Exploring Risk Factors for Cardiovascular Disease

Kumar Karthik Ankasandra Naveen

Sai Preetham Nagulapalli

Hari Priya Ramamoorthy

College of Professional Studies, Northeastern University

ALY 6040 Data Mining Applications

Professor. Justin Grosz

March 24, 2025

Introduction

Heart disease remains the leading cause of death in the United States, with approximately 697,000 fatalities in 2021 (CDC, 2023). This study analyzes the Personal Key Indicators of Heart Disease dataset from Kaggle, derived from CDC surveys, to identify key risk factors across U.S. states. This report covers data cleaning, exploratory analysis, hypothesis testing, and both linear and non-linear modeling to assess how demographic, health, and lifestyle factors influence heart disease prevalence and inform data-driven recommendations.

Data Cleaning

The dataset contains 246,022 records and 40 variables, capturing demographic, behavioral, and health-related factors. To support the needs of preventive health departments, the analysis focused on variables such as age, race/ethnicity, BMI, diabetes status, arthritis, asthma, pulmonary conditions, physical activity, and lifestyle attributes including smoking, alcohol use, sleep duration, and mental health indicators.

The cleaning process involved the following steps:

- **Missing Values:** An initial inspection showed no missing values, indicating the dataset was pre-processed and ready for analysis.
- **Feature Engineering:** Redundant columns were merged for clarity and efficiency. For example, history of heart attack, stroke, and angina were combined into a single binary variable.
- **Categorical Encoding:** Binary columns (e.g., Yes/No) were mapped to 0/1. Ordinal variables, such as general health, were numerically encoded to reflect their inherent order, and nominal variables like race/ethnicity were label encoded.

- **Standardization and Deduplication:** Column names were standardized for clarity, irrelevant features were dropped, and duplicate entries were removed to improve data quality.

After cleaning, the final dataset contained 25 variables, optimized for exploratory and predictive analysis of cardiovascular risk factors.

Exploratory Data Analysis

The exploratory phase began with a statistical overview and visual inspection of key numerical variables, including sleep hours, BMI, and physical and mental health days. The target variable, Any Cardio Event, which combines stroke, heart attack, and angina history into a binary indicator, was examined. Figure 1 illustrates a stark class imbalance, with far fewer individuals reporting cardiovascular events than those without. This imbalance presents a challenge for predictive modeling, as most algorithms may be biased toward the majority class. Addressing this would require resampling or weighting techniques to ensure fair evaluation during model training. This observation also highlighted that cardiovascular outcomes are relatively rare but critically important, underscoring the need to detect them early using nuanced data patterns. Boxplots revealed the presence of outliers, particularly in sleep hours, where some individuals reported durations exceeding 12 hours per day (Figure 2). These outliers were carefully reviewed to assess whether they reflected plausible health

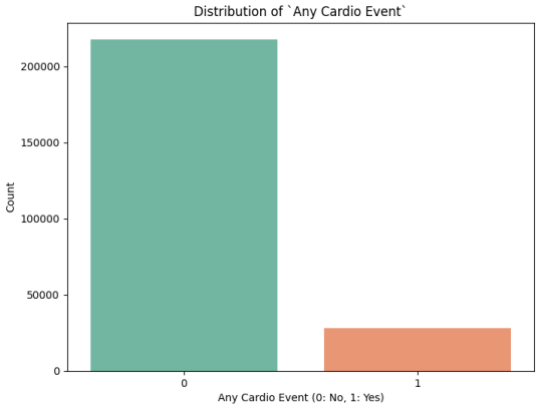


Figure 1. Distribution of Cardio Events

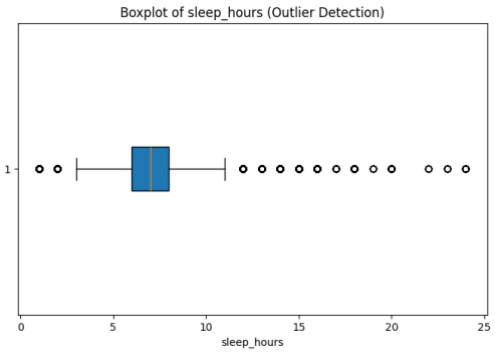


Figure 2. Box plot of sleep hours

conditions or data inconsistencies. Retaining these values was important, as they may represent edge cases relevant to heart disease risk modeling.

