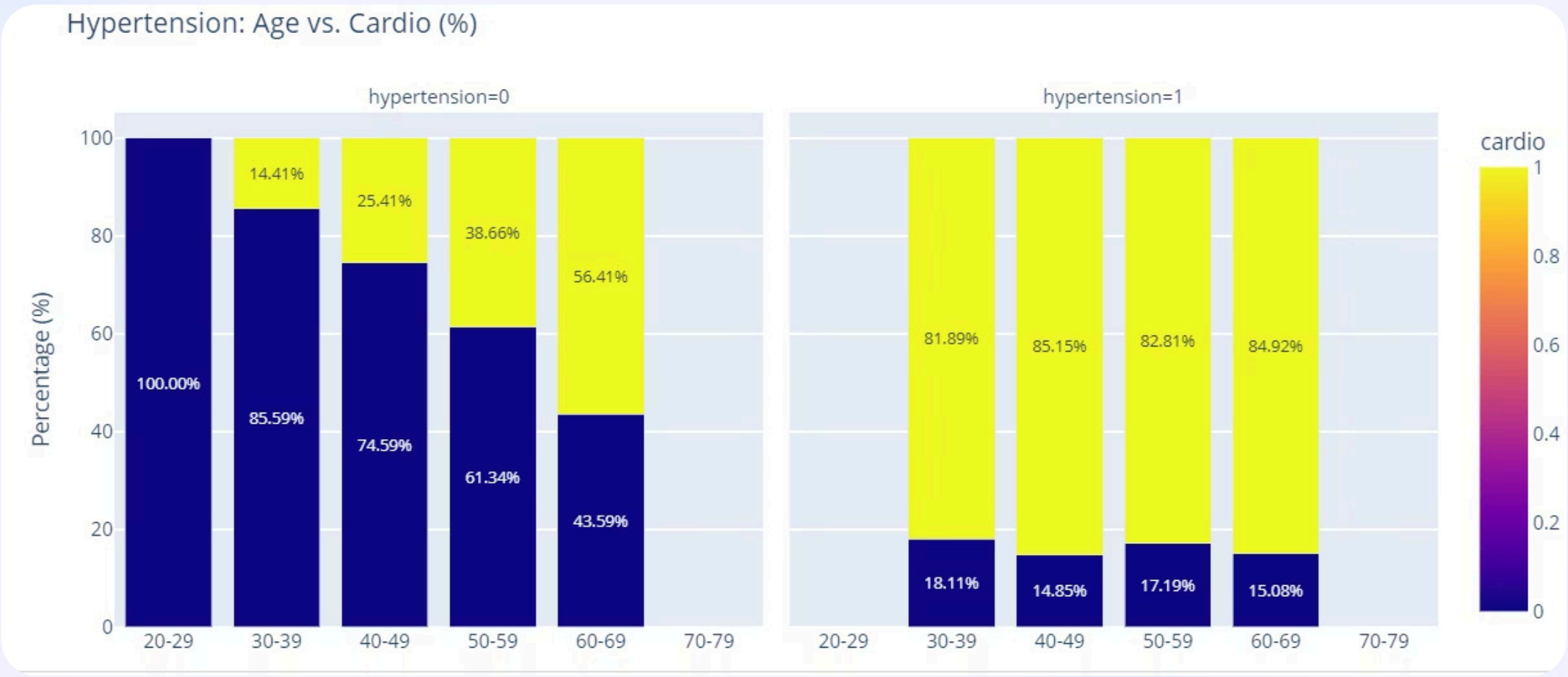
Pulse Pressure (PP = systolic - diastolic) and Mean Arterial Pressure (MAP = diastolic + 1/3(systolic - diastolic)) were calculated as key cardiovascular indicators.

Age Groups & PP Categories

Age was categorized into 6 groups (20-29, 30-39, etc.) and PP into 5 categories based on medical significance, from very low to very high pressure.

Exploratory Data Analysis

Data visualization revealed strong relationships between cardiovascular disease and several factors. BMI category and age group showed direct proportional relationships with CVD risk - people with higher BMI and in older age groups had significantly higher rates of cardiovascular disease.



Similarly, hypertension strongly correlated with CVD, especially in older age groups. Statistical tests confirmed dependencies between the target variable (cardio) and features like cholesterol, BMI category, hypertension, pulse pressure category, and age group (p-values ≈ 0).

