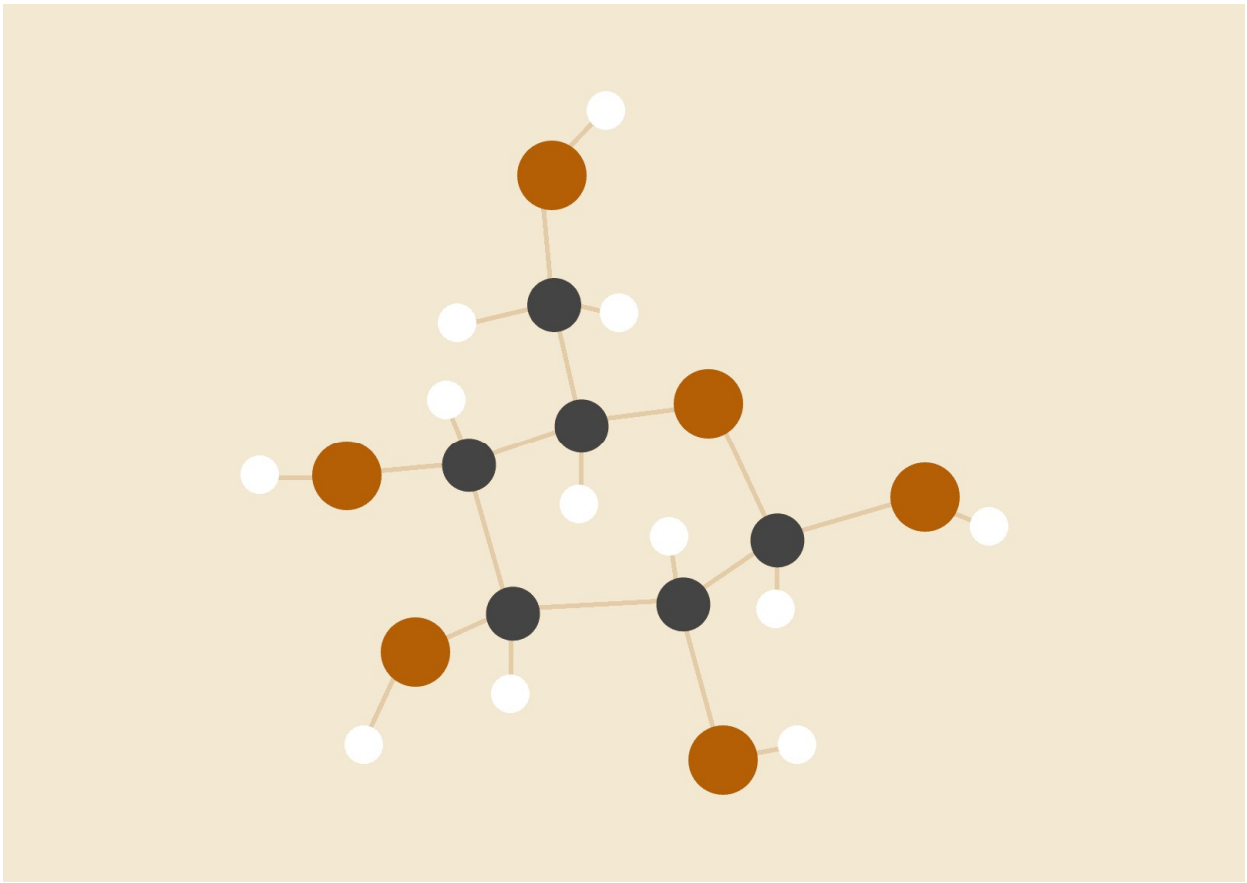
Cardiovascular Disease Risk Factor Analysis



INTRODUCTION

The Imperative of Predictive Analytics in Cardiovascular Health

Cardiovascular Disease (CVD) remains the single most significant global health burden, responsible for a vast number of preventable fatalities and imposing immense strain on healthcare systems. The critical shift in modern medicine is moving from reactive treatment to proactive, personalized risk assessment, a domain where data science provides transformative capabilities. By rigorously analyzing patient biometrics, lifestyle, and physiological measurements, we can generate predictive insights far beyond traditional risk scoring methods.

