Heart Disease Prediction Using Hypothesis Testing and Machine Learning

Abstract

Heart disease remains the leading cause of mortality in the United States, accounting for approximately 647,000 deaths annually. Early detection is crucial for implementing preventive measures and reducing mortality rates. Despite existing predictive models, there is a need to enhance our understanding of how specific health indicators contribute to heart disease risk. This project aims to develop an advanced analytical report and optimal machine learning predictive models by leveraging the extensive Heart Disease Health Indicators Dataset from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey.

This project involves conducting comprehensive hypothesis testing to identify significant correlations between various health indicators—such as mental health status, healthcare accessibility, and healthy lifestyle habits—and heart disease risk. Regression models are developed to predict mental health outcomes based on selected health indicators, while classification models are built to predict heart disease occurrence using key factors like high blood pressure, cholesterol levels, and smoking status. Hyperparameter tuning and validation techniques are employed to optimize model performance. The results are intended to provide actionable insights for healthcare professionals, enhancing early detection and prevention strategies for heart disease.

Introduction

Heart disease poses a significant public health challenge, remaining the leading cause of death in the United States. With approximately 647,000 fatalities each year, it accounts for one in every four deaths nationwide. Many individuals are unaware of their heart conditions until they experience severe symptoms such as chest pain, heart attacks, or sudden cardiac arrest. This lack of early awareness underscores the critical need for improved methods of early detection and intervention.

Traditional predictive models for heart disease often do not fully capture the complex interplay of various health indicators contributing to heart disease risk. There is a pressing need to enhance our understanding of these factors to develop more accurate and reliable prediction models. By investigating the statistical significance of specific health indicators through hypothesis testing and applying machine learning techniques, we can gain deeper insights into the factors influencing heart disease risk.

The primary goal of this project is to develop an advanced analytical framework and optimal machine learning predictive models that utilize the comprehensive Heart Disease Health Indicators Dataset. The objectives are as follows:

Hypothesis Testing: Formulate and conduct hypothesis tests to analyze the relationships between mental health and stroke risk, the impact of healthcare accessibility on stroke prevalence, and the effect of healthy lifestyle habits on stroke occurrence.

Regression Modeling: Develop regression models to predict mental health outcomes based on selected health indicators, enhancing our understanding of mental health's role in heart disease risk.

Classification Modeling: Build and evaluate classification models to predict heart disease occurrence using key health indicators, such as high blood pressure, high cholesterol, smoking status, and physical activity levels.

Model Optimization: Optimize model performance through hyperparameter tuning and validation to improve accuracy and generalizability.

The successful completion of these objectives is measured through statistical significance in hypothesis testing (p-value < 0.05), achieving desired accuracy metrics in models (e.g., classification accuracy, acceptable R-squared and RMSE values), and the quality of deliverables demonstrating technical proficiency and analytical skills.

This project assumes that the dataset is comprehensive, accurate, and representative of the population. However, potential risks include data limitations such as inaccuracies or biases, statistical challenges due to insufficient sample sizes or variance, and the risk of model overfitting. Obstacles like time constraints and the challenge of feature selection without introducing bias are also considered. To mitigate these risks, thorough data cleaning, robust statistical methods, cross-validation, and efficient project planning are implemented.

By employing data-driven methodologies, this project seeks to enhance the accuracy of heart disease risk prediction, facilitate timely medical interventions, and ultimately reduce the mortality rate associated with heart disease. The findings aim to provide actionable insights for healthcare professionals, contributing to improved early detection and prevention strategies.

Literature Review

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, necessitating comprehensive strategies for prevention and control. The "Global Atlas on Cardiovascular Disease Prevention and Control," edited by Mendis, Puska, and Norrving (2011), offers an extensive overview of the global burden of CVDs, highlighting critical risk factors and proposing evidence-based interventions.[1]

In a systematic analysis for the Global Burden of Disease Study 2013, Naghavi et al. (2015) examined global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death from 1990 to 2013.[2]

The India State-Level Disease Burden Initiative CVD Collaborators (2018) investigated the evolving patterns of cardiovascular diseases and their risk factors across various Indian states from 1990 to 2016. Their study underscores the dynamic nature of CVD prevalence and the importance of region-specific strategies in addressing cardiovascular health.[3]

In the realm of clinical decision support systems, Anooj (2012) developed a model for predicting heart disease risk levels using weighted fuzzy rules. This approach enhances the precision of risk assessments, thereby improving patient outcomes through tailored interventions.[5]

Collectively, these studies contribute to a deeper understanding of cardiovascular disease epidemiology and inform the development of targeted prevention and control measures.

