Empowering Health for a Brighter Future

Cardiovascular Disease Prediction: A Data Science Approach

reallygreatsite.com

Aldenaire & Partners

our team

Mariam Ibrahim faculty of computers & data science

Mariam Tamer faculty of computers & data science

perihane tarek

Mahmoud Abdelrazek

Adham Zakaria

Mohamed Negm

Background and Motivation

Cardiovascular disease encompasses a range of conditions affecting the heart and blood vessels, including coronary artery disease, hypertension, and heart failure. Risk factors for CVD include both modifiable elements (such as physical activity, smoking, and alcohol consumption) and non-modifiable factors (such as age and genetic predisposition). Traditional risk assessment methods often rely on clinical judgment and standardized scoring systems, which may not capture the complex, non-linear relationships between multiple risk factors.

Machine learning offers a powerful alternative by identifying patterns within large-scale health data that may be imperceptible to human analysis. By training models on historical patient data, we can develop predictive tools that support healthcare providers in identifying at-risk individuals earlier and more accurately.

Project Objectives

The primary objectives of this project are:

1. **Data Preparation: Clean and preprocess cardiovascular health data to ensure quality and reliability for subsequent analysis**
2. **Exploratory Analysis: Visualize and understand the relationships between various health indicators and cardiovascular disease prevalence**
3. **Predictive Modeling: Develop and evaluate machine learning models capable of predicting cardiovascular disease presence based on patient characteristics**
4. **Model Deployment: Create an accessible interface for model predictions through API development**

1.3 Dataset Description

The dataset comprises 70,000 patient records collected during medical examinations, with 11 input features and one binary target variable indicating the presence or absence of cardiovascular disease. Features are categorized into three types:

Objective Features (factual information):

Age (originally in days, converted to years)
Height (cm)
Weight (kg)
Gender (categorical)

Examination Features (medical measurements):

Systolic blood pressure (ap_hi)
Diastolic blood pressure (ap_lo)
Cholesterol level (1: normal, 2: above normal, 3: well above normal)
Glucose level (1: normal, 2: above normal, 3: well above normal)

Target Variable:

Cardiovascular disease presence (binary: 0 = absent, 1 = present)

Subjective Features (patient-reported):

Smoking status (binary)
Alcohol intake (binary)
Physical activity (binary)

All measurements were recorded at a single point in time during medical examination, providing a cross-sectional view of patient health status.

2. Phase 1: Data Preprocessing and Cleaning

Data quality is fundamental to the success of any machine learning project. Raw healthcare data often contains inconsistencies, missing values, and physiologically impossible measurements that can compromise model performance. This phase establishes a rigorous data cleaning pipeline to ensure the dataset is accurate, complete, and ready for analysis.

2.1 Data Loading and Initial Exploration

The dataset was obtained as a CSV file (cardio_train.csv) with semicolon delimiters. Initial exploration using pandas revealed the dataset structure, data types, and basic statistics:

```
import pandas as pd
df = pd.read_csv('cardio_train.csv', sep=';')
```

Preliminary examination using df.head(), df.info(), and df.describe() confirmed 70,000 records across 12 columns with no immediately apparent missing values. The dataset was reformatted and saved as cardio_train_separated.csv using standard comma delimiters to ensure compatibility with various analytical tools.

2.2 Data Quality Assessment

Missing Value Analysis: A comprehensive check using df.isnull().sum() confirmed zero missing values across all columns, eliminating the need for imputation strategies.
Column Standardization: To enhance code readability and maintain consistency with medical terminology, several columns were renamed:

- gluc → glucose
- alco → alcohol
- active → physically_active

Identifier Removal: The id column, serving solely as a record identifier without analytical value, was removed from the dataset.

2.3 Feature Transformation

Age Conversion: Age was originally recorded in days, which is uncommon in medical contexts and difficult to interpret. We converted this to years for intuitive understanding:

```
df['age_years'] = df['age'] / 365.25
df.drop('age', axis=1, inplace=True)
```

This transformation facilitates more meaningful analysis and aligns with standard medical practice.

2.4 Outlier Detection and Treatment

Outlier detection is particularly critical in medical datasets, as extreme values may represent either genuine biological variation or data entry errors. We employed a two-pronged approach:

Statistical Outlier Detection: The Interquartile Range (IQR) method was applied to continuous variables (height, weight, blood pressure measurements, age) to identify statistical outliers:

Q1 = 25th percentile

Q3 = 75th percentile

IQR = Q3 - Q1

Lower bound = Q1 - 1.5 × IQR

Upper bound = Q3 + 1.5 × IQR

Physiological Validity Filtering: Statistical outliers were not automatically removed, as they comprised a substantial portion of the dataset. Instead, we implemented domain-specific filters to remove physiologically impossible values:

- Blood Pressure: Removed records with systolic BP < 80 or > 200 mmHg, diastolic BP < 40 or > 140 mmHg
- Height: Removed records with height < 140 cm or > 210 cm
- Weight: Removed records with weight < 40 kg or > 200 kg

This selective approach preserved genuine biological variation while eliminating clear data quality issues, balancing data retention with analytical integrity.

2.5 Feature Engineering

To enhance the predictive power of our models, we created several derived features based on domain knowledge in cardiovascular health:

1. Body Mass Index (BMI):

```
df['bmi'] = df['weight'] / ((df['height'] / 100) ** 2)
```

BMI is a standard indicator of body composition and a known CVD risk factor.

2. Pulse Pressure:

```
df['pulse_pressure'] = df['ap_hi'] - df['ap_lo']
```

The difference between systolic and diastolic blood pressure provides information about arterial stiffness and cardiovascular health.

3. Health Index:

```
df['health_index'] = (df['physically_active'] * 1) - (df['smoke'] * 0.5) - (df['alcohol'] * 0.5)
```

A composite score capturing lifestyle factors, where positive behaviors (physical activity) increase the score and negative behaviors (smoking, alcohol) decrease it.

4. Cholesterol-Glucose Interaction:

```
df['cholesterol_gluc_interaction'] = df['cholesterol'] * df['glucose']
```

This interaction term captures the combined effect of metabolic risk factors, as elevated cholesterol and glucose often co-occur in metabolic syndrome.

5. Hypertension Indicator:

```
df['hypertension'] = ((df['ap_hi'] >= 130) | (df['ap_lo'] >= 90)).astype(int)
```

A binary flag based on clinical hypertension thresholds (systolic ≥130 mmHg or diastolic ≥90 mmHg), following current medical guidelines.

2.6 Final Dataset Preparation

Following all cleaning and transformation steps, the processed dataset was saved as cardio_train_new.csv. The final dataset contains:

- Valid, physiologically plausible measurements
- Consistent and meaningful column names
- No missing values or duplicate records
- Enhanced feature set including engineered variables
- Comprehensive documentation of all transformations

This cleaned dataset serves as the foundation for all subsequent analysis and modeling phases, ensuring that insights and predictions are based on high-quality, reliable data.

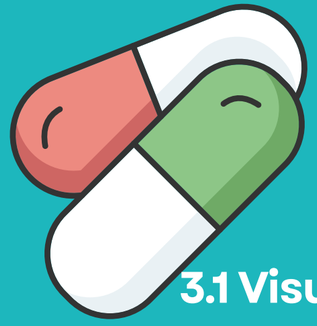
2.6 Final Dataset Preparation

Following all cleaning and transformation steps, the processed dataset was saved as cardio_train_new.csv. The final dataset contains:

- Valid, physiologically plausible measurements
- Consistent and meaningful column names
- No missing values or duplicate records
- Enhanced feature set including engineered variables
- Comprehensive documentation of all transformations

This cleaned dataset serves as the foundation for all subsequent analysis and modeling phases, ensuring that insights and predictions are based on high-quality, reliable data.

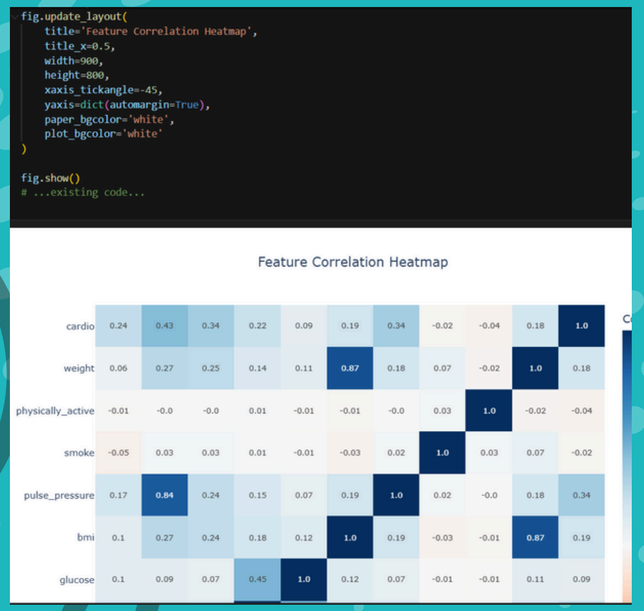
Phase 2: Data Visualization and Exploratory Analysis



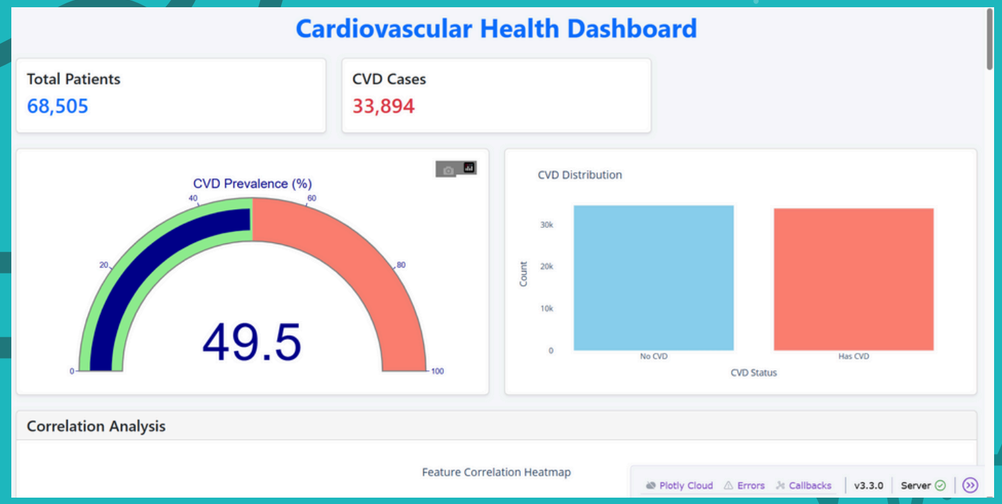
3.1 Visualization Infrastructure

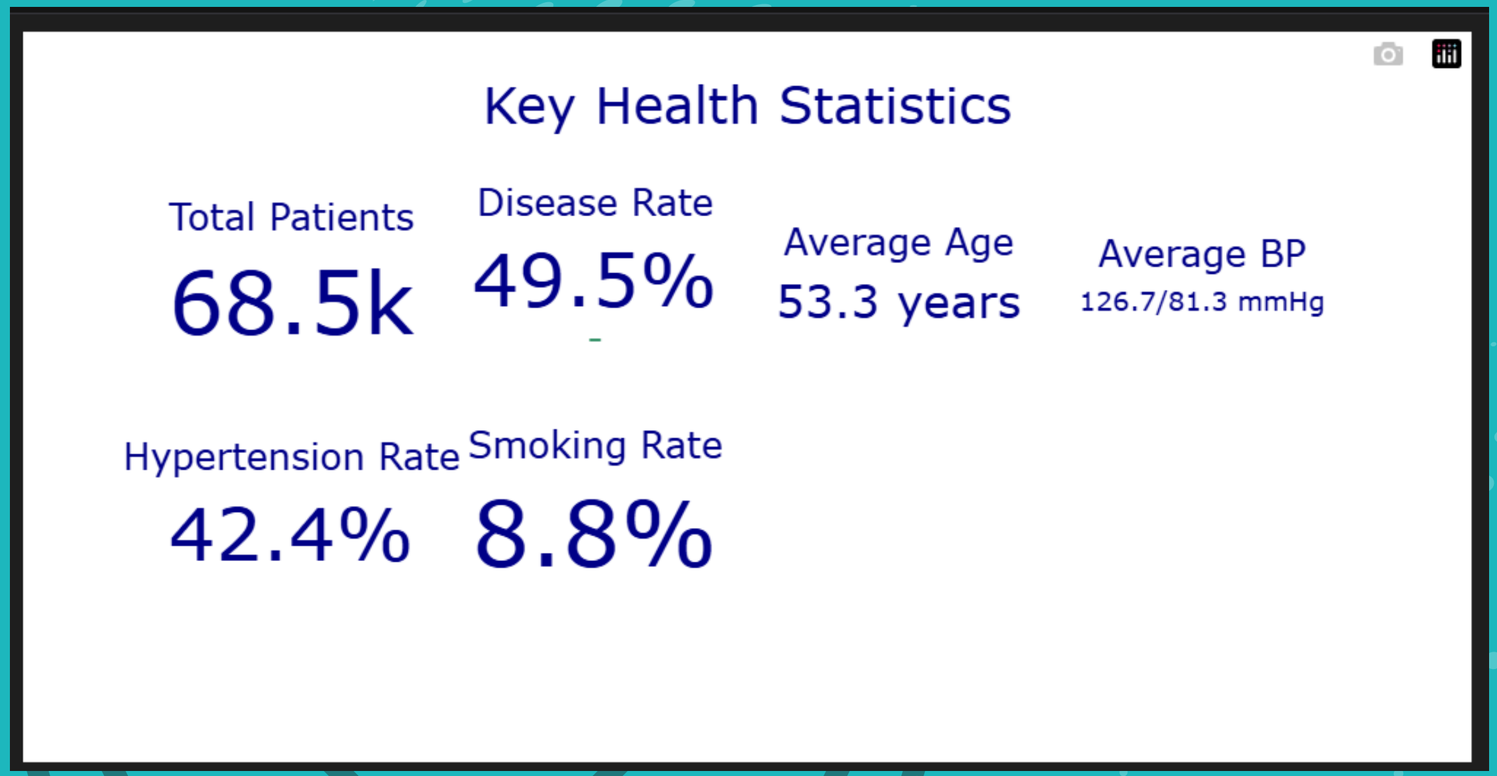
Technical Implementation: To facilitate comprehensive exploratory analysis, we developed a two-tier visualization system:

1. **Documentation Layer: A Jupyter notebook (visualization.ipynb) containing all visualization code, methodology, and static outputs. This serves as a complete record of our analytical process and ensures reproducibility of results.**



2.Interactive Dashboard: A production-ready web application built using Plotly for graphics rendering and Dash for interactivity. This dashboard was deployed to enable dynamic exploration of the dataset without requiring programming knowledge, making insights accessible to clinical stakeholders.