Correlation Analysis and Insights

To evaluate the linear relationships between variables, a correlation heatmap was generated (Figure 3). The highest correlations observed were between mental health days and depressive disorder (r = 0.42), and between arthritis and age category (r = 0.38). The target variable showed only weak to moderate correlation with individual predictors, such as age (r = 0.25), and near-zero correlation with most others. These findings reinforce the insight that no single factor can predict cardiovascular disease on its own. Instead, the condition results from the interaction of multiple health, demographic, and behavioral factors.

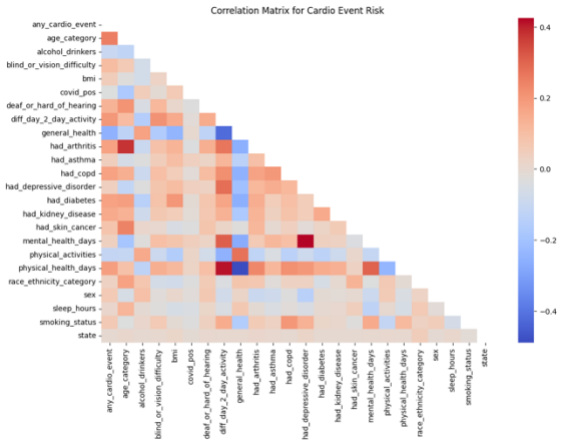


Figure 3. Correlation Matrix

This insight validated an earlier observation that cardiovascular events are multifactorial and require composite risk modeling rather than reliance on isolated indicators. For example, although age showed a modest correlation, it likely interacts with other variables such as diabetes, arthritis, or physical inactivity to increase risk. Thus, correlation analysis provided useful directional cues, but deeper statistical validation was required to establish which factors had significant influence.

Hypothesis Testing

To strengthen the understanding of potential risk factors, hypothesis testing was conducted on variables that showed relevance during exploratory analysis.

Hypothesis 1: Older individuals have a higher risk of experiencing a cardiovascular event.

The relationship between age and cardiovascular risk was explored based on clinical knowledge that heart function declines with age and that older individuals accumulate more risk factors over time. Figure 4 supports this hypothesis, with a noticeably higher median age (approximately 80) among individuals who experienced a cardiovascular event compared to those without (median age 60–64). These results provided strong evidence to reject the null hypothesis and supported the prioritization of elderly individuals for preventive screening and health education.

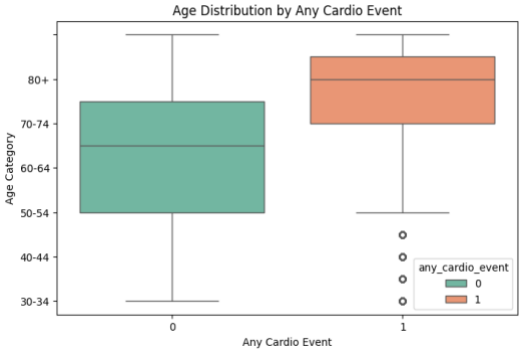


Figure 4. Box plot of age distribution by cardio events

Hypothesis 2: Higher Body Mass Index increases the likelihood of a cardiovascular event.

Obesity is widely known to contribute to cardiovascular disease through pathways such as increased blood pressure and insulin resistance. Figure 5 illustrates slightly higher BMI levels among affected individuals, but with substantial overlap between the two groups. While the difference is directionally consistent, the results suggest that BMI alone may not be a sufficient predictor. Therefore, the null hypothesis could not be confidently rejected, indicating that BMI should be considered as part of a broader risk profile.