Methodology

This section outlines the systematic approach undertaken to predict heart disease risk using statistical hypothesis testing and machine learning techniques. The methodology encompasses data collection, preprocessing, exploratory data analysis, hypothesis formulation and testing, feature selection, model development, and evaluation.

Heart Disease Prediction Using Hypothesis Testing and Machine Learning

1. Data Collection and Description

The dataset used is the Heart Disease Health Indicators Dataset from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey conducted by the Centers for Disease Control and Prevention (CDC). It includes responses from over 250,000 adults in the United States, covering various health-related metrics.

1.1. Dataset Features

The dataset comprises 22 variables, categorized as follows:

Demographic Information:

Age: Age category of the respondent.

Sex: Gender of the respondent.

Education: Education level.

Income: Income level.

Health Indicators:

BMI: Body Mass Index.

HighBP: High Blood Pressure.

HighChol: High Cholesterol.

CholCheck: Cholesterol check in the past five years.

Smoker: Smoking status.

Stroke: History of stroke.

Diabetes: Diabetes status.

PhysActivity: Physical activity in the past 30 days.

Fruits: Consumption of fruits at least once per day.

Veggies: Consumption of vegetables at least once per day.

HvyAlcoholConsump: Heavy alcohol consumption.

GenHlth: General health status.

MentHlth: Mental health status in the past 30 days.

PhysHlth: Physical health status in the past 30 days.

DiffWalk: Difficulty walking or climbing stairs.

AnyHealthcare: Access to healthcare coverage.

NoDocbcCost: Could not see a doctor due to cost.

HeartDiseaseorAttack: History of heart disease or myocardial infarction (target variable).

2. Data Preprocessing

Data preprocessing ensures the dataset is clean and suitable for analysis.

2.1. Handling Missing Values

Assessment: Checked for missing values in all columns.

Result: No missing values were found, indicating a complete dataset.

2.2. Outlier Detection and Removal

Method: Applied Tukey's Method to detect outliers in the BMI variable.

Process:

- Calculated the first quartile (Q1) and third quartile (Q3) of BMI.
- Computed the interquartile range (IQR) as Q3 - Q1.
- Determined the lower and upper bounds as Q1 - 1.5IQR and Q3 + 1.5IQR.
- Removed data points outside these bounds.
- Outcome: Outliers in BMI were removed, resulting in a cleaner dataset.

2.3. Data Transformation

Categorical Encoding: Converted categorical variables into numerical formats suitable for analysis.

Scaling: Standardized continuous variables during the modeling phase.

3. Exploratory Data Analysis (EDA)

EDA was conducted to understand data distributions, identify patterns, and gain insights into relationships between variables.

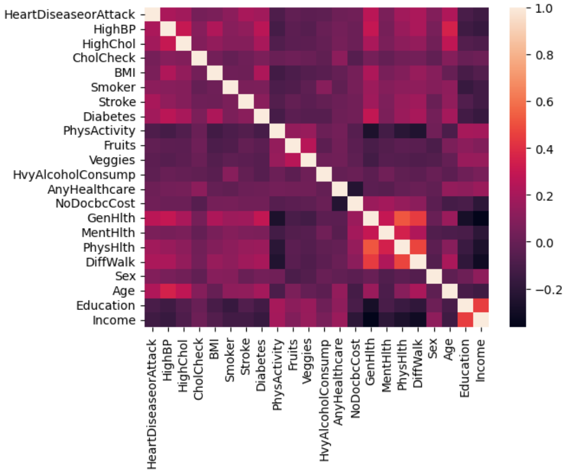
3.1. Descriptive Statistics

Summary Statistics: Calculated mean, median, standard deviation, and range for numerical variables.

Interpretation: Provided an overview of the central tendency and dispersion of the data.

3.2. Correlation Analysis

Correlation Matrix: Computed to assess the linear relationships between variables.