3.2 Dataset Summary Statistics

Before detailed visualization, we established key baseline metrics to characterize the dataset:

Population Overview:

- **Total Patients: 68,500 (after data cleaning)**
- **Disease Prevalence: 49.5% (nearly balanced dataset)**
- **Average Age: 53.3 years**
- **Average Blood Pressure: 126.7/81.3 mmHg**
- **Hypertension Rate: 42.4% (defined as BP ≥130/90 mmHg)**
- **Smoking Rate: 8.8%**

>>

Key Health Statistics

Total Patients	Disease Rate	Average Age	Average BP
68.5k	49.5%	53.3 years	126.7/81.3 mmHg

Hypertension Rate	Smoking Rate
42.4%	8.8%

These summary statistics immediately reveal several important characteristics. The dataset exhibits excellent class balance with approximately equal representation of healthy and CVD-positive patients, eliminating concerns about class imbalance that often plague medical classification tasks. The population skews toward middle-age and older adults, which appropriately reflects the demographic most at risk for cardiovascular disease. The relatively low smoking rate may reflect underreporting, a common issue with self-reported subjective features.

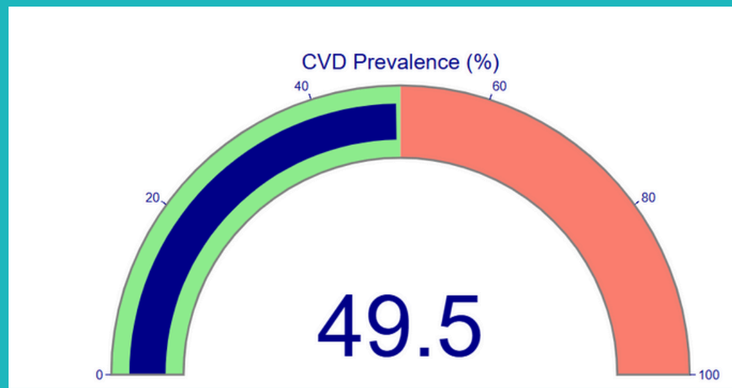
>>

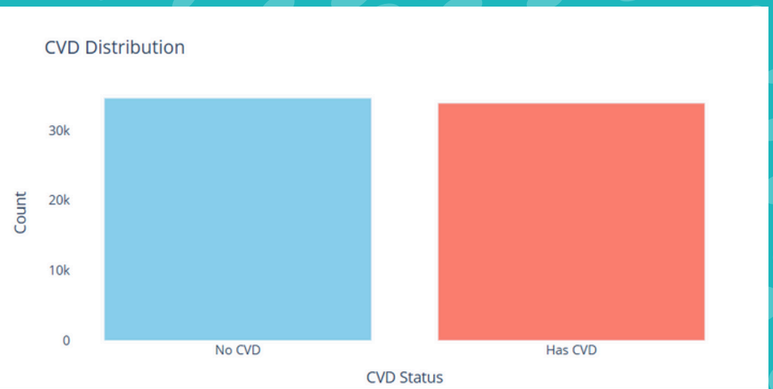
3.3.2 Interactive Variable Distribution Viewer

3.3.2 Interactive Variable Distribution Viewer

The dashboard implements an interactive histogram feature with bullet-point controls, allowing users to dynamically select any numerical variable (age, blood pressure measurements, BMI, weight, height, pulse pressure) and view its distribution instantly. This interface revealed:

- **Blood Pressure Variables: Both systolic and diastolic pressures show approximately normal distributions centered at clinical thresholds, with substantial right-skew toward hypertensive ranges**
- **BMI Distribution: Right-skewed distribution with modal values in the overweight category (25-30 kg/m²)**
- **Pulse Pressure: Near-normal distribution centered around 40 mmHg, consistent with healthy arterial compliance**

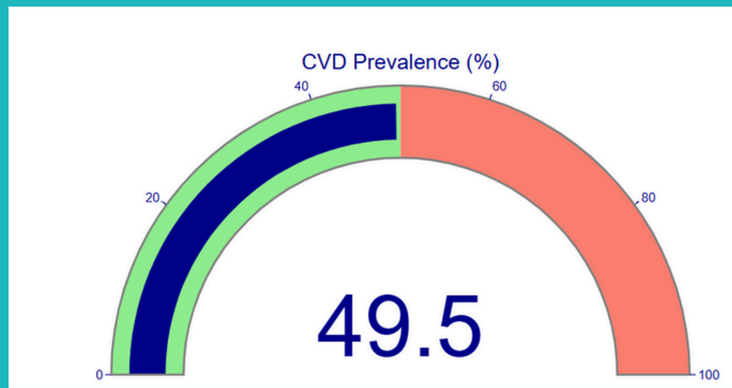


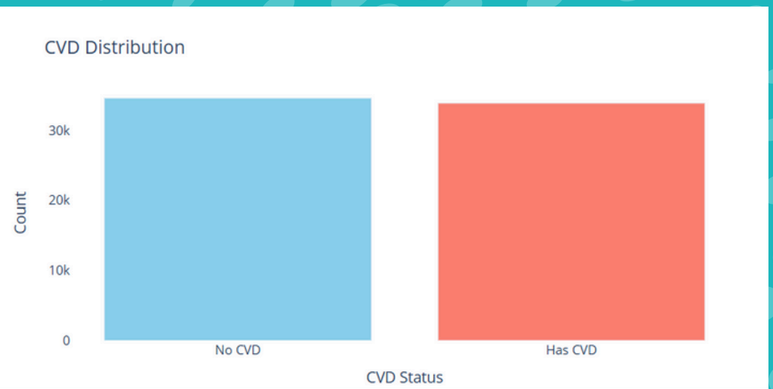


3.4 Target Variable Analysis

Disease Prevalence Assessment: A bar plot visualization confirms the dataset's class balance, with 49.5% of patients having cardiovascular disease and 50.5% classified as healthy. Additionally, a gauge chart provides an intuitive visual representation of disease prevalence, immediately communicating the near-perfect balance to stakeholders.

Implication for Modeling: The balanced nature of the target variable is highly advantageous for machine learning. It eliminates the need for resampling techniques (SMOTE, undersampling) or class-weight adjustments, and ensures that accuracy is a meaningful performance metric alongside precision and recall.

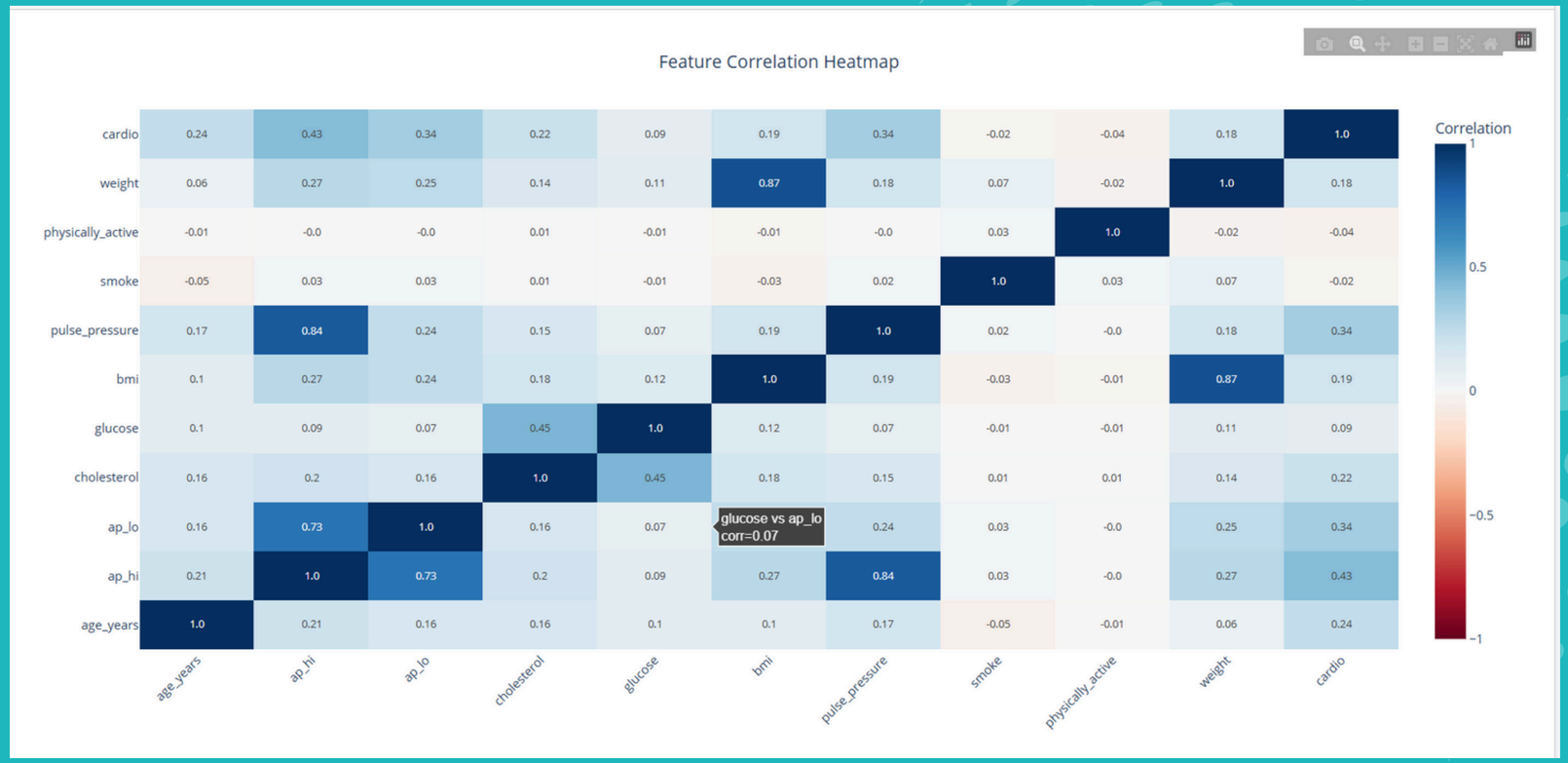




3.4 Target Variable Analysis

Disease Prevalence Assessment: A bar plot visualization confirms the dataset's class balance, with 49.5% of patients having cardiovascular disease and 50.5% classified as healthy. Additionally, a gauge chart provides an intuitive visual representation of disease prevalence, immediately communicating the near-perfect balance to stakeholders.

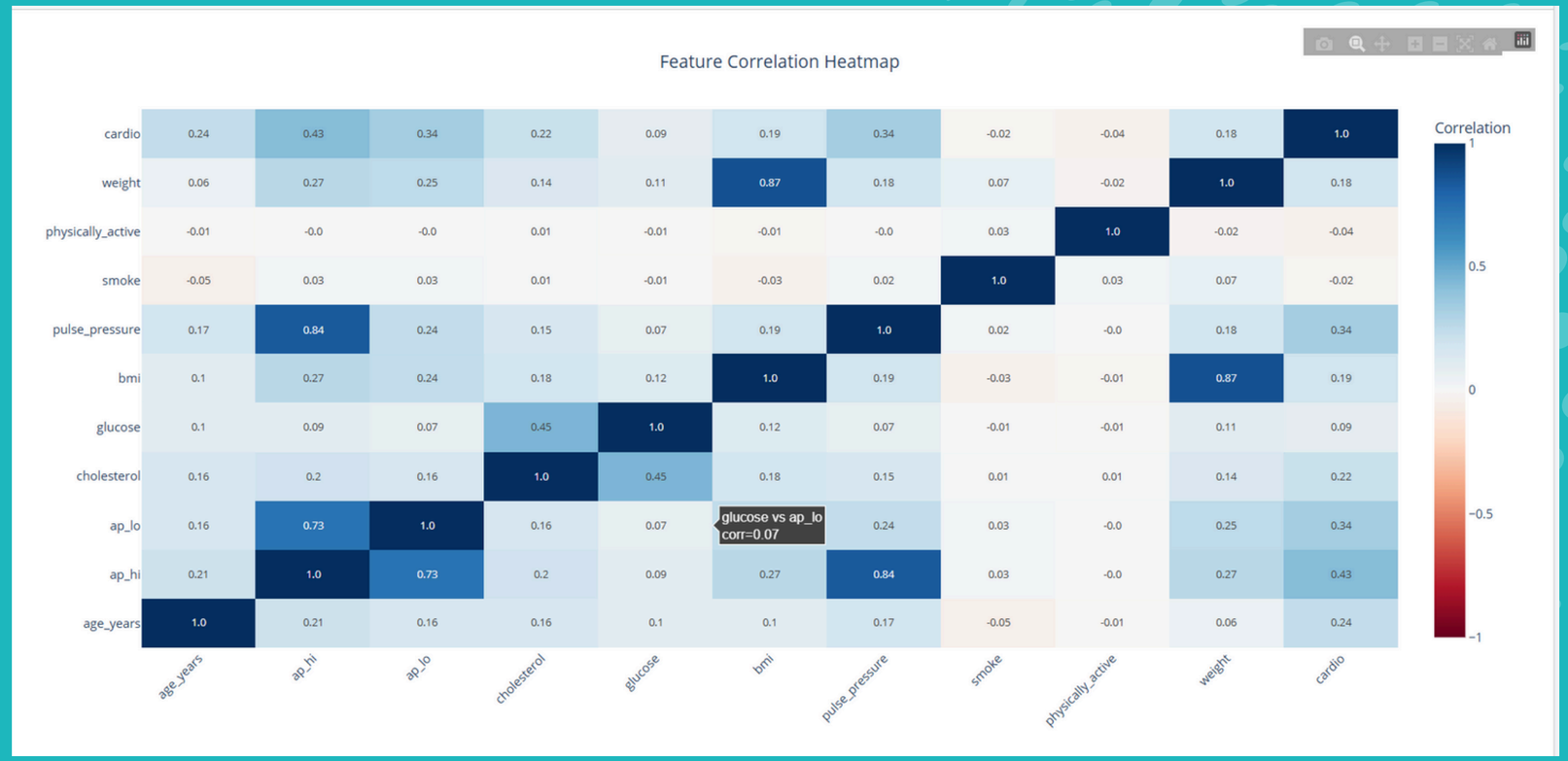
Implication for Modeling: The balanced nature of the target variable is highly advantageous for machine learning. It eliminates the need for resampling techniques (SMOTE, undersampling) or class-weight adjustments, and ensures that accuracy is a meaningful performance metric alongside precision and recall.