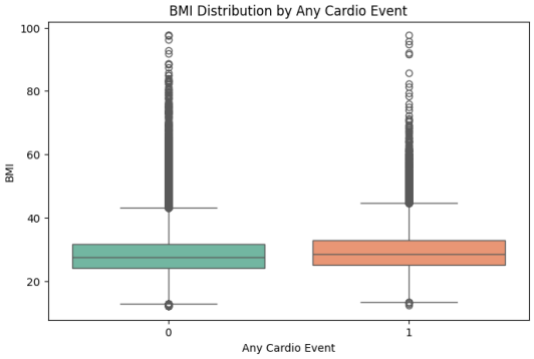


Figure 5.BMI distribution of cardio events

Hypothesis 3: Smoking and alcohol consumption increase cardiovascular event risk.

Behavioral risk factors were analyzed through the lens of vascular damage and elevated blood pressure caused by substance use. Figure 6 shows that individuals with a history of smoking or

alcohol consumption had a higher proportion of cardiovascular events compared to non-users. This supports rejection of the null hypothesis and reinforces the importance of public health policies aimed at reducing tobacco and alcohol use to minimize preventable heart disease.

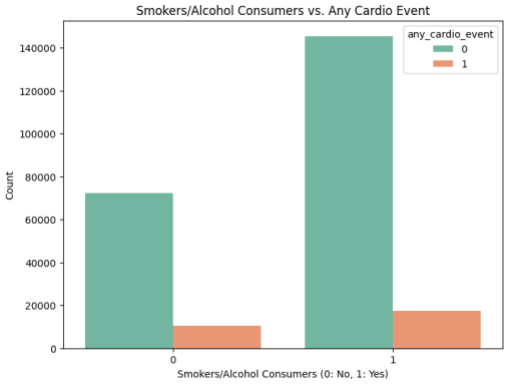


Figure 6. Smokers/Alcohol Consumers vs cardio events

Hypothesis 4: Individuals with diabetes (including gestational and pre-diabetic conditions) have a higher risk of experiencing a cardiovascular event.

Diabetes is a known contributor to heart disease, primarily due to blood vessel damage from high blood sugar levels. Figure 7 confirms this hypothesis, showing that individuals diagnosed with diabetes had a significantly higher prevalence of cardiovascular events. The null hypothesis was rejected, affirming the role of diabetes as a key modifiable risk factor and highlighting the need for early detection and management.