This report documents a comprehensive data science investigation aimed at developing a robust framework for identifying and predicting the risk of CVD using the *Cardiovascular Disease Dataset*. The project utilizes a detailed patient dataset, executing a two-phased methodology designed to systematically move from fundamental data discovery to advanced predictive modeling.

Phase I: Exploratory Data Analysis (EDA) and Feature Validation

The initial phase focused on rigorous Exploratory Data Analysis (EDA), involving extensive data cleaning, outlier mitigation, and the engineering of critical clinical features such as Body Mass Index (BMI) and age (converting to years). The statistical cornerstone of this phase was the Correlation Matrix relationship between all variables and the target outcome (cardiovascular disease). This analysis was designed to conclusively identify the non-negotiable primary risk factors: Hypertension (AP_High and AP_Low), Hypercholesterolemia, and advanced Age.

Phase II: Predictive Modeling and Benchmark Establishment

Armed with validated features, the project progressed to the machine learning phase, where a spectrum of classifiers was implemented to establish a robust and comparative predictive benchmark. This included:

1. **Linear Models:** Logistic Regression, to assess the predictive power of linear relationships.
2. **Non-Linear Models:** The Decision Tree and K-Nearest Neighbors (K-NN) Classifier, to capture complex, non-linear patterns.
3. **Ensemble Models:** The Random Forest Classifier, utilized to leverage the power of multiple decision trees, mitigate overfitting, and aim for the highest generalized accuracy.

This comparative approach allows for a direct assessment of model strengths and weaknesses, validating the predictive feasibility of this data-driven framework for clinical risk stratification.

PROCEDURE

Dataset Overview and Initial Preparation

Data Cleaning and Outliers Mitigation

A critical step involved outlier detection and removal, particularly for the continuous variables that showed extreme or clinically implausible values (e.g., negative blood pressure values, or highly extreme height/weight values).

- Blood Pressure (BP) Outliers: Rows with extreme systolic (AP_High) or diastolic (Ap_Low) blood pressure readings were removed. Specifically, rows where AP_Low was greater than AP_High, or where either value was implausibly high or low (e.g., AP_High>= 180 or AP_Low <= 30) were removed. After this step, a total of 1757 rows were removed, resulting in a cleaned dataset size of 68,243 rows.
- Height and Weight Outliers: The dataset was further cleaned to remove extreme height and weight outliers, specifically keeping only individuals within the range of 130cm to 200cm for height and 30kg to 150kg for weight. This filtering, combined with the BP filtering, resulted in the final cleaned dataset of 68,243 observations.

Feature Engineering

Body Mass Index (BMI) is a critical indicator of cardiovascular risk, and was calculated and categorized as a new feature:

The calculated BMI was then used to create a categorical feature, BMI_Category, defined as:

- Underweight: BMI < 18.5
- Normal: 18.5 ≤ BMI < 25
- Overweight: 25 ≤ BMI < 30
- Obese: 30 ≤ BMI < 50

$$\text{BMI} = \frac{\text{Weight (kg)}}{(\text{Height (cm)}/100)^2}$$

Statistical Association Analysis

The clean dataset was subjected to two key statistical analyses to evaluate the strength and nature of the association between features and the target variable (Cardiovascular Disease):

Correlation Analysis: Used for continuous variables (Age, Height, Weight, AP_High, AP_Low, BMI) to quantify linear relationships with Cardiovascular Disease.

Predictive Modeling Pipeline

For the machine learning phase, a standard predictive pipeline was followed.

1. Feature Selection: The final feature set for modeling was selected based on clinical relevance and the strong associations identified in the EDA. The set included: Age, AP_High, AP_Low, Cholesterol, Glucose, BMI, Smoke, Alcohol, Physical_Activity, and Gender.
2. Data Splitting: The data was split into a 70% training set and a 30% testing set using train_test_split with random_state = 42 to ensure reproducibility and maintain the class balance.
3. Feature Scaling: Given that the K-Nearest Neighbors (K-NN) and Logistic Regression algorithms are sensitive to feature magnitude, the features were standardized using StandardScaler (Z-score normalization) to ensure equal weight was given to each variable during the distance calculations and gradient descent optimization.