3.5 Correlation Analysis

A comprehensive correlation matrix heatmap was generated to quantify linear relationships between all numerical features and the target variable (CVD). The analysis revealed several critical findings:

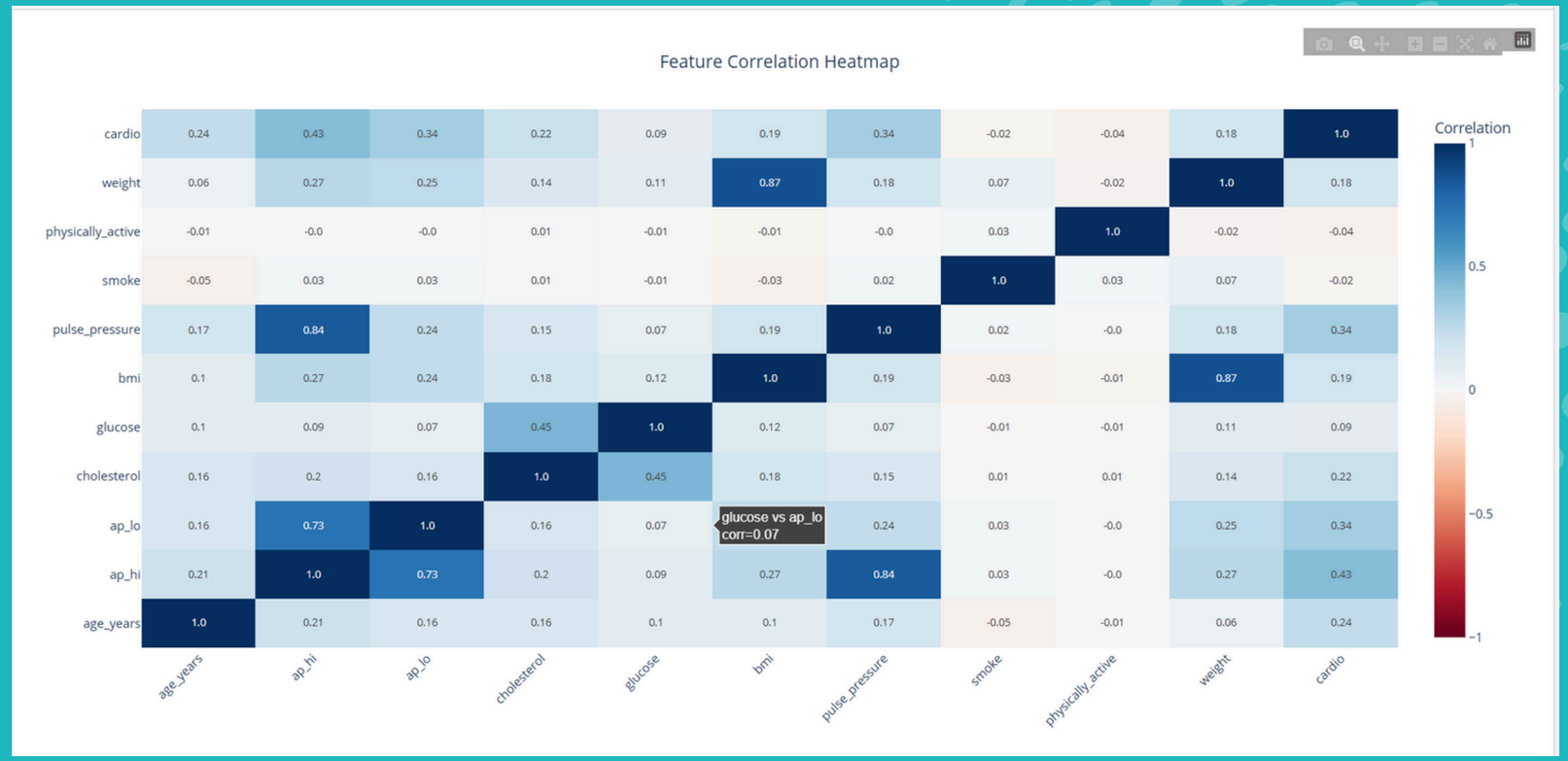
3.5 Correlation Analysis



Strong Positive Correlations with CVD:

- **Systolic Blood Pressure (ap_hi): r = 0.43 – strongest individual predictor**
- **Pulse Pressure: r = 0.34 – strong correlation reflecting arterial stiffness**
- **Cholesterol Level: r = 0.34 – metabolic risk factor confirmation**
- **Age: r = 0.24 – expected age-related disease progression**

3.5 Correlation Analysis



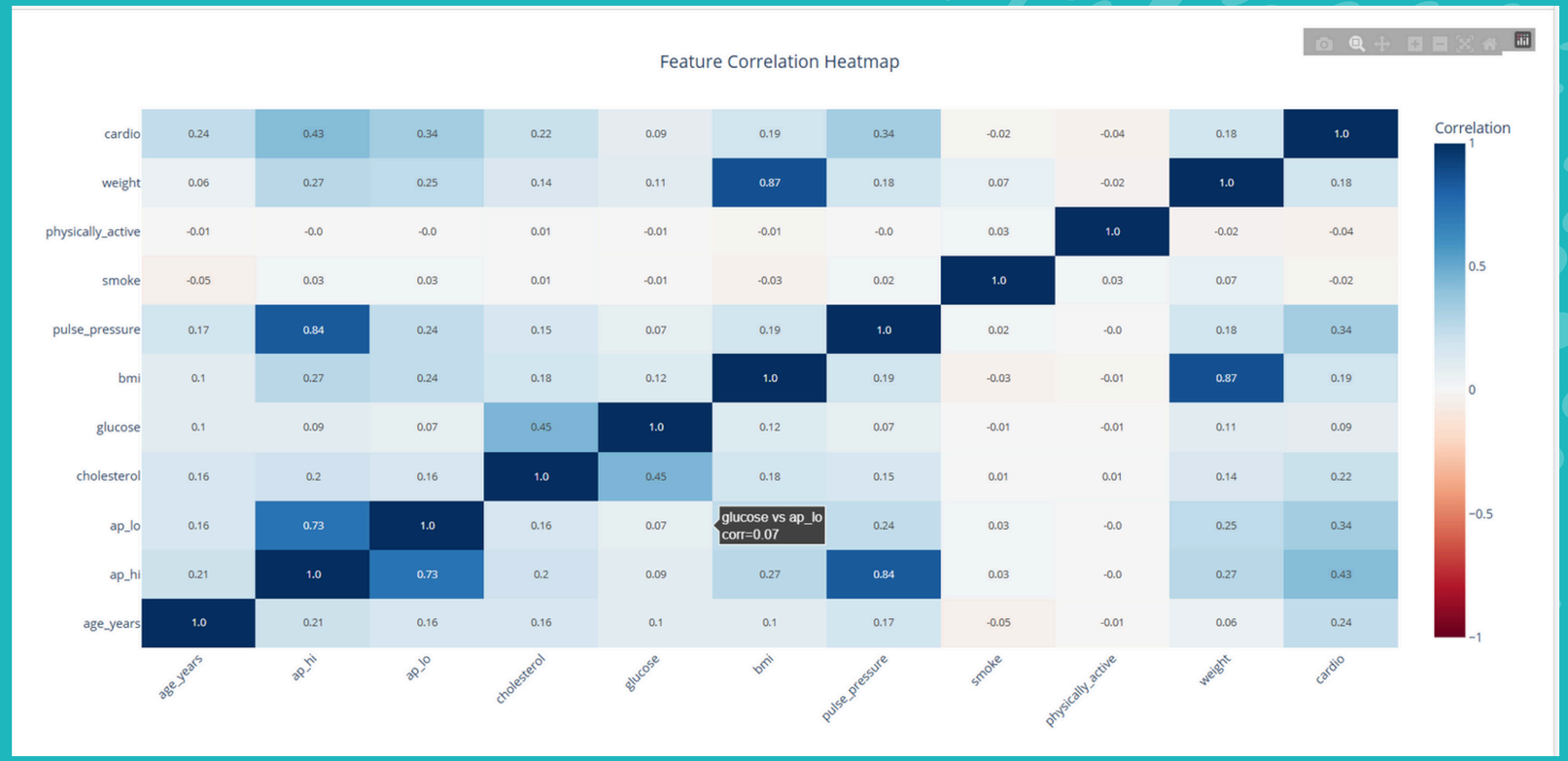
Moderate Positive Correlations:

- **BMI**: r = 0.19 – obesity's role in CVD risk
- **Diastolic Blood Pressure (ap_lo)**: r = 0.34
- **Weight**: r = 0.18

Weak or Negligible Correlations:

- **Physical Activity**: r ≈ -0.04 – surprisingly weak protective effect
- **Smoking**: r ≈ -0.02 – counterintuitively weak, suggesting data quality issues
- **Alcohol Consumption**: near-zero correlation

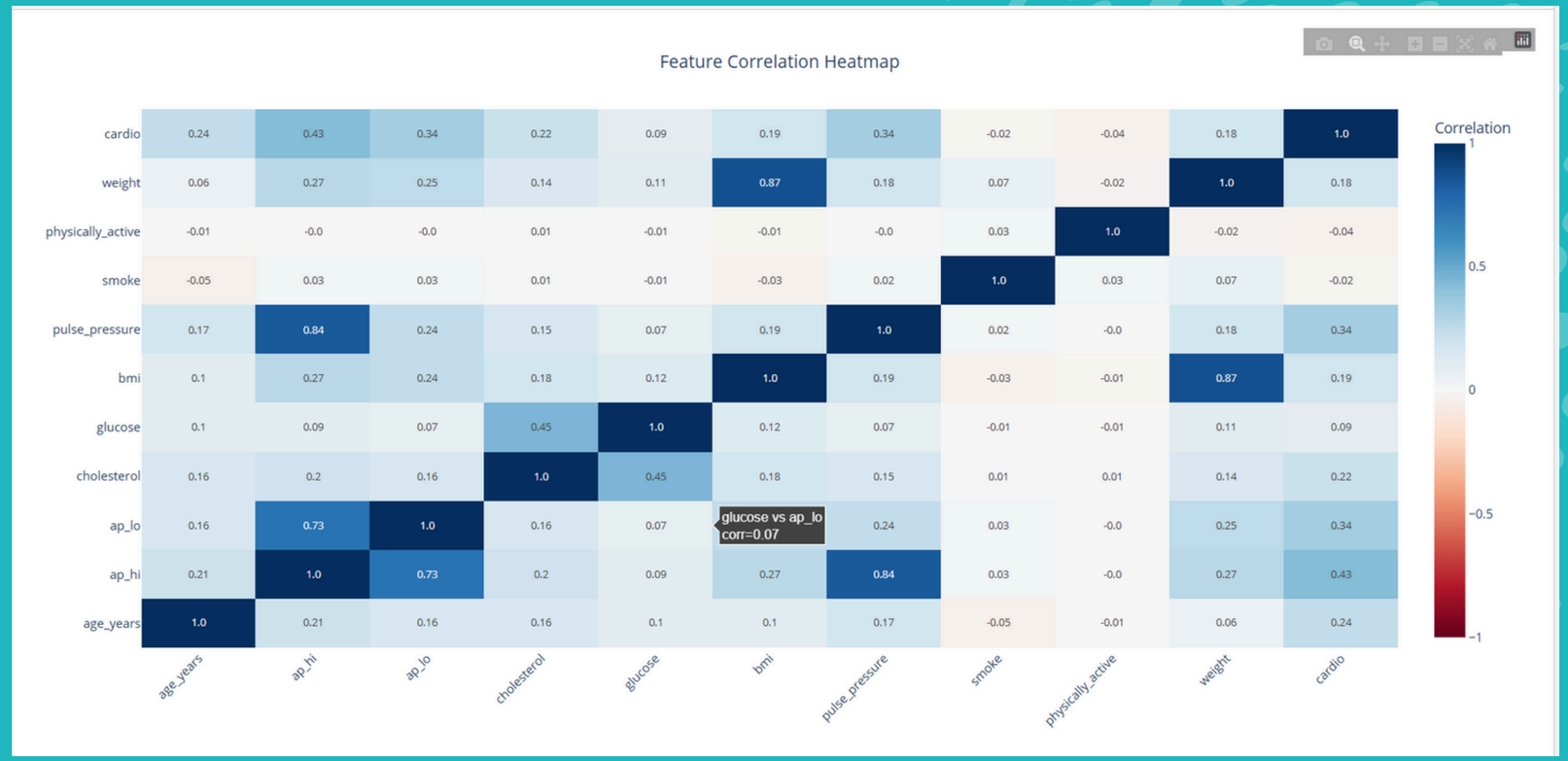
3.5 Correlation Analysis



Multicollinearity Detection: The correlation matrix also revealed expected intercorrelations:

- **Systolic and diastolic blood pressure: r = 0.73 (high, but both retained for clinical interpretability)**
- **BMI and weight: r = 0.87 (very high, suggesting potential redundancy)**
- **Pulse pressure and systolic BP: r = 0.84 (high due to mathematical relationship)**