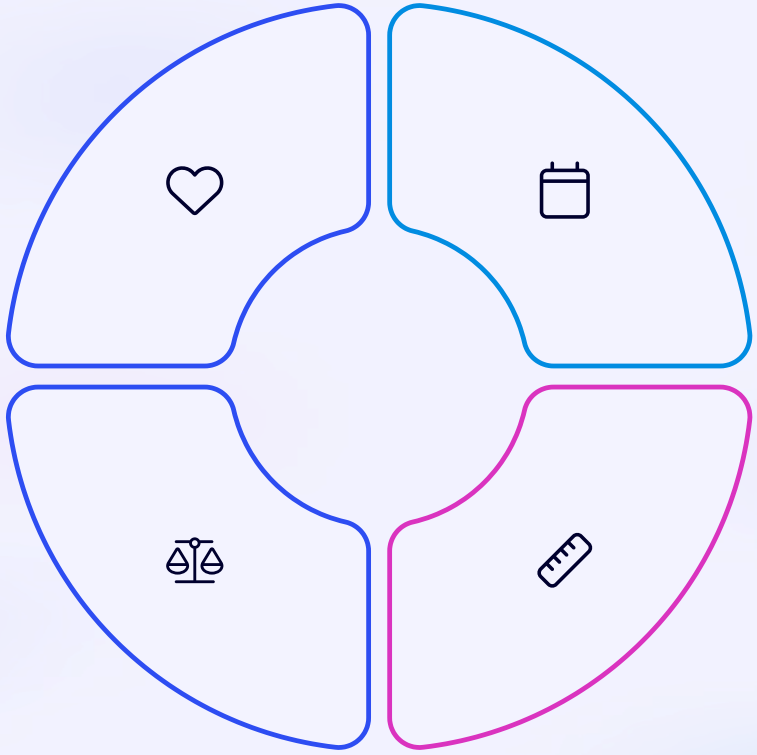
Feature Dependency Analysis

Strongest Correlations

Systolic pressure (ap_hi), Mean Arterial Pressure (MAP), Pulse Pressure (PP), and diastolic pressure (ap_lo) showed the strongest correlations with cardiovascular disease.

Target Balance

The target variable (cardio) was well-balanced in the dataset, allowing for effective model training without class imbalance issues.Both of them have a percentage 49.8%



Moderate Correlations

Age, BMI, and weight demonstrated moderate correlations with cardiovascular disease risk.

Weak Correlations

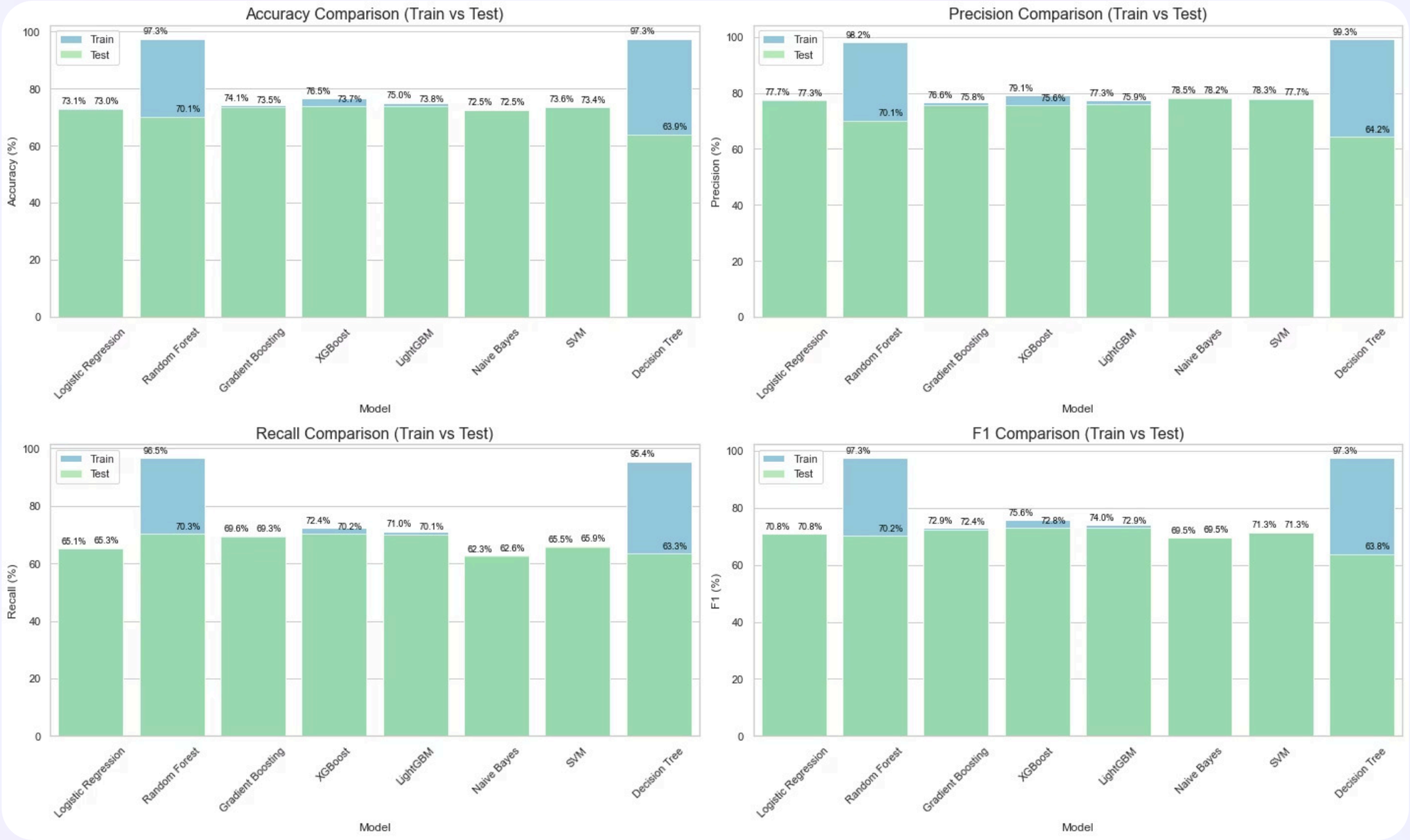
Height showed minimal correlation with cardiovascular disease, while gender had no statistically significant relationship (p-value > 0.05).