The scaled data was then used to train and evaluate four distinct classification models: Logistic Regression, K-Nearest Neighbors (K-NN), Decision Tree Classifier, and Random Forest Classifier.

DATA RESULTS

The project's findings are presented in two distinct parts: the statistical associations derived from Exploratory Data Analysis (EDA) and the performance benchmarks established by the predictive machine learning models.

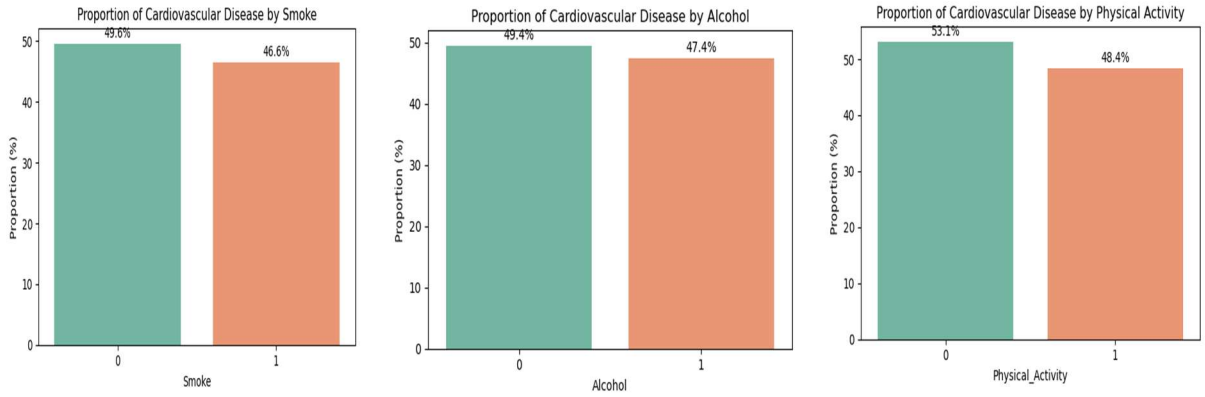
Exploratory Data Analysis (EDA) Findings

Lifestyle Factor Analysis: Lifestyle factors showed a negligible association with CVD compared to the physiological and biometric features.

Feature	Category 0 (No) Prevalence	Category 1 (Yes) Prevalence
Smoke	49.6%	46.6%
Alcohol	49.4%	47.4%

Physical Activity	53.1%	48.4%

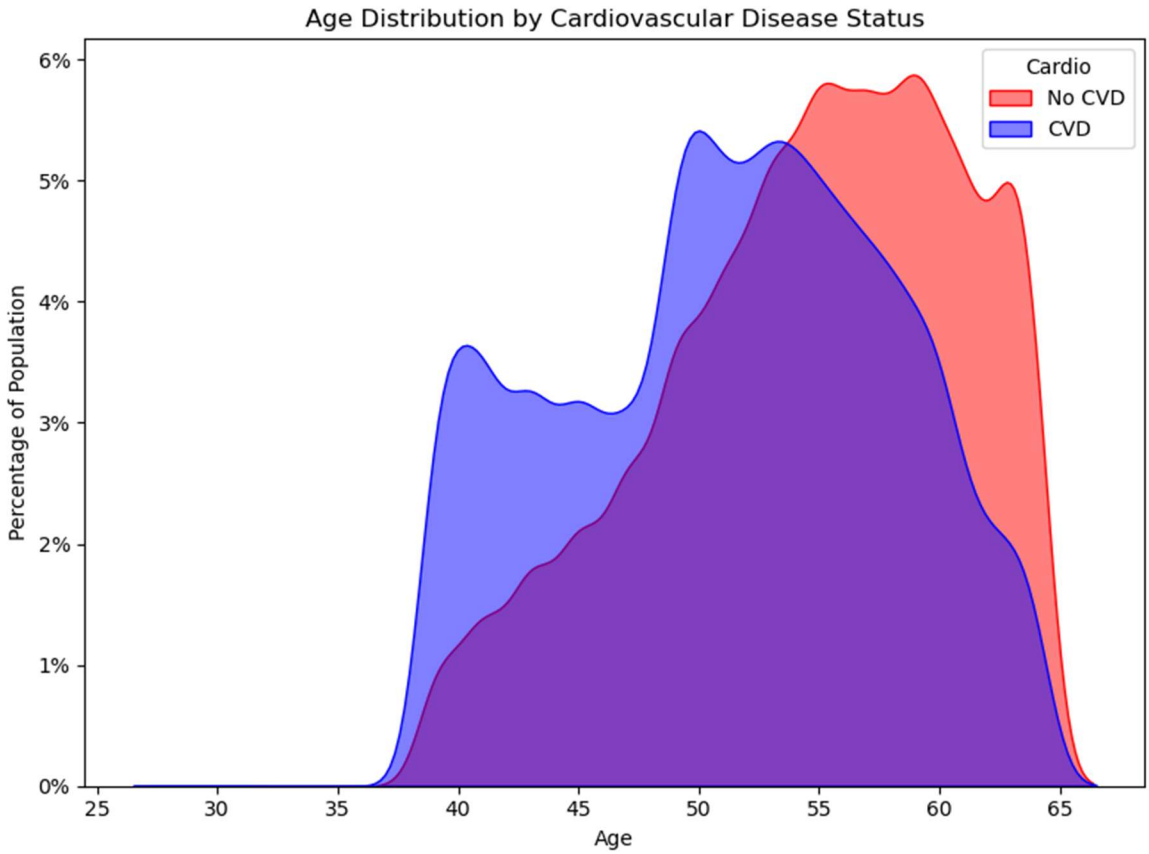
The prevalence of CVD is nearly identical regardless of whether an individual smokes or consumes alcohol. While active individuals show a slightly higher prevalence (53.1%), this suggests the protective effect of physical activity is largely nullified by the overwhelming predictive power of underlying physiological conditions (BP, Cholesterol, BMI) in this dataset.