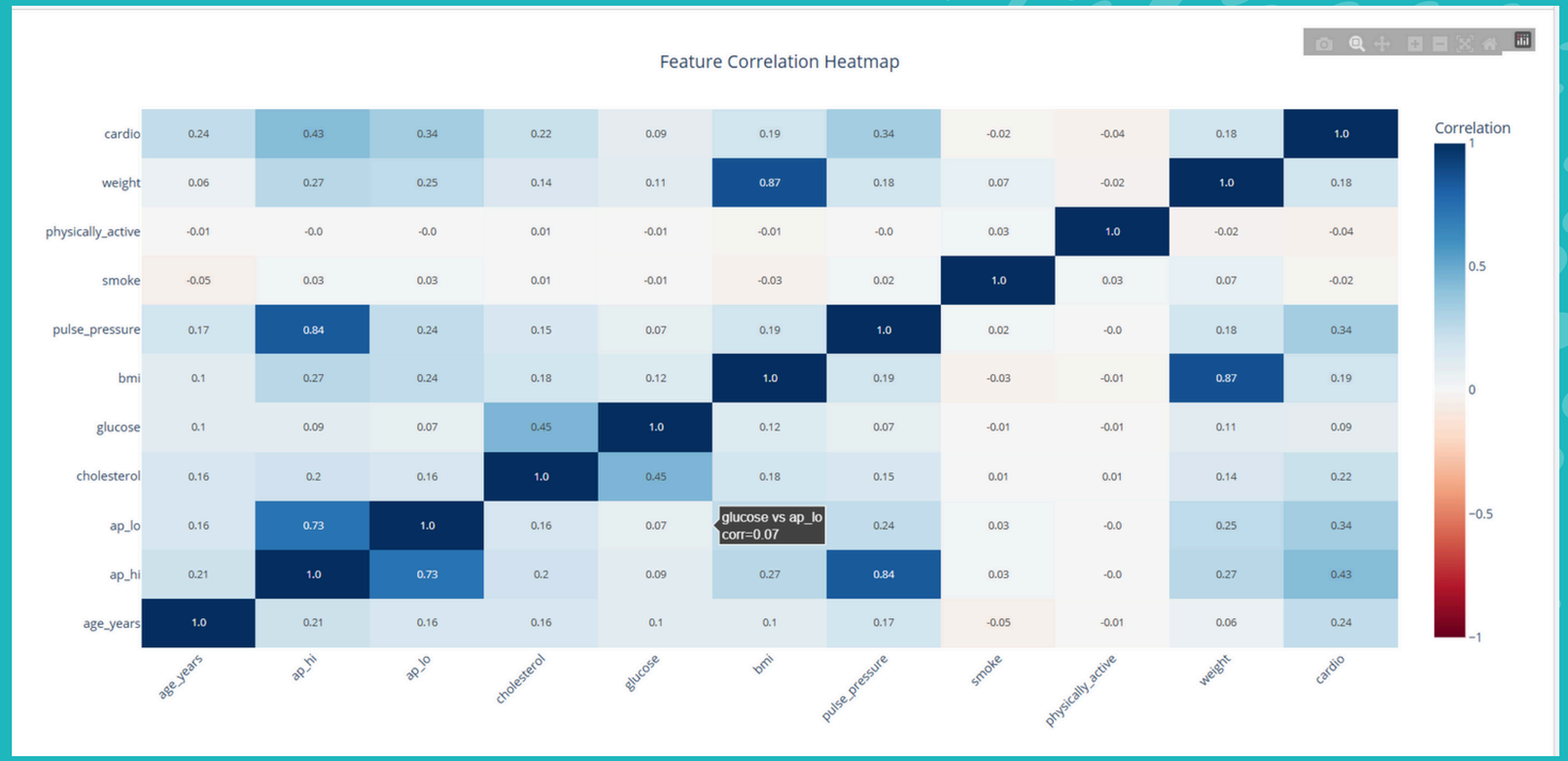
3.5 Correlation Analysis



Multicollinearity Detection: The correlation matrix also revealed expected intercorrelations:

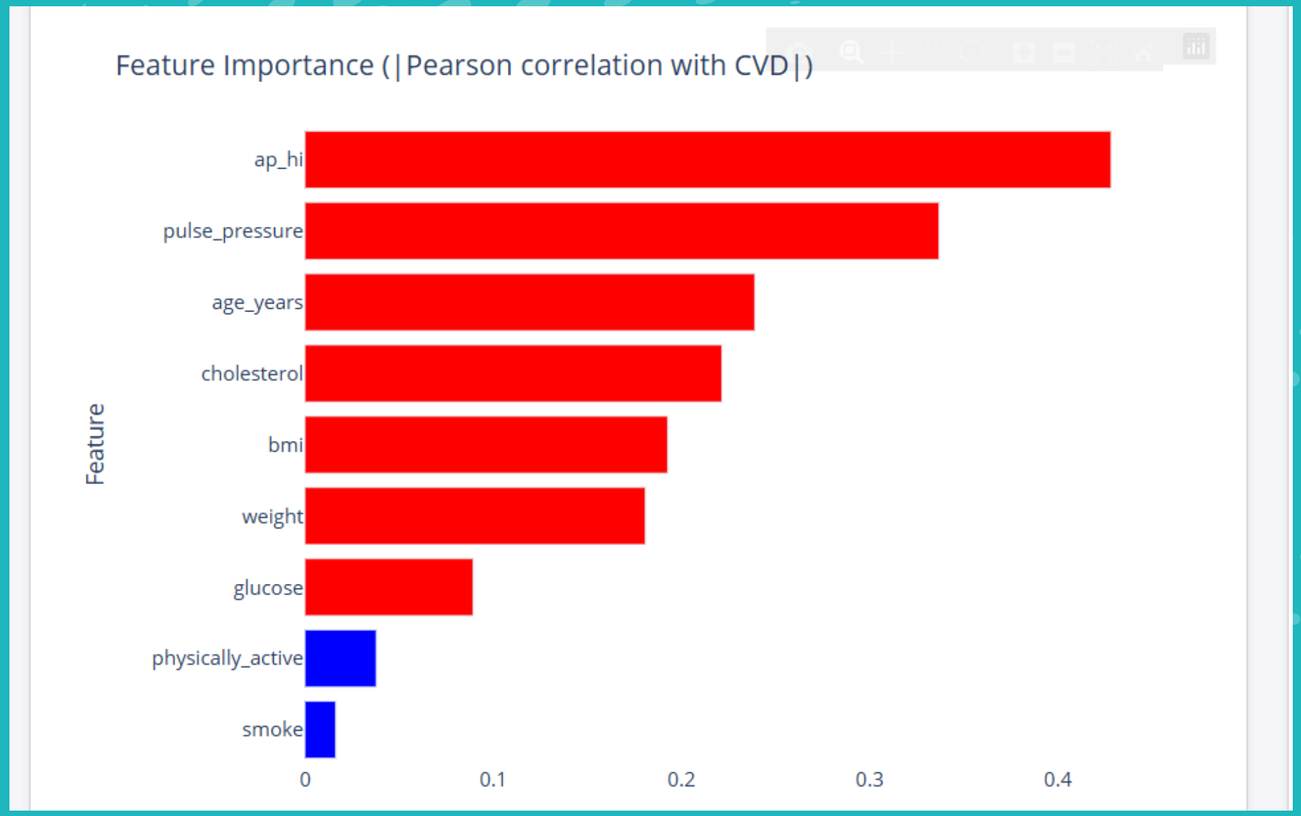
- **Systolic and diastolic blood pressure: r = 0.73 (high, but both retained for clinical interpretability)**
- **BMI and weight: r = 0.87 (very high, suggesting potential redundancy)**
- **Pulse pressure and systolic BP: r = 0.84 (high due to mathematical relationship)**

3.5 Correlation Analysis



Clinical Interpretation: The strong correlation of blood pressure metrics with CVD validates the dataset's clinical relevance. However, the near-zero correlations for lifestyle factors (smoking, physical activity, alcohol) are concerning and may reflect measurement error, underreporting bias, or the limitations of binary encoding for complex behaviors.

3.6 Feature Importance Analysis



3.6 Feature Importance Analysis

To identify which features contribute most to CVD prediction, we computed absolute Pearson correlation coefficients and visualized them as a ranked bar chart. This analysis complements the correlation matrix by providing a clear hierarchy of predictive features:

Top Predictive Features (ranked by correlation magnitude):

1. **Systolic Blood Pressure (ap_hi): Most important**
2. **Pulse Pressure: Second most important**
3. **Age: Third most important**
4. **Cholesterol: Fourth most important**
5. **BMI: Moderate importance**

Weight and Glucose: Lower-moderate importance

Negligible Features:

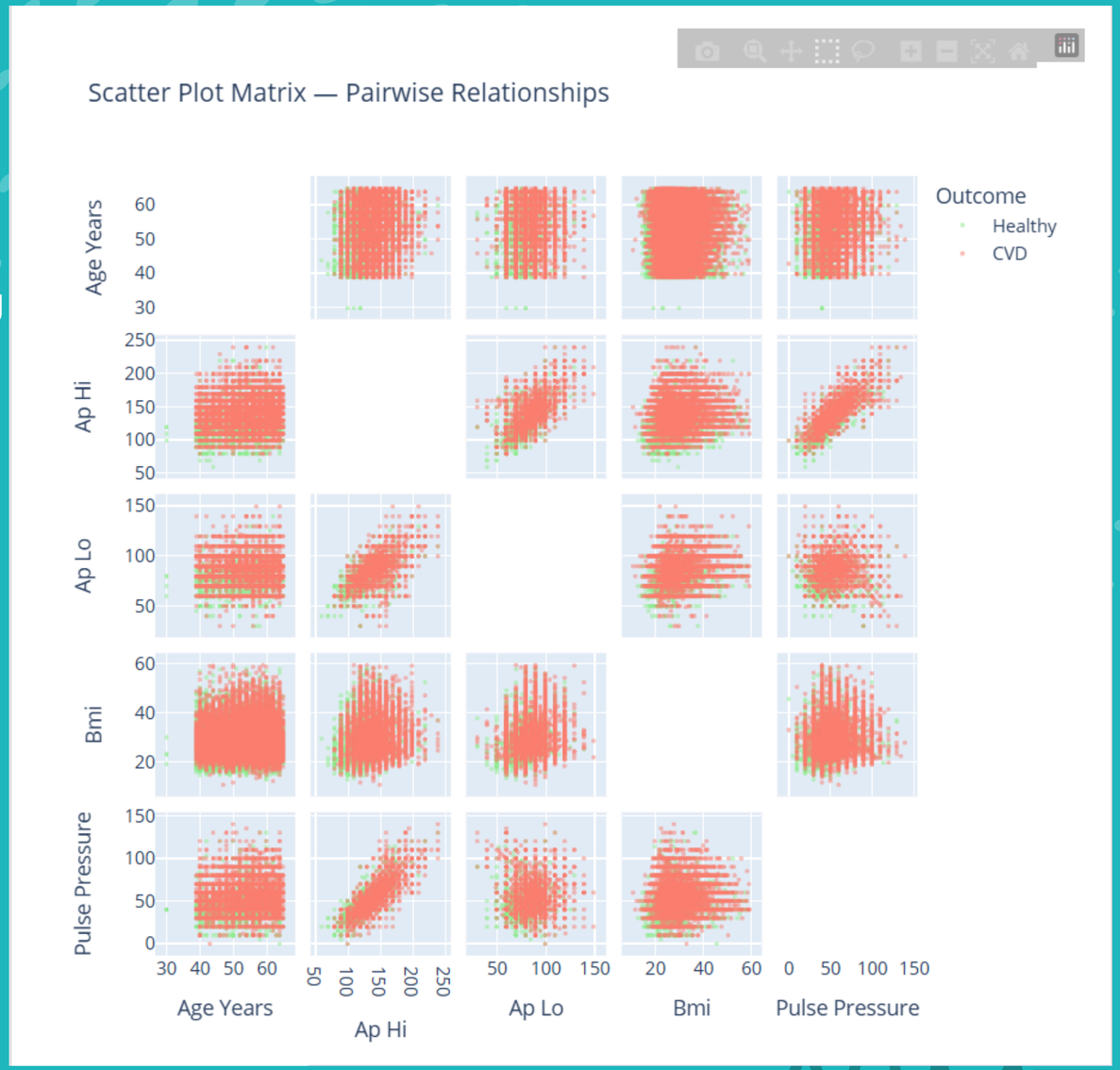
1. **Physical Activity: Extremely low correlation (shown in blue, indicating negative direction)**
2. **Smoking: Nearly zero predictive value in univariate analysis**

3.7 Bivariate and Multivariate Relationships

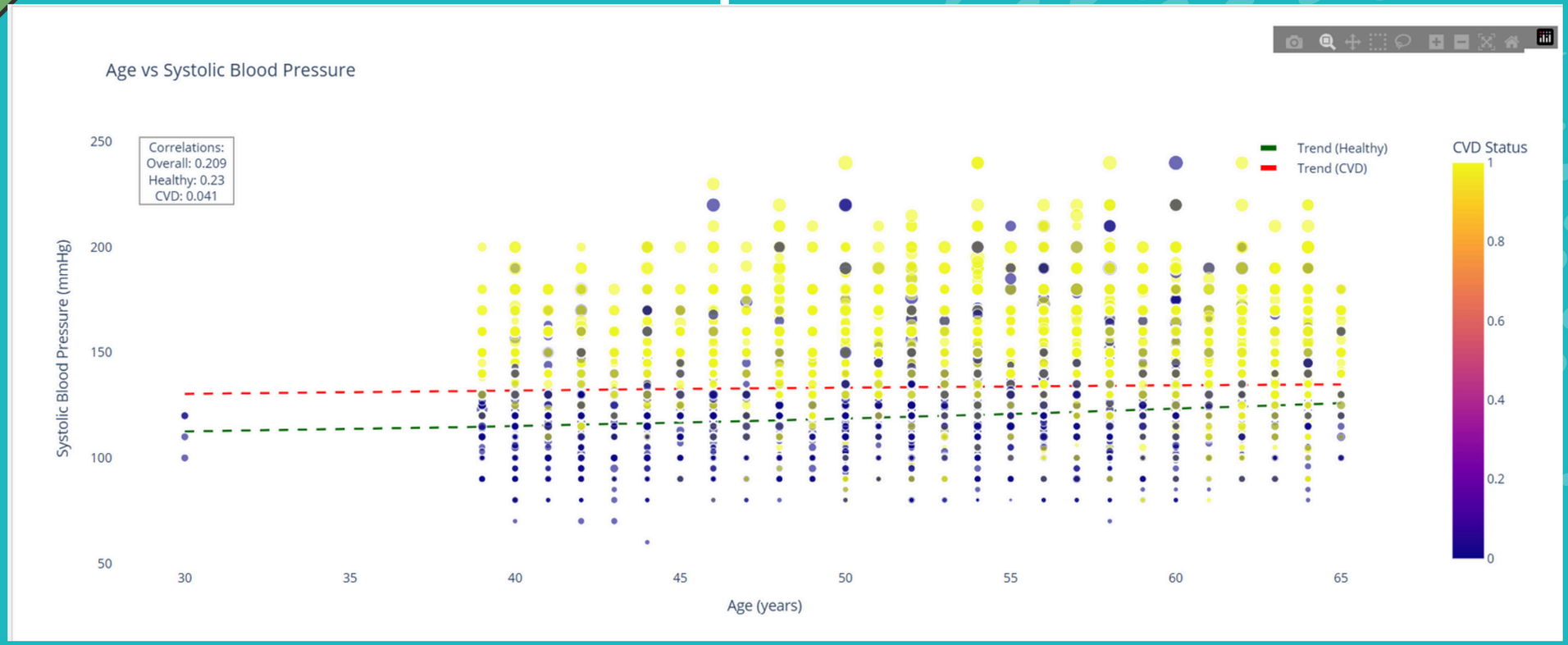
3.7.1 Scatter Plot Matrix

A comprehensive scatter plot matrix was created showing pairwise relationships between key continuous variables (age, systolic BP, diastolic BP, BMI), with points colored by CVD outcome (healthy = green, CVD = red). This visualization revealed:

- Blood Pressure Cluster Separation: Clear vertical stratification in BP-related plots, with CVD patients clustering at higher systolic and diastolic values
- Age-BP Interaction: Scatter plots of age vs. blood pressure show denser red regions (CVD patients) at higher ages and pressures simultaneously
- BMI Patterns: More diffuse relationship with CVD, suggesting BMI alone is less discriminative than blood pressure measurements



3.7 Bivariate and Multivariate Relationships

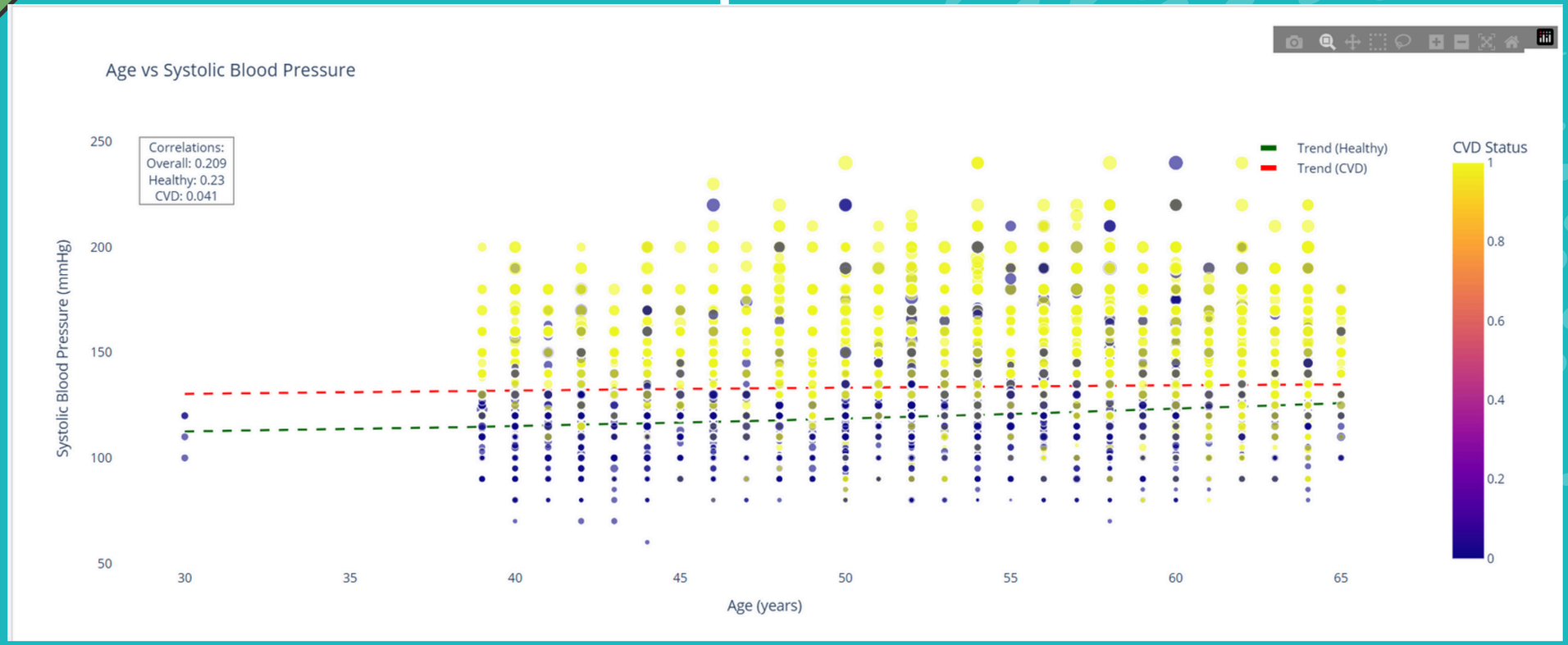


3.7.2 Age vs. Systolic Blood Pressure Analysis

A detailed scatter plot examining the age-systolic BP relationship with CVD status overlays demonstrated:

- Trend Lines: Separate regression lines for healthy and CVD groups show both populations experience age-related BP increases, but CVD patients maintain consistently higher BP at all ages
- Correlation Strength: Overall correlation of r = 0.209, with healthy patients showing r = 0.23 and CVD patients showing minimal correlation (r = 0.041), suggesting that within the CVD population, age-BP relationship weakens once disease is established

3.7 Bivariate and Multivariate Relationships

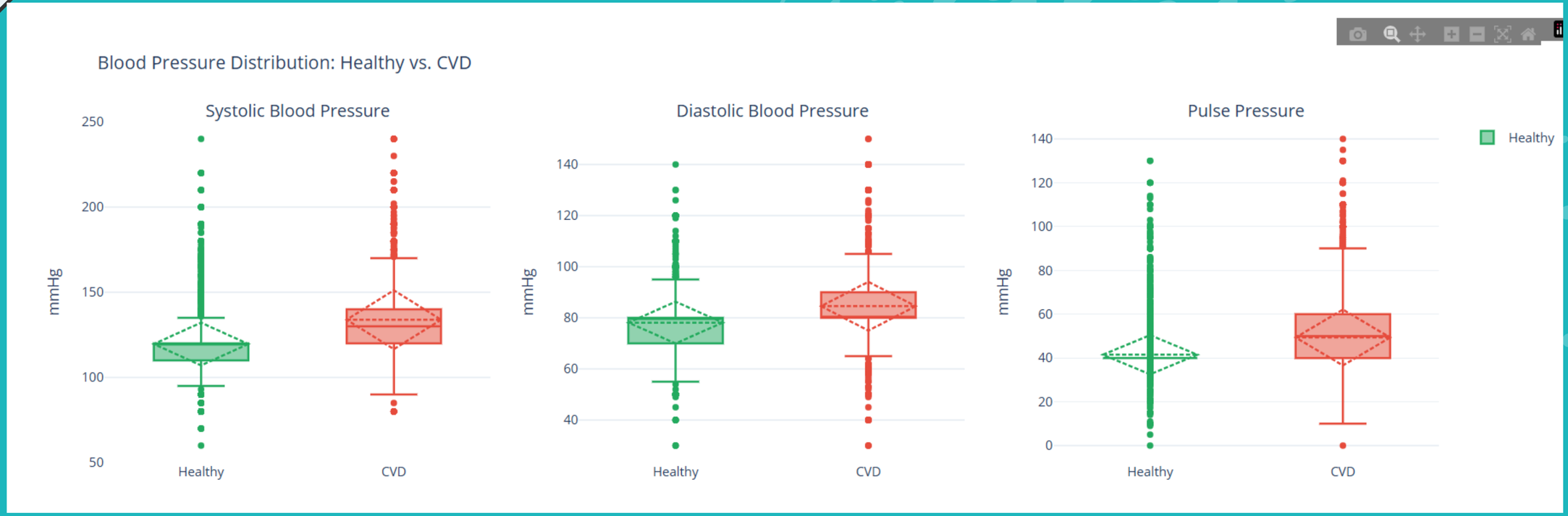


3.8 Blood Pressure Category Analysis

A detailed scatter plot examining the age-systolic BP relationship with CVD status overlays demonstrated:

- Trend Lines: Separate regression lines for healthy and CVD groups show both populations experience age-related BP increases, but CVD patients maintain consistently higher BP at all ages
- Correlation Strength: Overall correlation of r = 0.209, with healthy patients showing r = 0.23 and CVD patients showing minimal correlation (r = 0.041), suggesting that within the CVD population, age-BP relationship weakens once disease is established

3.8 Blood Pressure Category Analysis

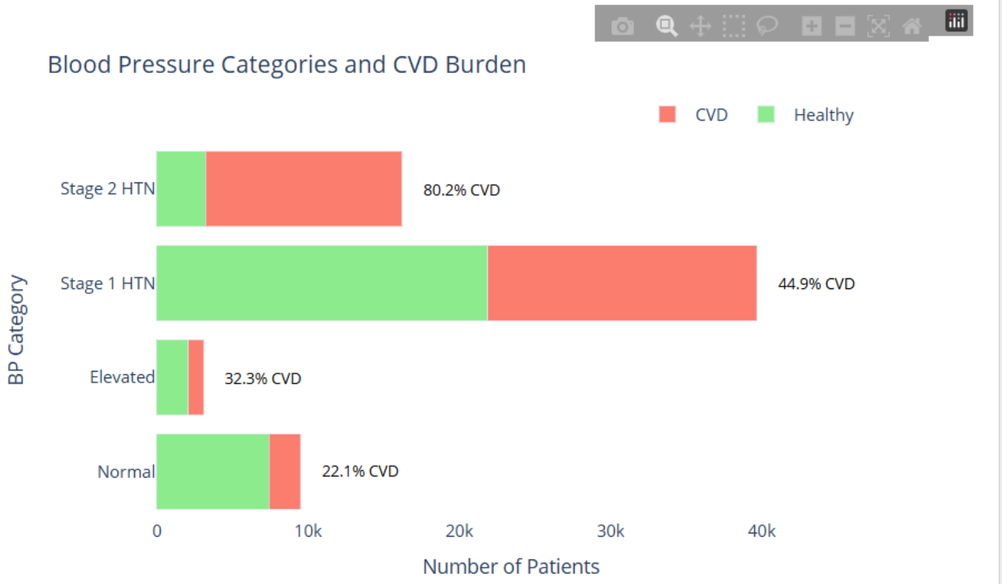


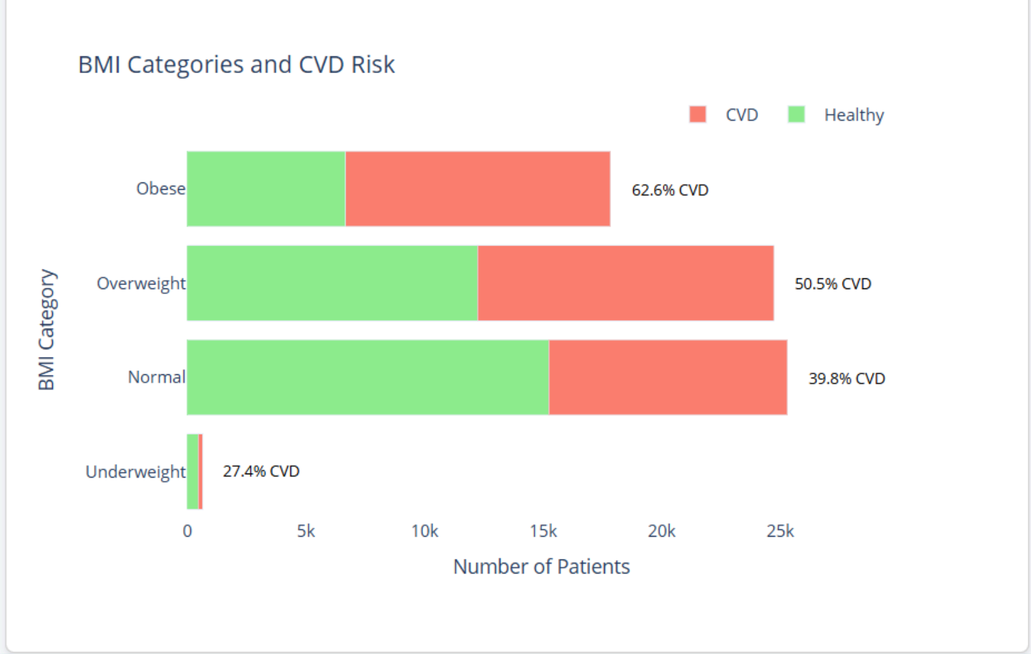
3.8.1 Distribution Comparison

Box plots comparing blood pressure metrics (systolic, diastolic, pulse pressure) between healthy and CVD groups revealed:

- **Systolic BP: Median for healthy patients ≈120 mmHg vs. CVD patients ≈135 mmHg (clear separation)**
- **Diastolic BP: Median for healthy ≈75 mmHg vs. CVD ≈85 mmHg**
- **Pulse Pressure: Median for healthy ≈42 mmHg vs. CVD ≈52 mmHg (indicating reduced arterial compliance in CVD group)**
- **Overlap: Despite clear median differences, substantial overlap exists between distributions, explaining why single-threshold rules are insufficient for CVD prediction**