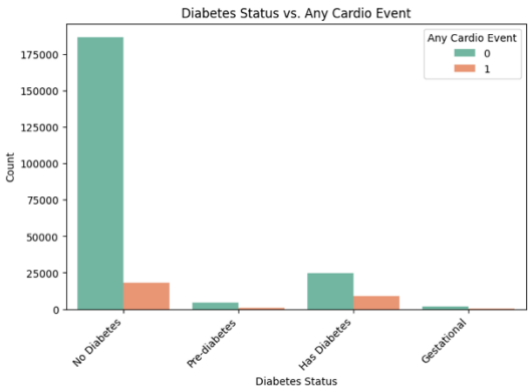


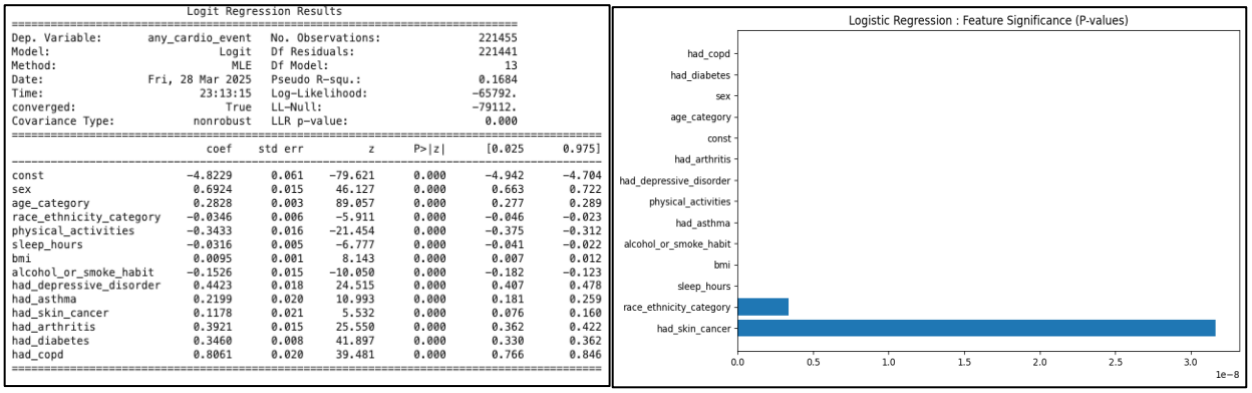
Figure 7. Diabetes vs cardio events

Together, the results of hypothesis testing confirmed that age, diabetes, smoking, and alcohol use are statistically associated with cardiovascular disease, while BMI showed only limited predictive value in isolation. These findings supported earlier patterns observed during exploratory analysis and guided the development of more sophisticated models to detect cardiovascular risk in the population.

ML Modeling

To evaluate the factors contributing to cardiovascular risk and to validate the findings from exploratory analysis, three supervised machine learning models were implemented: Logistic Regression, Random Forest, and a Random Forest model trained on Principal Component Analysis (PCA)-transformed data. These models were selected to capture both linear and non-linear patterns in the data. Due to the significant class imbalance in the target variable, under-sampling was applied to the majority class before training. Each model was trained using a 70-30 train-test split and evaluated based on accuracy, confusion matrix results, and the interpretability of their outputs. These results offer practical insights for identifying at-risk populations and enhancing public health strategies.

Model 1: Logistic Regression



Logit Regression Results

Dep. Variable:	any_cardio_event	No. Observations:	221455
Model:	Logit	Df Residuals:	221441
Method:	MLE	Df Model:	13
Date:	Fri, 28 Mar 2025	Pseudo R-squ.:	0.1684
Time:	23:13:15	Log-Likelihood:	-65792.
converged:	True	LL-Null:	-79112.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.8229	0.061	-79.621	0.000	-4.942	-4.704
sex	0.6924	0.015	46.127	0.000	0.663	0.722
age_category	0.2828	0.003	89.057	0.000	0.277	0.289
race_ethnicity_category	-0.0346	0.006	-5.911	0.000	-0.046	-0.023
physical_activities	-0.3433	0.016	-21.454	0.000	-0.375	-0.312
sleep_hours	-0.0316	0.005	-6.777	0.000	-0.041	-0.022
bmi	0.0095	0.001	8.143	0.000	0.007	0.012
alcohol_or_smoke_habit	-0.1526	0.015	-10.050	0.000	-0.182	-0.123
had_depressive_disorder	0.4423	0.018	24.515	0.000	0.407	0.478
had_asthma	0.2199	0.020	10.993	0.000	0.181	0.259
had_skin_cancer	0.1178	0.021	5.532	0.000	0.076	0.160
had_arthritis	0.3921	0.015	25.550	0.000	0.362	0.422
had_diabetes	0.3460	0.008	41.897	0.000	0.330	0.362
had_copd	0.8061	0.020	39.481	0.000	0.766	0.846

Figure 8. Logit Model Results: Significant Predictors plotted as Bar chart : (low P-Values) : copd, diabetes , Age, sex, arthritis