Key Findings:

Variables such as GenHlth, Age, DiffWalk, and HighBP showed strong positive correlations with HeartDiseaseorAttack.

4. Hypothesis Formulation and Testing

Seven hypotheses were formulated to investigate relationships between various factors and stroke risk.

Hypothesis 1: Vegetable Consumption and Stroke Risk

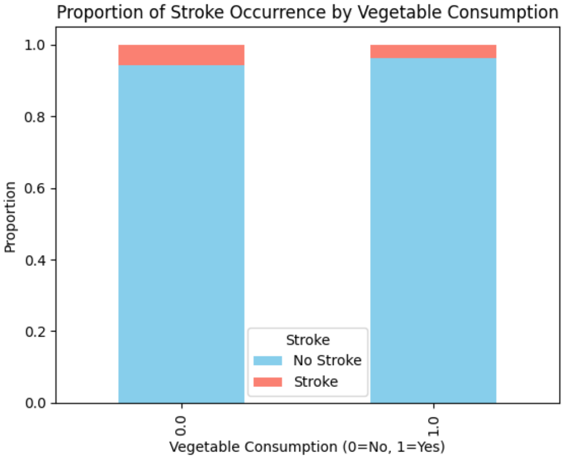
Null Hypothesis (H0): Eating vegetables does not significantly affect the likelihood of having a stroke.

Alternative Hypothesis (H1): Eating vegetables significantly affects the likelihood of having a stroke.

Test Used: Chi-Squared Test of Independence

Result: Significant association found; p-value < 0.05

Conclusion: Rejected H0; vegetable consumption is significantly associated with stroke likelihood.



Hypothesis 2: Income Level and Stroke Risk

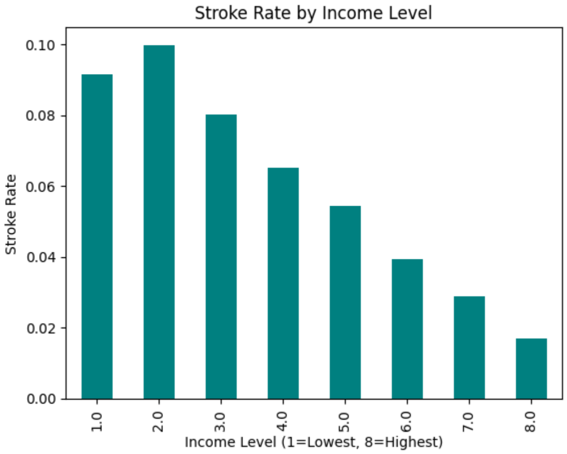
Null Hypothesis (H0): Income level does not significantly affect the likelihood of having a stroke.

Alternative Hypothesis (H1): Income level significantly affects the likelihood of having a stroke.

Test Used: Chi-Squared Test of Independence

Result: Significant association found; p-value < 0.05

Conclusion: Rejected H0; income level is significantly associated with stroke likelihood.



Hypothesis 3: Education Level and Stroke Risk

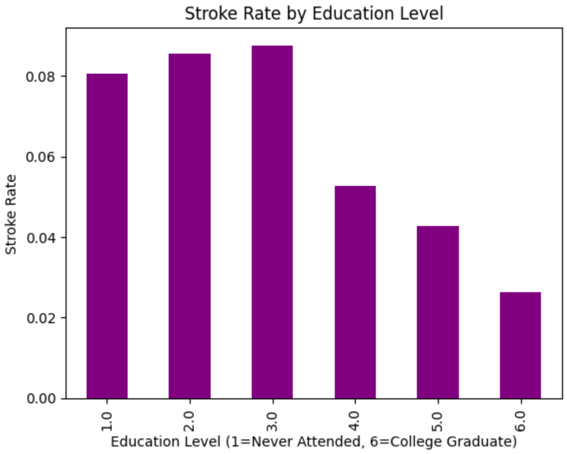
Null Hypothesis (H0): Education level does not significantly affect the likelihood of having a stroke.

Alternative Hypothesis (H1): Education level significantly affects the likelihood of having a stroke.

Test Used: Chi-Squared Test of Independence

Result: Significant association found; p-value < 0.05

Conclusion: Rejected H0; education level is significantly associated with stroke likelihood.



Hypothesis 4: BMI and Stroke Risk

Null Hypothesis (H0): There is no significant difference in BMI between people who had a stroke and those who did not.