3.8 Blood Pressure Category Analysis





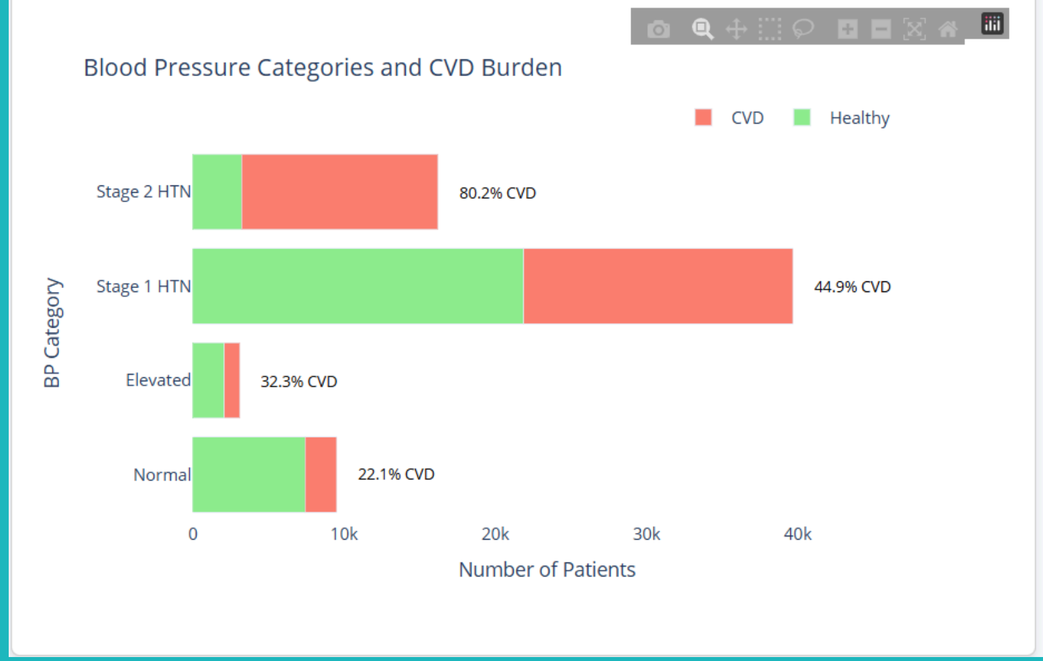
3.8.2 Hypertension Staging and Disease Burden

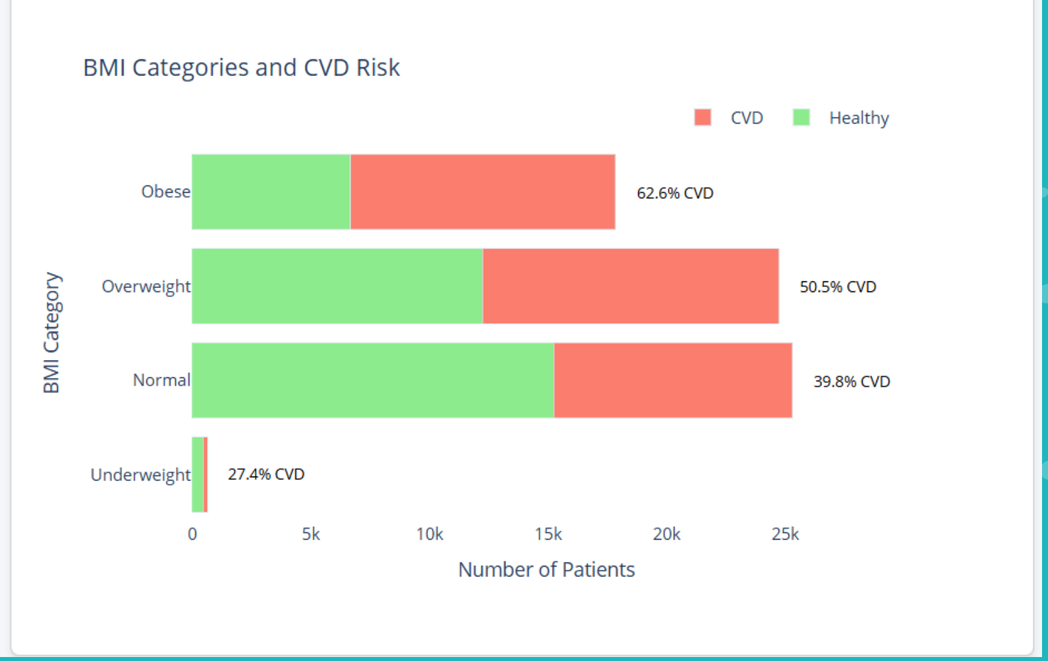
A stacked bar chart analyzing CVD prevalence across clinical blood pressure categories revealed striking patterns:

- Normal BP (<120/80): 22.1% CVD prevalence
- Elevated BP (120-129/<80): 32.3% CVD prevalence
- Stage 1 Hypertension (130-139/80-89): 44.9% CVD prevalence
- Stage 2 Hypertension (≥140/90): 80.2% CVD prevalence

Critical Insight: The dramatic increase in CVD prevalence with hypertension stage (especially Stage 2) reinforces that blood pressure management is central to cardiovascular health. Even patients with "normal" BP still show 22.1% CVD prevalence, indicating other risk factors are at play.

3.8 Blood Pressure Category Analysis





Critical Insight: The dramatic increase in CVD prevalence with hypertension stage (especially Stage 2) reinforces that blood pressure management is central to cardiovascular health. Even patients with "normal" BP still show 22.1% CVD prevalence, indicating other risk factors are at play.

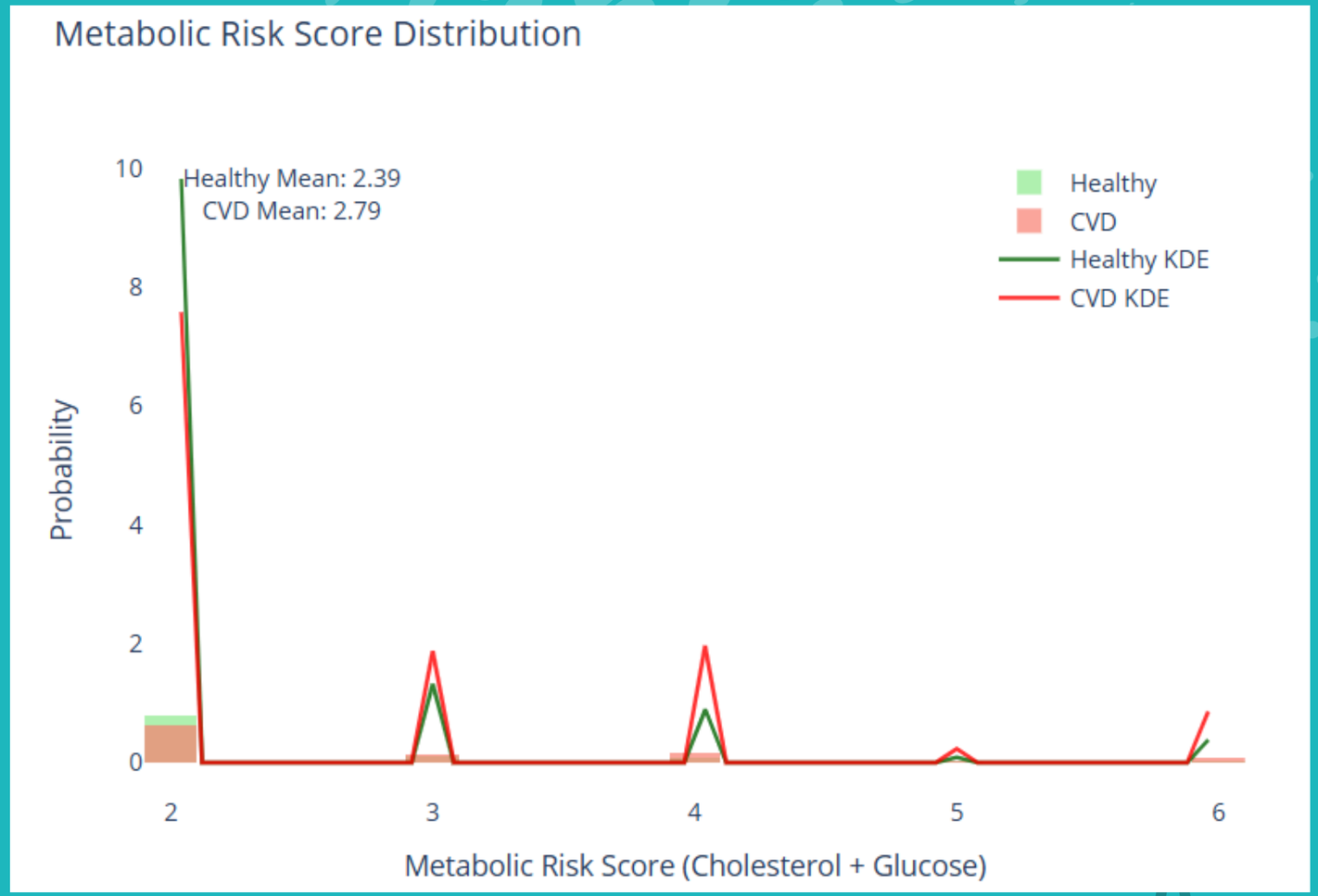
3.9 Metabolic Risk Analysis

3.9.1 Metabolic Risk Score Distribution

We visualized the distribution of the cholesterol-glucose interaction term (metabolic risk score) using kernel density estimation (KDE) curves for healthy and CVD populations:

- **Healthy patients: Mean metabolic score = 2.39**
- **CVD patients: Mean metabolic score = 2.79**
- **Distribution shift: CVD curve is right-shifted, indicating higher metabolic dysfunction**

This validates our feature engineering decision to include the cholesterol-glucose interaction term, as it captures combined metabolic risk.



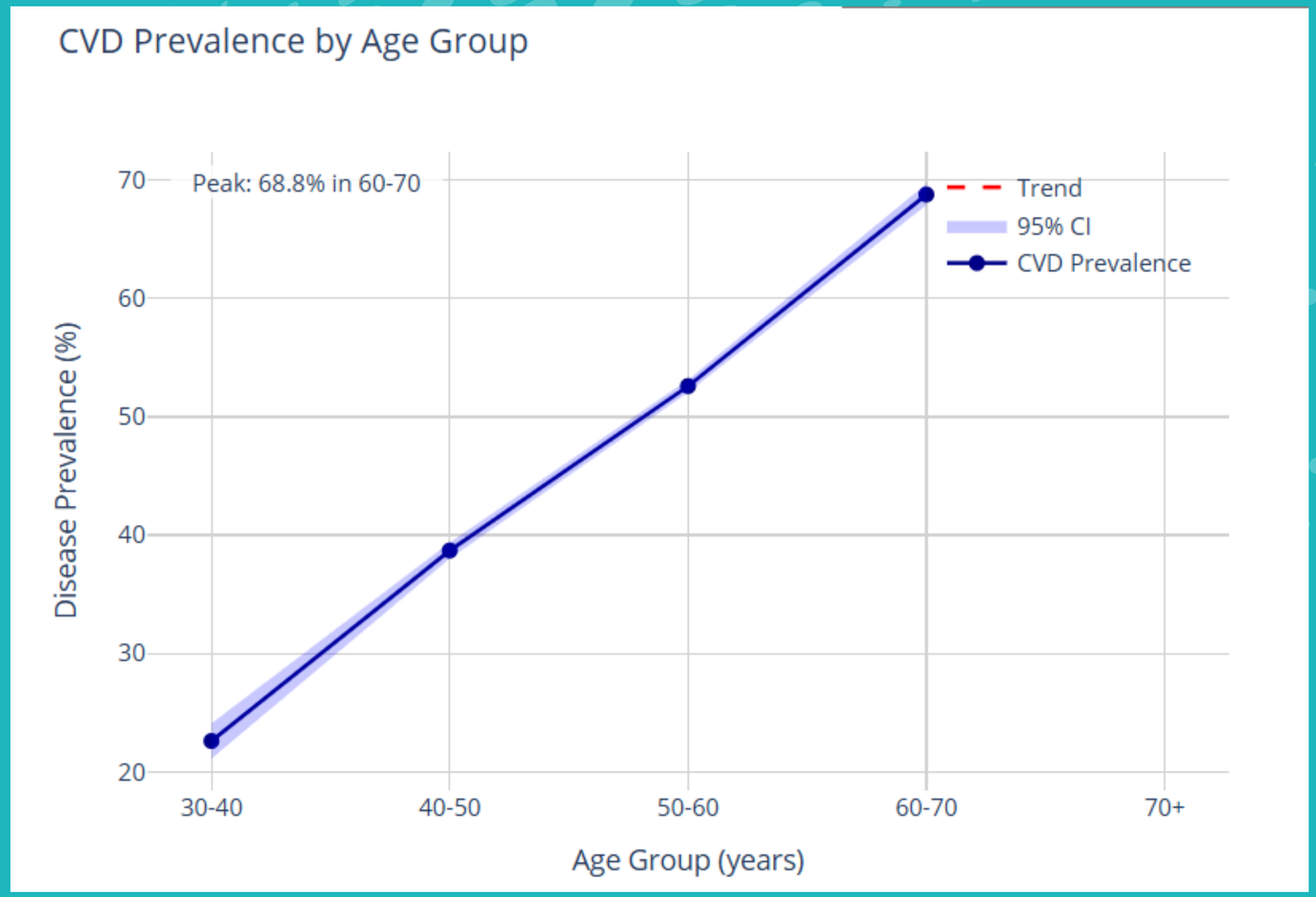
3.9 Metabolic Risk Analysis

captures combined metabolic risk.

3.9.2 Age-Stratified CVD Prevalence

A line plot showing CVD prevalence by age group demonstrated:

- Age 30-40: ~22% CVD prevalence
- Age 40-50: ~38% prevalence
- Age 50-60: ~53% prevalence
- Age 60-70: Peak at ~68.8% prevalence
- Age 70+: Slight decrease, likely due to survivorship bias



The near-linear increase in CVD prevalence with age (r^2 trend fit indicates strong linear relationship) confirms age as a critical risk factor and justifies its prominence in predictive models.