The logistic regression model revealed several key predictors of cardiovascular risk, with p-values providing insight into their significance. Further, the bar chart in Figure 8, was plotted to identify significant predictors based on p-values. Chronic obstructive pulmonary disease (COPD) emerged as a highly significant factor, with a p-value approaching zero ($p=0.00$), indicating a strong and

reliable association with cardiovascular events. Similarly, sex, age category, diabetes, arthritis, and depressive disorder all recorded p-values well below the conventional significance threshold ($p<0.05$), confirming their critical roles as risk factors. Protective variables, such as engagement in physical activity (p=0.91) and adequate sleep (p=0.008), also showed statistically significant relationships, supporting their influence in lowering cardiovascular risk. In contrast, some variables, including BMI (p=0.057) and alcohol or smoking habits (p=0.056), presented higher p-values that exceeded the 0.05 threshold, suggesting that their relationships with cardiovascular events are weaker or possibly confounded by other factors. The consistently low p-values of key health conditions and demographic variables underline their importance in accurately assessing cardiovascular risk within the model.

A logistic regression model was fit based on the important prectictors identified by P-values. The model achieved an accuracy of 72.71% and a cross-validation score of 72.52%, with an F1-score of 0.79. The confusion matrix revealed 6,390 true positives and 5,920 true negatives, meaning the model correctly identified a large number of individuals with and without cardiovascular events. However, it also misclassified 2,595 individuals as at-risk when they were not (false positives) and failed to identify 2,026 at-risk individuals (false negatives). While overall performance was balanced, the false negatives are of particular concern in a healthcare context, where undetected risk could delay life-saving interventions. This model is valuable for its interpretability and can assist public health professionals in understanding which conditions most influence cardiovascular risk.

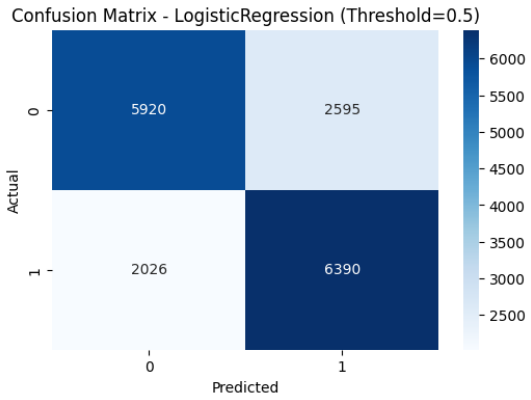


Figure 9. Confusion matrix of logistic regression

Model 2 – Random Forest

The Random Forest model was trained using the same preprocessed dataset and was particularly effective in capturing non-linear interactions between features. Feature importance analysis showed that BMI was the most influential predictor, with a score of 0.4063, followed by age category and sleep hours. Conditions like diabetes and arthritis also had substantial importance, consistent with known risk factors. Race/ethnicity had a more moderate impact, while COPD, which was the top predictor in the logistic regression model, showed relatively less importance here, suggesting that its influence may be more linear.

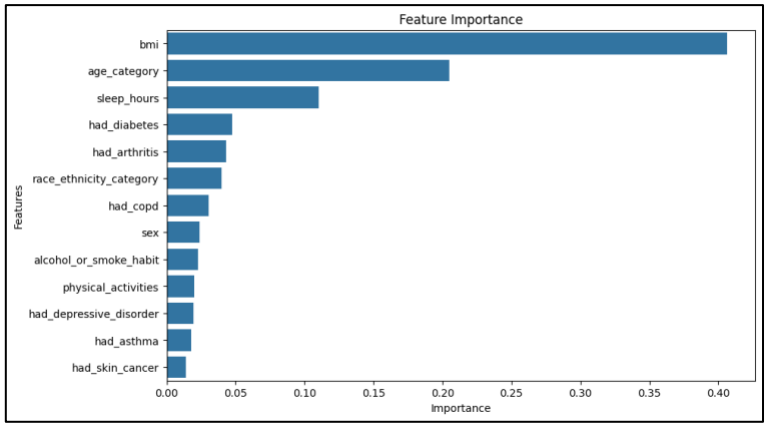


Figure 9. Feature importance of random forest

The confusion matrix(Figure 10) showed 6,005 true positives and 5,667 true negatives. However, there were 2,848 false positives and 2,411 false negatives slightly more than in the logistic regression model. The model achieved an accuracy of 68.94% and a cross-validation score of 68.55%, with an F1-score of 0.70. Although the model captured more complex relationships, its slightly higher false negative rate further emphasizes the trade-off between model complexity and reliability in identifying high-risk individuals. Despite