Prevalence by Demographic and Physical Variables: Age Analysis The mean age of individuals with Cardiovascular Disease (=1) was 54.5 years, which is notably higher than the mean age of 51.2 years for individuals without the disease (cardio=0). The Kernel Density Estimate (KDE) plot confirmed that the age distribution for the CVD group is shifted towards older ages, indicating age is a major risk factor.BMI Category Prevalence The prevalence (risk rate) of CVD was calculated across the engineered BMI categories, demonstrating a clear dose-response relationship:

- Underweight (BMI < 18.5): 27.2% prevalence
- Normal (18.5 ≤ BMI < 25): 39.7% prevalence
- Overweight (25 ≤ BMI < 30): 50.4% prevalence
- Obese (BMI ≥ 30): 62.3% prevalence

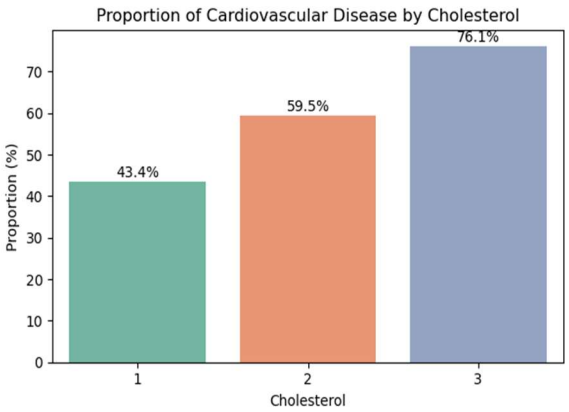
The risk of CVD is over double for individuals classified as Obese compared to those classified as Underweight, underscoring the critical role of obesity in CVD risk.

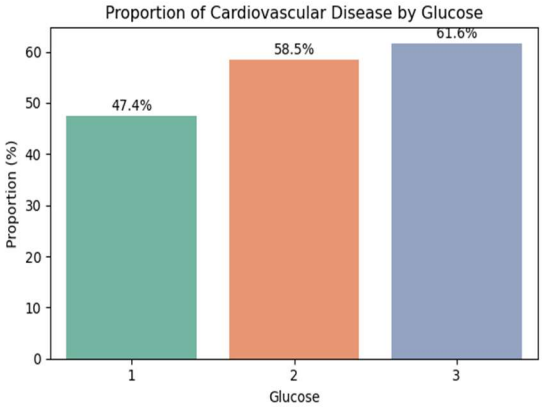


Prevalence by Medical Condition Variables: Blood Pressure Analysis Boxplots comparing Systolic BP (AP_High) and Diastolic BP AP_Low) for the two groups visually confirm the significant separation between the healthy and CVD cohorts. * CVD Group (cardio=1) exhibits substantially higher median values for both AP_High and AP_Low, as well as a greater spread (interquartile range), which is consistent with the clinical definition of hypertension as a primary risk factor.