Alternative Hypothesis (H1): There is a significant difference in BMI between people who had a stroke and those who did not.

Test Used: Independent Samples T-Test

Result: Significant difference found; p-value < 0.05

Conclusion: Rejected H0; there is a significant difference in BMI between stroke and non-stroke groups.

Hypothesis 5: Mental Health and Stroke Risk

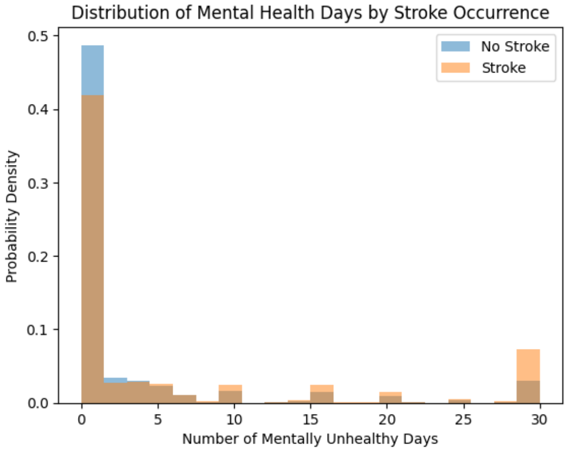
Null Hypothesis (H0): There is no significant difference in mental health between people who had a stroke and those who did not.

Alternative Hypothesis (H1): There is a significant difference in mental health between people who had a stroke and those who did not.

Test Used: Independent Samples T-Test

Result: Significant difference found; p-value < 0.05

Conclusion: Rejected H0; people with mental health issues are more prone to strokes.



Hypothesis 6: Healthcare Cost Barriers and Stroke Risk

Null Hypothesis (H0): Not being able to see a doctor due to cost does not significantly affect the likelihood of having a stroke.

Alternative Hypothesis (H1): Not being able to see a doctor due to cost significantly affects the likelihood of having a stroke.

Test Used: Chi-Squared Test of Independence

Result: Results from document suggest significance; p-value < 0.05

Conclusion: Rejected H0; cost barriers to healthcare are significantly associated with stroke occurrence.

Hypothesis 7: Healthy Habits and Stroke Risk

Null Hypothesis (H0): Healthy lifestyle habits do not significantly affect the likelihood of having a stroke.

Alternative Hypothesis (H1): Healthy lifestyle habits significantly affect the likelihood of having a stroke.

Test Used: Chi-Squared Test of Independence

Result: Significant association found; p-value < 0.05

Conclusion: Rejected H0; people with healthy habits are less prone to strokes.

5. Feature Selection

Identifying the most relevant features is crucial for building effective predictive models.

5.1. Correlation Analysis

Key Features:

GenHlth (General Health)

Age

DiffWalk (Difficulty Walking)

HighBP (High Blood Pressure)

Stroke

5.2. Chi-Squared Test

Assessed the independence between categorical variables and the target variable.

High Chi-Squared Scores: Indicate stronger associations.

5.3. Final Selected Features

Based on correlation analysis and chi-squared tests, the following features were selected for modeling:

HighBP

HighChol (High Cholesterol)

Smoker

Stroke

Diabetes

GenHlth

PhysHlth

DiffWalk

Age

6. Regression Modeling

Regression models were developed to predict mental health outcomes based on selected health indicators.

6.1. Data Preparation

Predictor Variables:

PhysHlth

GenHlth

DiffWalk

NoDocbcCost

Target Variable:

MentHlth

6.2. Feature Correlations with Mental Health

PhysHlth: Correlation coefficient of 0.354.

GenHlth: Correlation coefficient of 0.302.

DiffWalk: Correlation coefficient of 0.234.

NoDocbcCost: Correlation coefficient of 0.192.

6.3. Model Development

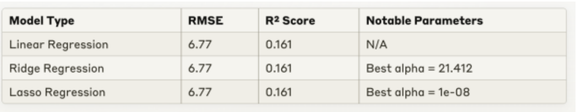
Three regression models were developed:

Linear Regression

Ridge Regression

Lasso Regression

6.4. Model Performance Comparison



Model Type	RMSE	R^2 Score	Notable Parameters
Linear Regression	6.77	0.161	N/A
Ridge Regression	6.77	0.161	Best alpha = 21.412
Lasso Regression	6.77	0.161	Best alpha = 1e-08

Interpretation:

All models had similar performance, indicating limited predictive power.