Dash Board Demo

Google Docs

DataVisualization.mp4

Model Selection and Evaluation



Model Selection and Evaluation

Trails Summary:

- Trial 1—> compared among different binary classification models and filtered on Light GBM model. (all features used)
- Trial 2—> Used 7 top features with the same models but all models are overfitted.
- Trial 3—> Used voting clf model but it is overfitted.
- Trial 4—> Tried to drop feature by feature and calculate different metrics (Precision, Recall, F1-Score & Accuracy).

Conclusion

- The most scores happended with model Light GBM without alchohol feature.
- By using Random search to get the best parameters of Light GBM the results are attached in next slide.

Model Selection and Evaluation

74%

Best Model Accuracy

LightGBM model without the alcohol consumption feature

0.76

Precision for CVD

Ability to correctly identify positive cases

0.70

Recall for CVD

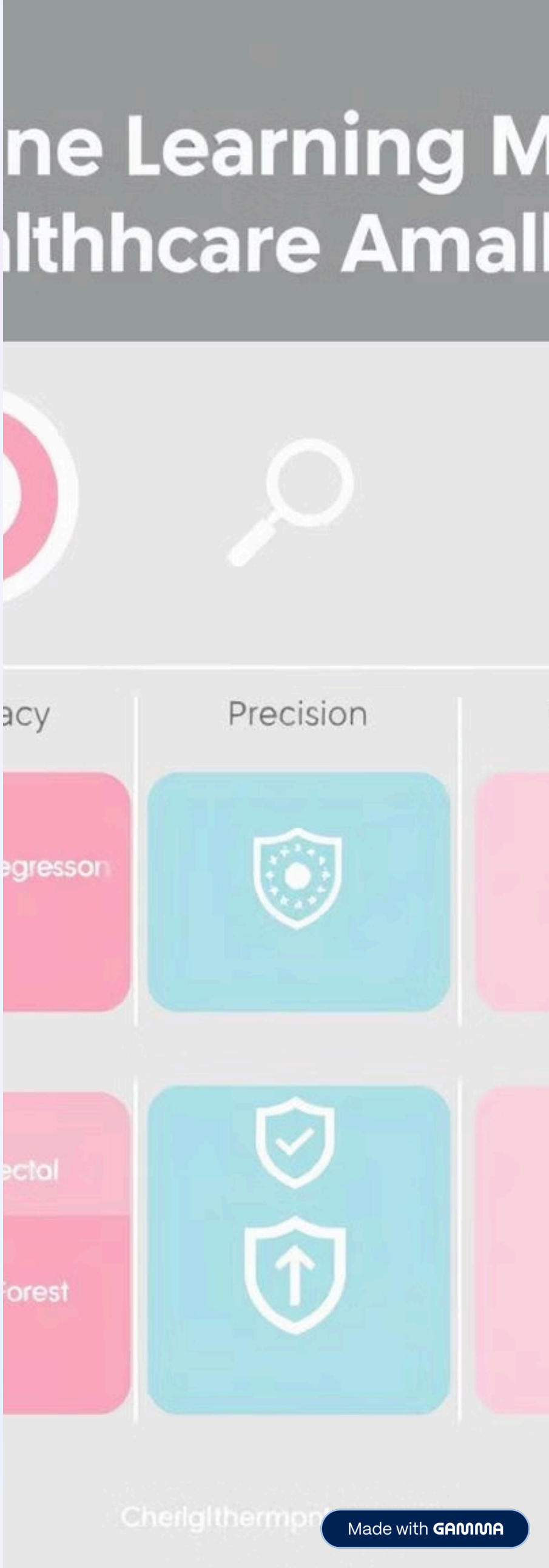
Proportion of actual positives correctly identified