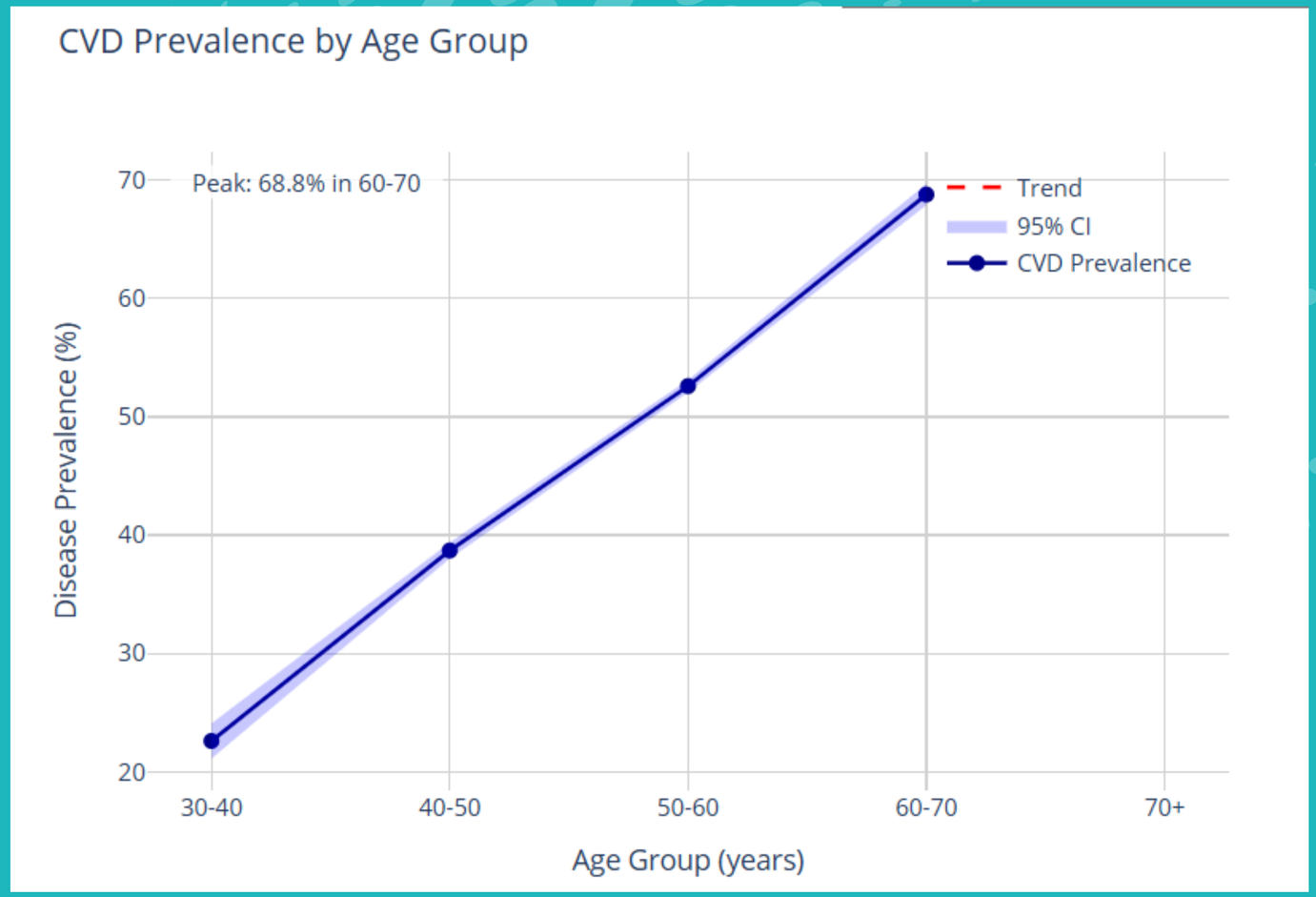
3.9 Metabolic Risk Analysis

captures combined metabolic risk.

3.9.2 Age-Stratified CVD Prevalence

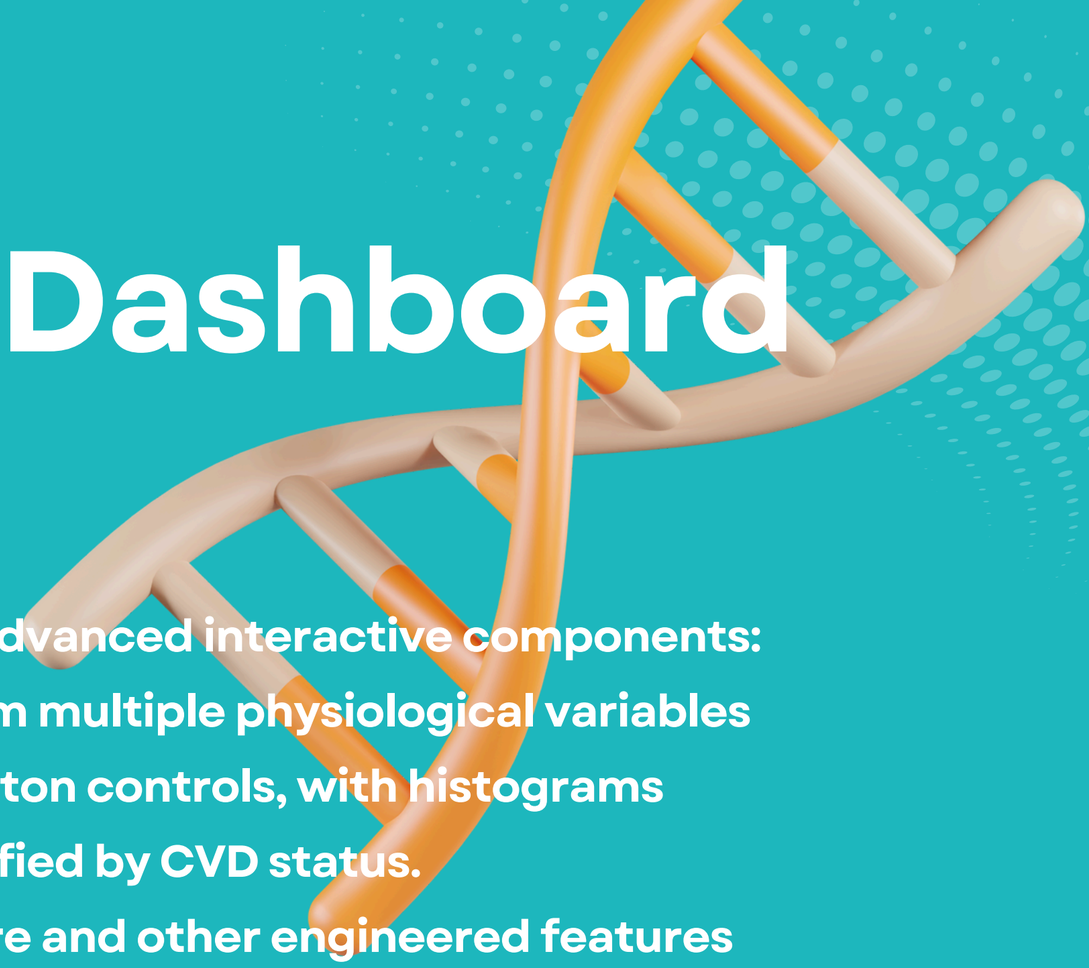
A line plot showing CVD prevalence by age group demonstrated:

- Age 30-40: ~22% CVD prevalence
- Age 40-50: ~38% prevalence
- Age 50-60: ~53% prevalence
- Age 60-70: Peak at ~68.8% prevalence
- Age 70+: Slight decrease, likely due to survivorship bias



The near-linear increase in CVD prevalence with age (r^2 trend fit indicates strong linear relationship) confirms age as a critical risk factor and justifies its prominence in predictive models.

3.10 Interactive Dashboard Features



3.10 Interactive Dashboard Features

The deployed Dash dashboard includes several advanced interactive components:

Dynamic Variable Selection: Users can select from multiple physiological variables (blood pressure, BMI, age, weight, height) via button controls, with histograms updating in real-time to show distributions stratified by CVD status.

Distribution Viewer: histograms for pulse pressure and other engineered features display overlapping distributions for healthy vs. CVD populations, allowing visual assessment of feature discriminative power.

Patient Data Table: A searchable, sortable table displays individual patient records with all features and calculated values, enabling case-by-case examination and quality verification.

3.11 Key Insights and Implications for Modeling

The visualization phase yielded several actionable insights:

1.Blood Pressure Dominance: Systolic blood pressure and pulse pressure emerge as the most powerful predictors, suggesting models will heavily weight these features

2.Lifestyle Factor Limitations: The weak univariate correlations for smoking and physical activity suggest these features may need careful handling or transformation, though they should not be discarded due to potential non-linear effects

3.Feature Engineering Success: Engineered features (BMI, pulse pressure, metabolic risk score, hypertension indicator) show strong relationships with CVD, validating their inclusion

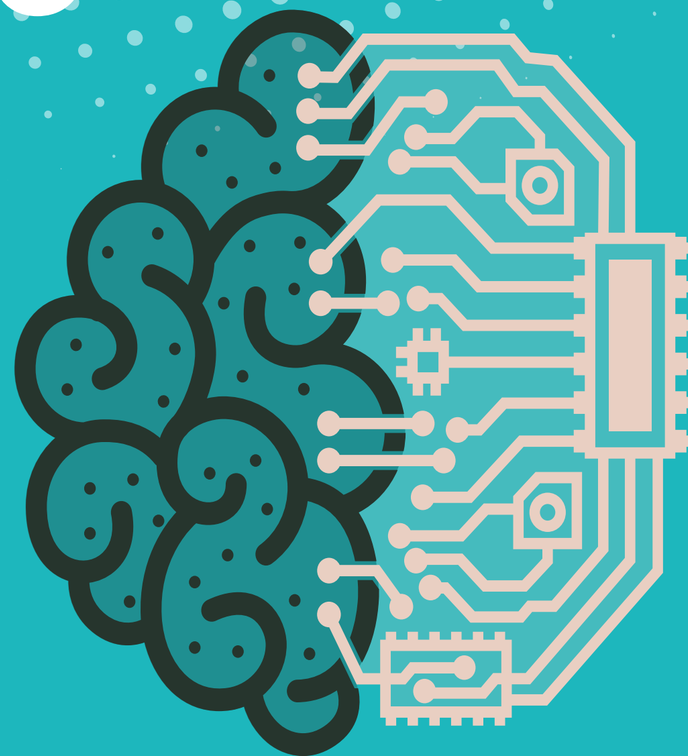
4.Age-Disease Gradient: The strong monotonic relationship between age and CVD prevalence suggests age will be critical for model calibration

5.Class Balance Advantage: The nearly perfect 50-50 split eliminates sampling concerns and ensures robust model training

6.Multicollinearity Considerations: High correlations between weight-BMI and BP measurements suggest feature selection or regularization may be needed to prevent redundancy in linear models

These insights directly informed our approach to machine learning model development in Phase 3, guiding feature selection, model choice, and evaluation strategies.

Phase 3: Cardiovascular Disease Prediction Using Machine Learning



machine learning Goal

THE GOAL OF THIS PROJECT IS TO BUILD A MACHINE LEARNING MODEL THAT PREDICTS WHETHER A PERSON HAS CARDIOVASCULAR DISEASE BASED ON MEDICAL AND DEMOGRAPHIC FEATURES.

Key objectives

1. Prepare and preprocess the dataset
2. Train multiple machine learning models
3. *Compare their performance*
4. *Tune the best models*
5. *Optimize the classification threshold*
6. *Select the final, most accurate model*

>>

workflow:

1. **DATA SPLITTING** : 75% training, 25% testing with stratification to maintain class balance
2. **FEATURE SCALING** StandardScaler applied to normalize features for optimal model performance
3. **MODEL TRAINING** Five algorithms tested: Logistic Regression, Random Forest, Gradient Boosting, Decision Tree, and XGBoost
4. **HYPERPARAMETER TUNING** RandomizedSearchCV with 5-fold cross-validation (30-40 iterations per model)
5. **THRESHOLD OPTIMIZATION** Custom threshold tuning to maximize F1 score for balanced predictions

Models Evaluated

MODEL	TYPE	KEY CHARACTERISTICS
Logistic Regression	Linear	Fast, interpretable baseline model
Random Forest	Ensemble	Robust, handles non-linear
Gradient Boosting	Ensemble	Sequential learning, high accuracy
Decision Tree	Tree-based	Simple, interpretable structure
XGBoost	Ensemble	Advanced boosting, regularization support

Why Multiple Models

Different algorithms capture different patterns in the data. Comparing them helps identify the best approach for this specific problem.

Initial Model Results

I trained each model on the scaled training set and evaluated them on the test set

Key observations

LOGISTIC REGRESSION

- **Train Accuracy** = 0.781
- **Test Accuracy** = 0.780
- **F1 = 0.757**

What it means:

- Moderate performance
- Logistic Regression assumes linear relationships between features and disease
- Our dataset has complex interactions (age + blood pressure + cholesterol together)
- So the model cannot capture all complexities → moderate performance
- Advantage: easy to interpret, useful for doctors

>>

Initial Model Results

DECISION TREE

- **Train Accuracy** = 0.977
- **Test Accuracy** = 0.779
- **F1** = 0.777

What it means:

- Very high training accuracy : overfitting
- Performance drops on the test set
- Decision Tree memorizes training data rules

RANDOM FOREST

- **Train Accuracy** = 0.977
- **Test Accuracy** = 0.813
- **F1** = 0.801

What it means:

- Ensemble of Decision Trees :reduces overfitting
- Performs better than a single tree on test data

>>

Initial Model Results

GRADIENT BOOSTING

- **Train Accuracy = 0.833**
- **Test Accuracy = 0.831**
- **F1 = 0.80**

What it means:

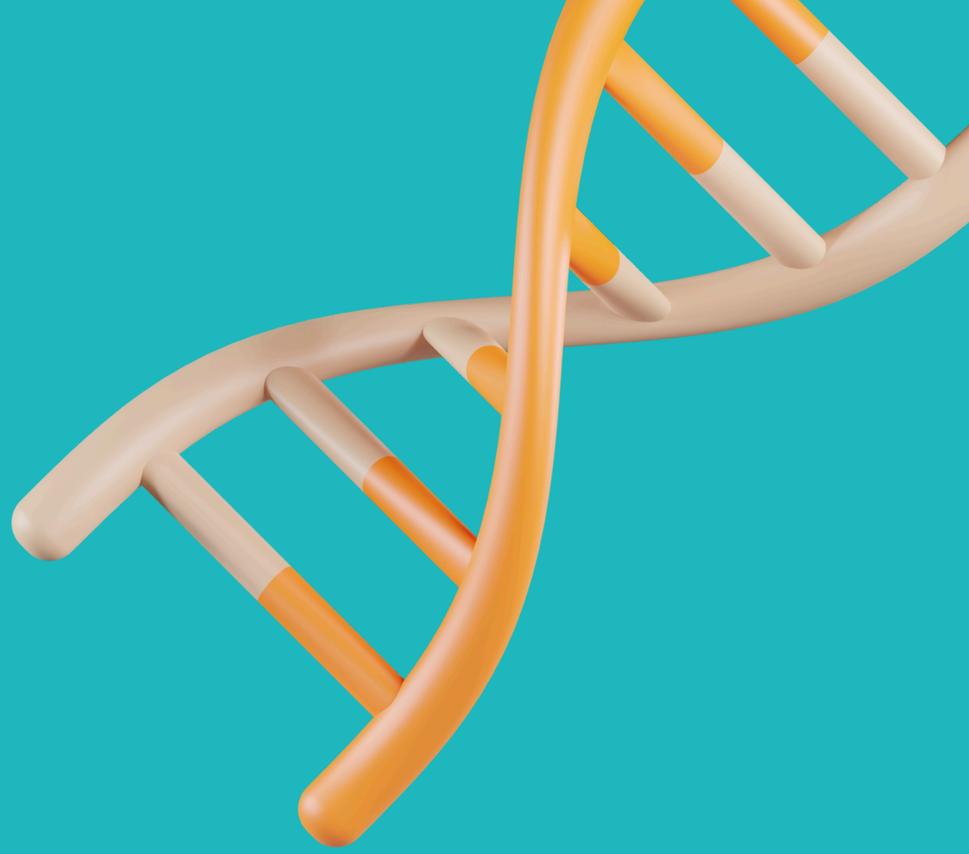
- Model corrects its mistakes step by step
- High accuracy on test set
- High precision : most predicted positives are correct
- Slightly lower recall : may miss some patients

XGBOOST

- **Train Accuracy = 0.854**
- **Test Accuracy = 0.834**
- **F1 = 0.814**
- **ROC-AUC = 0.909**

What it means:

- Best model overall
- Balances precision and recall



>>

Why Hyperparameter Tuning

After the initial comparison, I applied RandomizedSearchCV to improve model performance by tuning their hyperparameters.