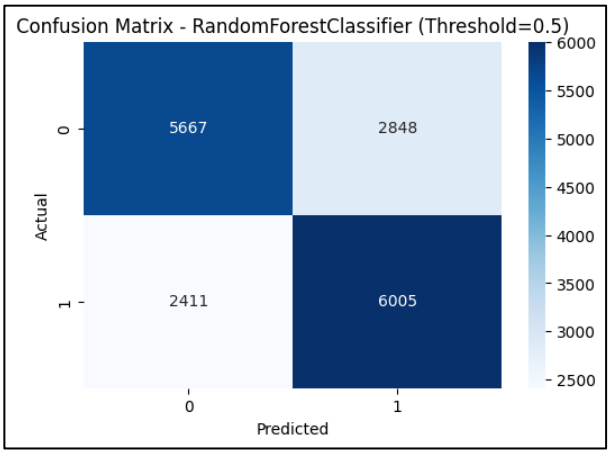


Figure 10. Confusion matrix for random forest

this, the Random Forest model offers better predictive accuracy for some variables and can complement logistic regression by identifying patterns missed in simpler models.

Model 3: PCA with Random Forest

To evaluate model performance in a reduced feature space and to address potential multicollinearity, Principal Component Analysis (PCA) was applied before training a Random Forest model. Four principal components were selected, capturing over 95% of the dataset's variance. As in Figure 12, These components were primarily composed of variables such as age category, BMI, physical health days, and mental health days. PC5, the most influential component, was driven by age and health condition metrics, emphasizing the significant role of both physical and mental well-being in cardiovascular risk. Demographic factors such as race/ethnicity and state appeared with lower influence across the components.