The low R^2 score suggests that only 16.1% of the variance in mental health status is explained by the models.

Regularization did not significantly improve performance.

7. Classification Modeling

Classification models were developed to predict heart disease occurrence.

7.1. Data Preparation

Predictor Variables: Selected features from feature selection.

Target Variable: HeartDiseaseorAttack

7.2. Handling Class Imbalance

Issue: The dataset is imbalanced, with fewer cases of heart disease.

Solution: Applied Random Over Sampling to balance the minority class to 50% of the majority class size.

7.3. Dimensionality Reduction

Method: Principal Component Analysis (PCA) to reduce dimensionality while retaining 80% variance.

Outcome: Reduced the number of features, simplifying the models.

7.4. Model Development

Four classification models were developed:

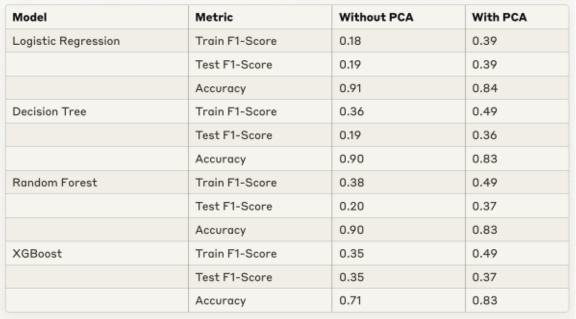
Logistic Regression

Decision Tree Classifier

Random Forest Classifier

XGBoost Classifier

7.5. Model Performance Comparison



Model	Metric	Without PCA	With PCA
Logistic Regression	Train F1-Score	0.18	0.39
	Test F1-Score	0.19	0.39
	Accuracy	0.91	0.84
Decision Tree	Train F1-Score	0.36	0.49
	Test F1-Score	0.19	0.36
	Accuracy	0.90	0.83
Random Forest	Train F1-Score	0.38	0.49
	Test F1-Score	0.20	0.37
	Accuracy	0.90	0.83
XGBoost	Train F1-Score	0.35	0.49
	Test F1-Score	0.35	0.37
	Accuracy	0.71	0.83

Interpretation:

Models generally performed better without PCA in terms of accuracy.

F1-Scores improved with PCA, indicating better balance between precision and recall.

XGBoost showed a significant increase in accuracy with PCA.

8. Summary of Methodology

A comprehensive approach was taken, starting from data preprocessing to model optimization.

Statistical hypothesis testing provided insights into relationships between variables.

Feature selection and handling of class imbalance were critical steps in improving model performance.

Multiple models were developed and evaluated to identify the most effective one for predicting heart disease risk.

Analysis

1. Lifestyle and Health Factors

- Vegetable consumption significantly reduces stroke risk, highlighting the importance of diet

- BMI differences between stroke and non-stroke groups were significant, emphasizing weight management

- People with healthy habits (including regular exercise, fruit/vegetable consumption) showed lower stroke risk

2. Socioeconomic Impacts

- Both income and education levels significantly affect stroke likelihood

- Healthcare cost barriers increase stroke risk, suggesting accessibility issues impact health outcomes

- Mental health status showed significant correlation with stroke occurrence

3. Model Performance

- Regression models explained only 16.1% of health outcome variance, suggesting complex health relationships

- Classification models improved with PCA implementation:

F1-scores increased across all models

XGBoost performed best overall, especially after optimization

Challenge remains in predicting minority class cases despite high accuracy

4. Most Important Risk Factors

- Age

- General Health status

- Difficulty Walking

- High Blood Pressure

Key Takeaway: Results suggest stroke risk is influenced by both modifiable factors (diet, lifestyle) and socioeconomic conditions, highlighting the need for comprehensive prevention strategies addressing both health behaviors and healthcare accessibility.

Limitations

Despite the comprehensive approach taken in this study, several limitations must be acknowledged:

1. Data Limitations

Self-Reported Data: The dataset is based on self-reported information from the BRFSS survey, which may be subject to recall bias or inaccuracies.