This helps

- Reduce overfitting
- Improve generalization
- Optimize model structure
- Increase accuracy and recall

Hyperparameter Optimization

RandomizedSearchCV was used to efficiently explore hyperparameter space:

1

Random Forest

Parameters tuned:

- n_estimators: 100-300
- max_depth: 5-20
- min_samples_split: 2-10
- max_features: sqrt, log2

2

Random Forest

Parameters tuned:

- n_estimators: 100-300
- learning_rate: 0.01-0.2
- max_depth: 3-7
- subsample: 0.7-1.0

3

Logistic Regression

Parameters tuned:

- C (regularization): 0.01-10
- penalty: L2
- solver: lbfgs
- max_iter: 1000

4

XGBoost

Parameters tuned:

- learning_rate: 0.01-0.9
- max_depth: 3-10
- subsample: 0.7-1.0
- regularization: gamma, alpha, lambda

Evaluation Metrics

ACCURACY

Overall Correctness

Percentage of correct predictions (both positive and negative)

PRECISION

Positive Predictive Value

Of all predicted disease cases, how many were actually positive?

RECALL

True Positive Rate

Of all actual disease cases, how many did we correctly identify?

F1 SCORE

Harmonic Mean

Balance between precision and recall - critical for medical diagnosis

ROC-AUC

ROC-AUC measures how well the model distinguishes between patients with the disease and those without it

Medical Context

In cardiovascular disease prediction, high recall is crucial to avoid missing true cases (false negatives), while maintaining reasonable precision to avoid unnecessary treatments.

Best Models After Tuning

I evaluated the best model from each tuning search.

```
Logistic Regression : Train Accuracy: 0.781 | Test Accuracy: 0.780
Random Forest : Train Accuracy: 0.843 | Test Accuracy: 0.833
Gradient Boosting : Train Accuracy: 0.844 | Test Accuracy: 0.835
XGBoost : Train Accuracy: 0.839 | Test Accuracy: 0.835
                 Model  Accuracy  Precision    Recall        F1   ROC-AUC
2    Gradient Boosting  0.835371   0.929660  0.725443  0.814953  0.912015
3              XGBoost  0.834629   0.936968  0.717324  0.812565  0.912039
1        Random Forest  0.833371   0.953760  0.700515  0.807753  0.910732
0  Logistic Regression  0.780171   0.845027  0.685878  0.757180  0.856303
```

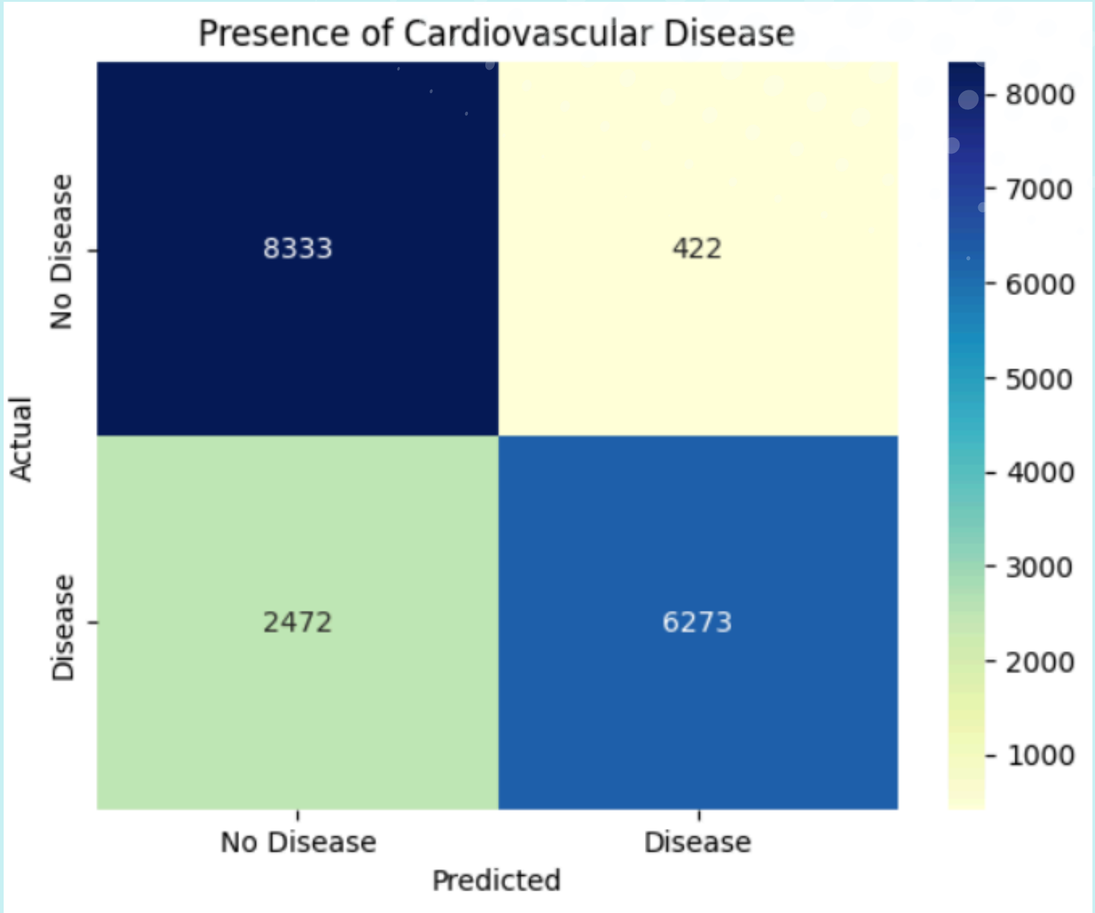
Key findings:

- XGBoost remained the strongest model
- Gradient Boosting and Random Forest were close
- Logistic Regression improved but was still behind

Confusion Matrix

The confusion matrix shows that the model performs well overall, but it still misses some actual disease cases (false negatives), which is critical in medical diagnosis.

- **True negative** : 8333 cases correctly predicted as not infected
- **False Positives** :422 cases incorrectly predicted as infected
- **True positive** :6273 cases correctly predicted as infected
- **False Negatives** :2472 cases incorrectly predicted as not infected



Threshold Optimization

Default classification threshold (0.5) was optimized for each model to maximize F1 score:

Process

Tested thresholds from 0.0 to 1.0 in 0.01 increments
Selected threshold that yielded highest F1 score for each model
This balances the trade-off between precision and recall
best threshold is 0.40

Lower Threshold

↑ Recall (fewer missed cases)
↓ Precision (more false alarms)

Higher Threshold

↑ Precision (fewer false alarms)
↓ Recall (more missed cases)

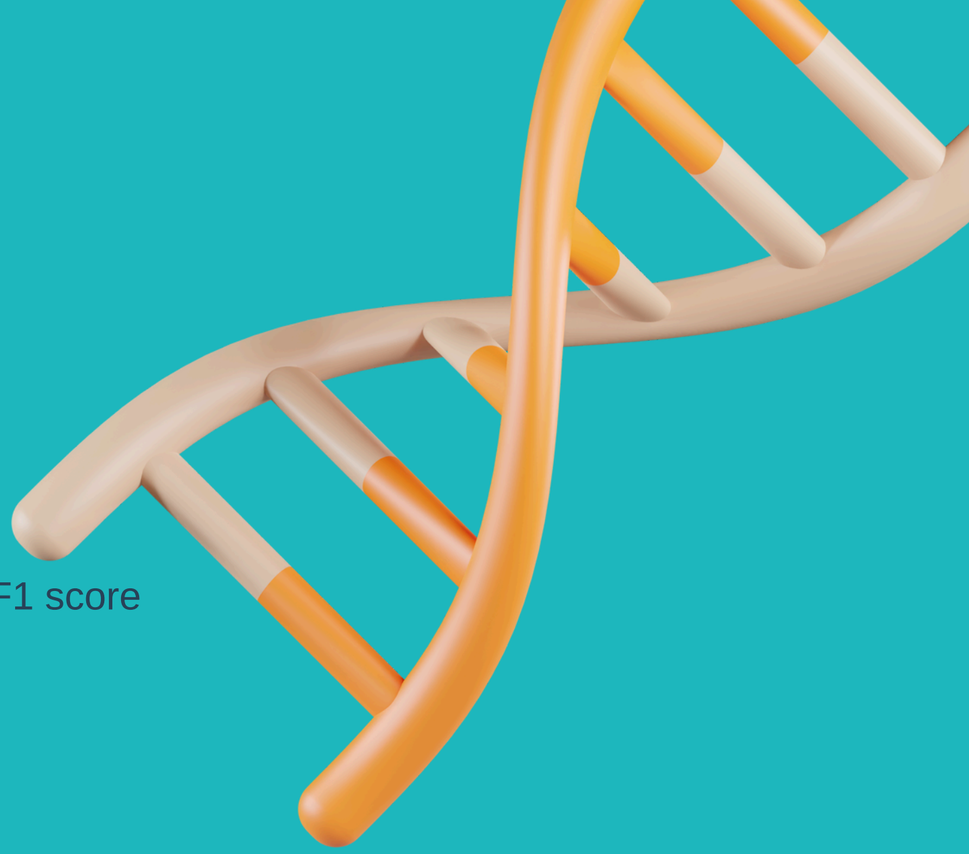
Key Results & Conclusions

Best Model: XGBoost

XGBoost was selected as the final model after threshold optimization based on highest F1 score

Key Findings

- Hyperparameter tuning significantly improved model performance across all algorithms
- Threshold optimization provided better balance between precision and recall
- Ensemble methods (Random Forest, Gradient Boosting, XGBoost) outperformed simpler models
- Feature scaling was essential for optimal performance



>>

Phase 4: MLOps, Deployment,and Monitoring

Introduction to the Final Phase

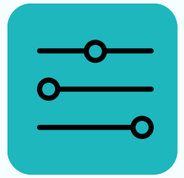
The final phase of this project focuses on transforming the trained machine learning models into a fully operational, scalable, and maintainable system that can be used by healthcare professionals in real-world scenarios. After completing data cleaning, feature engineering, and model training earlier in the workflow, this final stage ensures that the model is not only accurate but also reliable, reproducible, and accessible through modern deployment techniques.

MLOps Implementation

In this project, MLOps was implemented using MLflow to ensure that every experiment, model version, hyperparameter configuration, and evaluation metric was tracked and stored. Instead of training a single model, the pipeline was redesigned to evaluate and compare five different algorithms—Logistic Regression, Random Forest, Gradient Boosting, Decision Tree, and XGBoost. Each model was trained through a unified pipeline that included consistent preprocessing using a ColumnTransformer and StandardScaler.

MLflow to Track:

Experiments



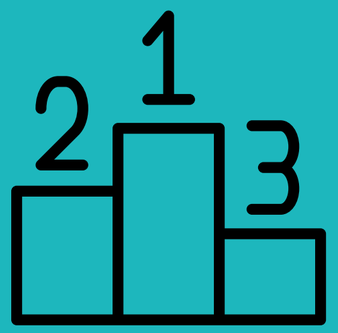
Hyperparameters

Metrics

Stored Artifacts

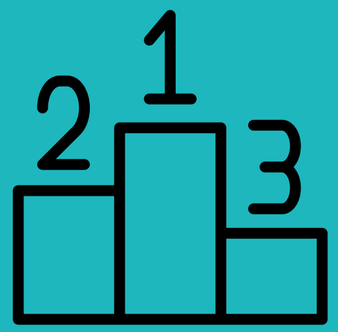
Model Comparison, Selection, and Registry Integration

Once all models were trained and their metrics logged, an automated comparison module ranked them based on accuracy (with the option to switch to metrics like F1-Score or AUC). This allowed the pipeline to objectively select the best-performing model without manual intervention.



Model Comparison, Selection, and Registry Integration

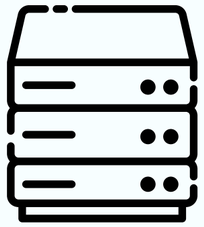
After identifying the top model, it was automatically registered into the MLflow Model Registry. Registering the model formalizes it as the official versioned model for production. MLflow assigned it a version number and transitioned it into the "Production" stage, making it easy to deploy and track over time. This system also supports future retraining cycles where newer versions can replace older ones while keeping historical versions archived.



Local Artifact Saving and Metadata Generation

The pipeline saved several useful artifacts locally. These included the serialized best model pipeline (cvd_best_pipeline.joblib), a metadata file describing the model's properties:

- features
- accuracy
- version
- run ID



These files provide a quick, offline way to inspect the model and support deployment in environments where the MLflow server may not be running. Local artifact generation ensures flexibility and creates a fallback mechanism for deployment.

Experiment Tracking with MLflow

Aldenaire & Partners

Thank You

Thank You for Your Attention

At Aldenaire & Partners, we are dedicated to improving healthcare and enhancing lives. We are here to support you in every step of your health journey. We look forward to serving you in 2030 and beyond.