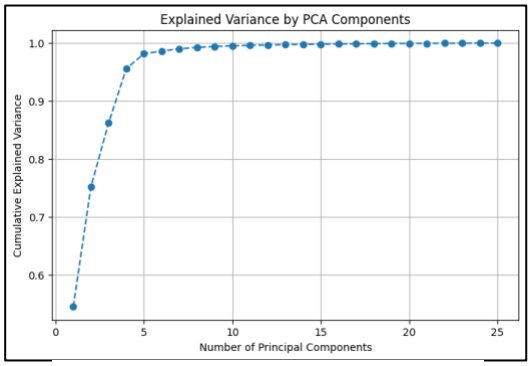


Figure 11. Variance by PCA components

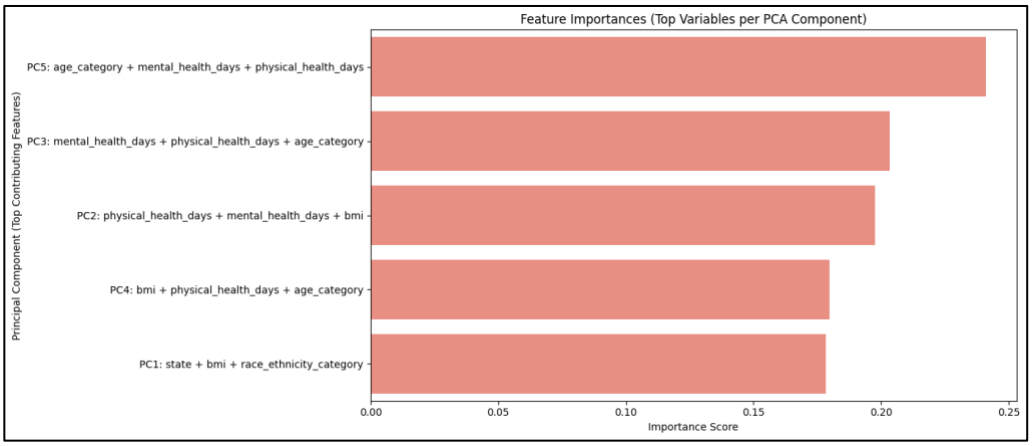


Figure 11. Top Contributing factors for PCA

Although the model trained on PCA-transformed data demonstrated improved classification accuracy on paper, its confusion matrix revealed a major limitation: it failed to correctly identify 7,829 individuals who were truly at risk, resulting in a significantly higher number of false negatives. The model achieved an accuracy of 88.03% and a cross-validation score of 68.55%, with an F1-score of 0.12. This shows that model has poor sensitivity making it unsuitable for real-world public health use, where early detection of at-risk individuals is critical.

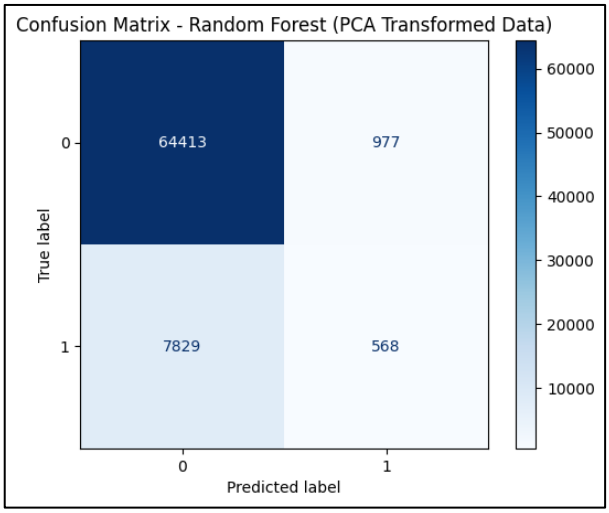


Figure 12. Confusion matrix for random forest (PCA Dataset)

The modeling outcomes demonstrate that cardiovascular risk is influenced by a combination of demographic, clinical, and behavioral factors. Key predictors such as age, BMI, diabetes, arthritis, and sleep quality consistently contributed across models. The logistic regression model offered transparency into which variables most directly influence risk, while the Random Forest model captured more complex interactions, especially for lifestyle features. In contrast, the PCA-based model, though technically accurate in classifying low-risk individuals, was less effective at identifying high-risk cases due to a high false negative rate.

These results help validate the patterns observed in exploratory analysis and highlight the importance of using a combination of linear and non-linear models to better understand cardiovascular risk. Each model provides unique insights—logistic regression for interpretability, Random Forest for complexity, and PCA for dimensionality control—collectively enriching the

overall understanding of the data. These interpretations lay the groundwork for the final section, where practical next steps and data-driven recommendations are outlined based on these findings.

Conclusion

This study applied a comprehensive data mining approach to the Heart Risk dataset, incorporating data cleaning, exploratory analysis, hypothesis testing, and machine learning models to understand the complex relationships between demographic, behavioral, and clinical factors and cardiovascular event risk.

Table 1. Model comparison

Metric	**Logistic Regression (Under-sampled)**	**Random Forest (Under-sampled)**	**Random Forest with PCA (Original Dataset)**
Accuracy	72.12%	68.94%	88.03%
Precision	0.71	0.68	0.3663
Recall	0.76	0.71	0.071
F1-Score	**0.73**	0.70	0.1214
Mean CV Score	72.52%	68.55%	87.95%