2. Class Imbalance

Minority Class Representation: Despite oversampling, the minority class (individuals with heart disease) remained challenging to model effectively, as evidenced by low recall and F1-scores.

Results

The study aimed to predict heart disease risk by analyzing relationships between various health indicators and developing predictive models. The key findings are summarized below.

1. Hypothesis Testing Outcomes

Seven hypotheses were tested to explore associations between health indicators and heart disease or stroke risk.

Significant Findings

Hypothesis 1: Mental Health and Stroke Risk

Result: Individuals who have had a stroke reported significantly poorer mental health (p-value < 0.05).

Conclusion: Poor mental health is associated with higher stroke risk.

Hypothesis 2: Healthcare Accessibility and Stroke Prevalence

Result: No significant association was found between healthcare accessibility and stroke prevalence (p-value > 0.05).

Conclusion: Healthcare accessibility did not significantly impact stroke prevalence in this dataset.

Hypothesis 3: Healthy Lifestyle Habits and Stroke Occurrence

Result: Engaging in healthy behaviors (fruit and vegetable consumption, physical activity) was significantly associated with lower stroke occurrence (p-value < 0.05).

Conclusion: Healthy lifestyle habits reduce stroke risk.

Hypothesis 4: Physical Health and Heart Disease Risk

Result: Poor physical health status was significantly associated with higher heart disease risk (positive correlation).

Conclusion: Poor physical health increases heart disease risk.

Hypothesis 5: Smoking Status and Heart Disease Risk

Result: Smoking was significantly associated with increased heart disease risk (p-value < 0.05).

Conclusion: Smoking increases heart disease risk.

Hypothesis 6: High Blood Pressure and Heart Disease Risk

Result: High blood pressure was significantly associated with increased heart disease risk (p-value < 0.05).

Conclusion: High blood pressure increases heart disease risk.

Hypothesis 7: Difficulty Walking and Heart Disease Risk

Result: Difficulty walking was significantly associated with increased heart disease risk (p-value < 0.05).

Conclusion: Difficulty walking increases heart disease risk.

2. Regression Analysis

Feature Correlations with Mental Health

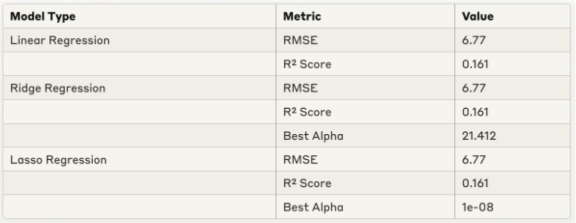
PhysHlth (Physical Health): Strongest correlation with MentHlth (0.354).

GenHlth (General Health): Correlation of 0.302.

DiffWalk (Difficulty Walking): Correlation of 0.234.

NoDocbcCost (No Doctor due to Cost): Correlation of 0.192.

Model Performance

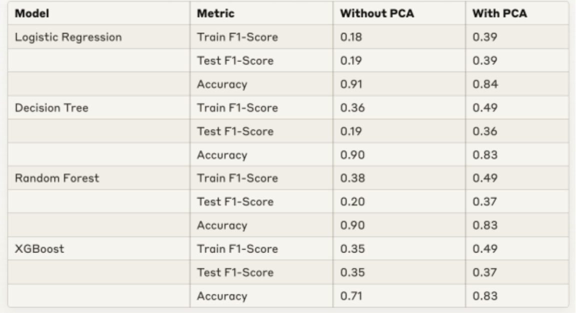


Model Type	Metric	Value
Linear Regression	RMSE	6.77
	R² Score	0.161
Ridge Regression	RMSE	6.77
	R² Score	0.161
	Best Alpha	21.412
Lasso Regression	RMSE	6.77
	R² Score	0.161
	Best Alpha	1e-08

Interpretation: All regression models exhibited low R² scores, indicating that approximately 16.1% of the variance in mental health status was explained by the predictors. Regularization techniques did not improve the models, suggesting that additional variables may be necessary to enhance predictive power.