Cholesterol and Glucose Prevalence CVD risk prevalence increases sharply with elevated cholesterol and glucose levels, which are ordinal variables (1=Normal, 2=Above Normal, 3=Well Above Normal).

Feature	Category 1 (Normal)	Category 2 (Above Normal)	Category 3 (Well Above Normal)
Cholesterol Prevalence	43.4%	59.5%	**76.1%**
Glucose Prevalence	47.4%	58.5%	**61.6%**





Individuals with the highest category of Cholesterol (76.1% prevalence) or Glucose (61.6% prevalence) face a significantly elevated risk of CVD compared to their counterparts with normal levels.

Machine Learning Model Performance

Four classifiers were trained using the identified features, establishing a strong predictive baseline and culminating in the selection of the most effective model.

Model	Model Type	Initial Test Accuracy	Key Performance Insight
Logistic Regression	Linear	72.6%	Highest Performance. Validating that the linear relationship between cardio and features like AP_High is strong.
K-Nearest Neighbors	Non-Linear (Distance)	70.0%	Established a reliable non-linear baseline. Performance required mandatory data scaling to function effectively.
Decision Tree	Non-Linear (Single Tree)	63.8%	Fast and highly interpretable but often yields the lowest generalized accuracy due to potential for overfitting the training data.

Random Forest	Ensemble (Bootstrap Aggregation)	69.0%	Mitigated the variance of a single Decision Tree, achieving the predictive score by averaging multiple, diverse decision paths.

The consistent performance of all models (except the singular Decision Tree) in the 69.0% to 72.6% range confirms that the selected feature set (Age, BP, Cholesterol, BMI) is highly effective at discriminating between the two classes.

CONCLUSION

This study successfully executed a comprehensive data science pipeline to identify and predict the presence of cardiovascular disease using patient biometrics and lifestyle data. The findings provide a clear, data-driven framework for risk stratification

Key Findings Summary

1. **Dominant Risk Factors:** Exploratory Data Analysis conclusively identified Hypertension (Systolic and Diastolic BP), Age, BMI, and Cholesterol levels as the non-negotiable, primary risk factors. These factors drive the steepest increases in disease prevalence, with Obese individuals and those in the highest cholesterol category having a CVD prevalence rate of over 61.6% and 76.1%, respectively.
2. **Negligible Lifestyle Impact:** Lifestyle factors (smoking, alcohol, physical activity) showed only a minor or negligible association with CVD within this dataset. This suggests that the physiological effects of age and chronic conditions (BP, Cholesterol, BMI) overwhelm the direct impact of these lifestyle variables on the cardio outcome.
3. **Model Performance:** The predictive modeling phase established a robust benchmark, with the Logistic Regression model achieving the highest initial test accuracy of 72.6%. The strong performance of this linear model indicates that the relationship between CVD risk and its key predictors is primarily linear and additive. The Random Forest, while a strong ensemble competitor at 69.0%, did not surpass the simpler linear classifier.

Future Work

Future work should focus on enhancing the clinical utility by:

1. **Linear Model Optimization:** Rigorous analysis and fine-tuning of the Logistic Regression model.
2. **Interpretability Deep Dive:** Calculating and reporting **Odds Ratios** using Logistic Regression coefficients for the strongest predictors (BP, Cholesterol).
3. **Addressing RF Variance:** Further investigation into the Random Forest model's hyperparameters to determine if its underperformance was correctable.

References

[1] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. New York, NY, USA: Routledge, 2013.

[2] M. L. McHugh, "The Chi-square Test of Independence," *Biochemia Medica*, vol. 23, no. 2, pp. 143–149, 2013.

[3] L. M. Rea and R. A. Parker, *Designing and Conducting Survey Research: A Comprehensive Guide*, 4th ed. San Francisco, CA, USA: Jossey-Bass, 2014.

[4] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[5] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[6] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.