Table 1 presents a summary of predictive performance across the three models. Although the PCA-based Random Forest model achieved the highest accuracy at 88.03%, it demonstrated poor recall (0.071) and an F1-score of only 0.1214. This imbalance indicates that the model struggled to correctly identify high-risk individuals, making it less reliable for healthcare applications where minimizing false negatives is crucial. In contrast, the Logistic Regression model, with a lower accuracy of 72.12%, achieved the best F1-score (0.73) and recall (0.76), demonstrating a better balance between identifying at-risk individuals and avoiding false positives. The Random Forest model had comparable but slightly weaker performance in both precision and recall. Given this evaluation, the under-sampled Logistic Regression model is preferred for its consistent and

interpretable results, especially as its high recall minimizes false negatives, in the context of cardiovascular disease screening.

Based on the Logistic Regression coefficients and Feature Importance from the Random Forest model, age, health conditions like diabetes, pulmonary disease and lifestyle factors like physical activity, BMI, sleep hours emerged as the strongest predictors of cardiovascular risk, aligning with EDA insights.

Based on these insights, the following recommendations are proposed for public health policymakers and data owners: Health department should implement **targeted preventive programs** by introducing early screening initiatives for older adults and individuals with pre-existing conditions like diabetes and COPD to detect cardiovascular risks early and enable timely interventions. Additionally, **promoting on healthy lifestyles through public awareness campaigns**, especially for high-risk elderly group in community centers, emphasizing regular physical activity, balanced nutrition, and adequate sleep is crucial in reducing cardiovascular risk factors can support preventive efforts. Ensuring **accessibility to healthcare for high-risk individuals**, particularly in underserved or rural areas, is essential by providing affordable screenings, medications, and lifestyle coaching to mitigate risks before they escalate into severe conditions. These strategies will help reduce disparities in cardiovascular health outcomes and enhance preventive care, ultimately contributing to a healthier population with a lower incidence of heart-related diseases.

Further research could focus on developing Personalized Health Risk Assessment Tools leveraging built Logistic Regression model to create user-friendly risk calculators and clinical decision support systems. By adjusting the classification threshold (e.g., lowering it from 0.5 to 0.3) of the

current model, these tools can minimize false negatives, improving early detection and intervention. Such models can empower both individuals and healthcare professionals by providing clear insights into risk factors and enabling tailored prevention strategies. Additionally, integrating more granular behavioral data—such as dietary habits, stress levels, and healthcare access—could further enhance predictive accuracy and personalization of interventions.

In conclusion, this analysis demonstrates the value of integrating statistical modeling with real-world health data to inform actionable public health strategies. By identifying the strongest predictors of cardiovascular events and evaluating model performance with a focus on recall, this study supports more effective risk detection and prevention planning. With continued refinement and broader data integration, predictive models like those explored here can serve as critical tools for improving cardiovascular health outcomes at the population level.

Reference

1. Centers for Disease Control and Prevention. (2023). Heart disease facts and statistics. U.S. Department of Health & Human Services. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html
2. Pytlak, K. (2023). Personal key indicators of heart disease [Data set]. Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
3. Centers for Disease Control and Prevention. (2023a). Heart disease risk factors. https://www.cdc.gov/heart-disease/risk-factors/index.html
4. Centers for Disease Control and Prevention. (2023b). Diabetes and your heart. https://www.cdc.gov/diabetes/diabetes-complications/diabetes-and-your-heart.html

5. National Heart, Lung, and Blood Institute. (2023a). Smoking and cardiovascular disease. https://www.nhlbi.nih.gov/health/heart/smoking
6. National Heart, Lung, and Blood Institute. (2023b). Racial and ethnic disparities in cardiovascular health. https://pmc.ncbi.nlm.nih.gov/articles/PMC8493650/
7. National Institute on Aging. (2023). Heart health and aging. https://www.nia.nih.gov/health/heart-health/heart-health-and-aging
8. Centers for Disease Control and Prevention. (2025). Preventing chronic disease. https://www.cdc.gov/pcd/issues/2025/24_0270.htm