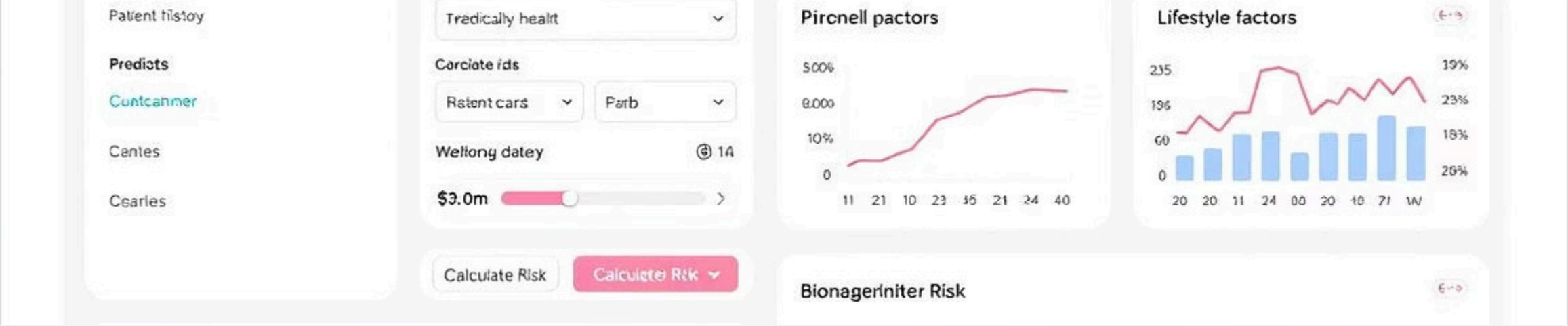
0.73

F1 Score

Harmonic mean of precision and recall

Multiple classification models were evaluated including Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM, Naive Bayes, SVM, and Decision Tree. After extensive testing and parameter tuning using RandomizedSearchCV, the LightGBM model without the alcohol consumption feature was selected as the best performer with an F1 score of 0.73.

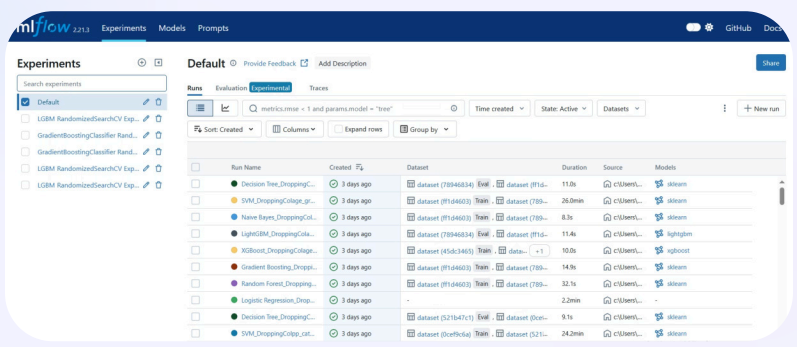




Model Tracking and Deployment

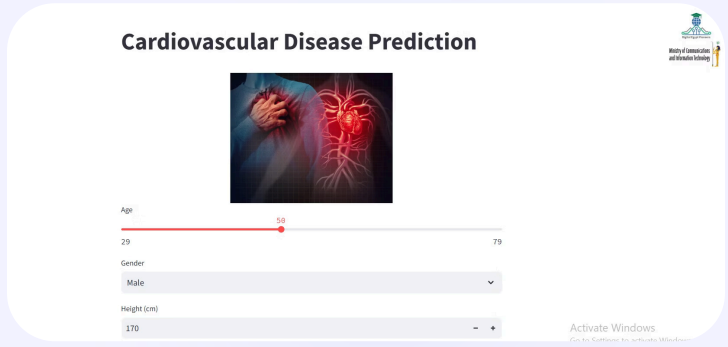
MLflow Integration

MLflow was implemented to track model performance, parameters, and metrics across all experiments, enabling better reproducibility and comparison.



Streamlit Application

A user-friendly web interface was created using Streamlit, allowing healthcare professionals to input patient data and receive cardiovascular risk predictions.



Model Serialization

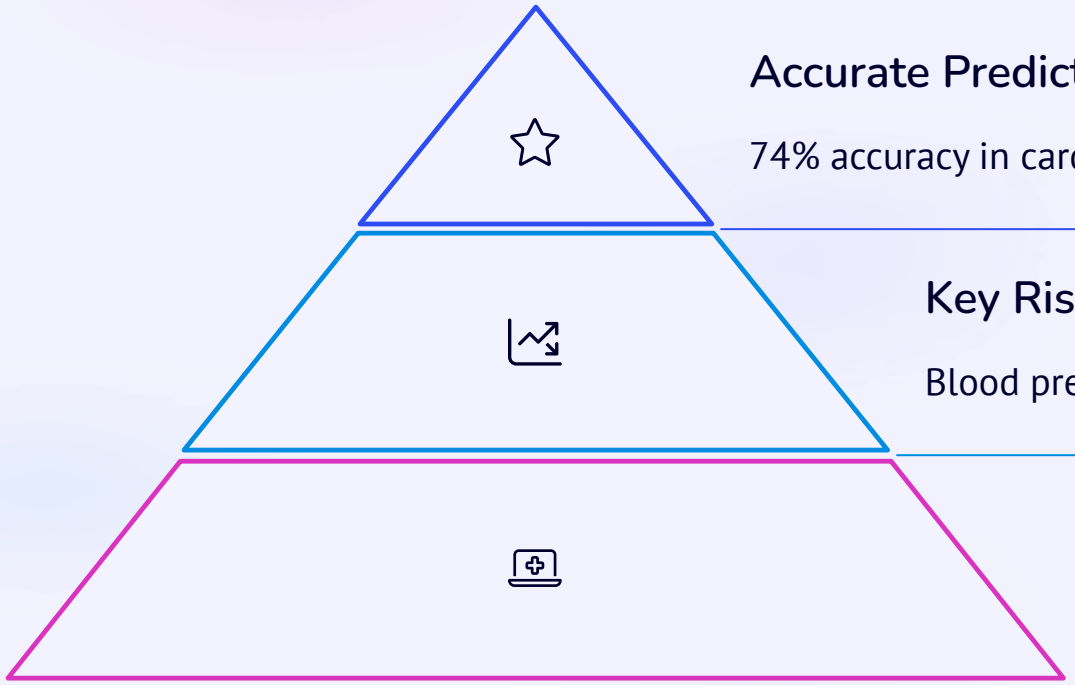
The best performing LightGBM model and standardization parameters were serialized as PKL files for deployment.

Demployment Demo

Google Docs

Deployment vedio.mp4

Conclusion and Future Work



Accurate Prediction

74% accuracy in cardiovascular disease risk assessment

Key Risk Factors

Blood pressure, age, and BMI identified as strongest predictors

Practical Application

Deployed as an accessible tool for healthcare professionals

The HealthCare Predictive Analysis project successfully developed a model that can identify cardiovascular disease risk with 74% accuracy. Blood pressure measurements (systolic, diastolic, MAP, and pulse pressure) emerged as the strongest predictors, followed by age and BMI.

Future work could include incorporating additional biomarkers, expanding the dataset with more diverse populations, and developing more sophisticated ensemble models to further improve prediction accuracy. The current implementation provides a solid foundation for clinical decision support in cardiovascular risk assessment.