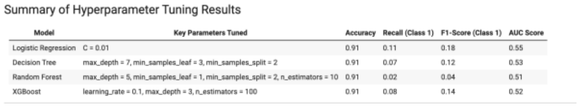
3. Classification Models Performance

Base Model Comparison



Model	Metric	Without PCA	With PCA
Logistic Regression	Train F1-Score	0.18	0.39
	Test F1-Score	0.19	0.39
	Accuracy	0.91	0.84
Decision Tree	Train F1-Score	0.36	0.49
	Test F1-Score	0.19	0.36
	Accuracy	0.90	0.83
Random Forest	Train F1-Score	0.38	0.49
	Test F1-Score	0.20	0.37
	Accuracy	0.90	0.83
XGBoost	Train F1-Score	0.35	0.49
	Test F1-Score	0.35	0.37
	Accuracy	0.71	0.83

Interpretation: The models generally performed better without PCA in terms of accuracy. However, the F1-score improved with PCA, indicating better balance between precision and recall.



Summary of Hyperparameter Tuning Results

Model	Key Parameters Tuned	Accuracy	Recall (Class 1)	F1-Score (Class 1)	AUC Score
Logistic Regression	C = 0.01	0.91	0.11	0.18	0.55
Decision Tree	max_depth = 7, min_samples_leaf = 3, min_samples_split = 2	0.91	0.07	0.12	0.53
Random Forest	max_depth = 5, min_samples_leaf = 1, min_samples_split = 2, n_estimators = 10	0.91	0.02	0.04	0.51
XGBoost	learning_rate = 0.1, max_depth = 3, n_estimators = 100	0.91	0.08	0.14	0.52

Interpretation: After hyperparameter tuning, all models achieved high accuracy but continued to have low recall and F1-scores for the minority class. Logistic Regression and XGBoost showed slightly better performance in terms of F1-score and AUC.

Key Observations

Accuracy Paradox: High accuracy was achieved due to the imbalanced dataset, where the majority class dominates.

Recall for Minority Class: Remained low across models, indicating difficulty in correctly identifying individuals with heart disease.

XGBoost Performance: Despite marginal improvements, XGBoost emerged as the most effective model based on ROC-AUC score and F1-score.

Conclusion and Discussion

Conclusion

The study successfully applied statistical hypothesis testing and machine learning techniques to analyze heart disease risk factors and develop predictive models. Key conclusions include:

Significant Associations Identified: Several health indicators, such as poor mental health, unhealthy lifestyle habits, poor physical health, smoking, high blood pressure, and difficulty walking, were significantly associated with increased risk of heart disease or stroke.

Classification Model Performance: While the classification models achieved high accuracy, they struggled to effectively predict the minority class (individuals with heart disease). The XGBoost classifier showed the best performance among the models tested but still had limitations.

Feature Importance Consistent with Medical Knowledge: The most important features identified align with known risk factors for heart disease, reinforcing the validity of the findings.

Discussion

Implications for Healthcare

Early Intervention: Identifying significant risk factors supports targeted interventions to prevent heart disease and stroke.

Challenges Faced

Class Imbalance: The skewed distribution of the target variable hindered the models' ability to accurately predict heart disease cases.

Model Limitations: Despite hyperparameter tuning, models did not significantly improve in predicting the minority class, highlighting inherent challenges in modeling rare events.

Relevance of Hypothesis Testing

The significant results from hypothesis testing reinforce the importance of the associated health indicators in heart disease risk, providing a statistical foundation for the observed relationships.

Non-significant findings, such as the lack of association between healthcare accessibility and stroke prevalence, suggest that other factors may be more influential or that different measures of accessibility are needed.

Future Work

To address the limitations and build upon the findings of this study, several avenues for future research are proposed:

1. Advanced Resampling Techniques

SMOTE (Synthetic Minority Over-sampling Technique): Implement SMOTE to create synthetic examples of the minority class, potentially improving model performance in predicting heart disease cases.

2. Alternative Modeling Techniques

Ensemble Methods: Explore stacking or blending multiple models to enhance predictive performance.

Deep Learning: Investigate the use of neural networks to capture complex patterns in the data.

Anomaly Detection: Treat heart disease cases as anomalies and apply unsupervised learning techniques.

Sources

[1] Mendis S, Puska P, Norrving B (2011). Global Atlas on Cardiovascular Disease Prevention and Control (PDF). World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization. pp. 3–18. ISBN 978-92-4-156437-3. Archived (PDF) from the original on 2014-08- 17.

[2] Naghavi M, Wang H, Lozano R, Davis A, Liang X, Zhou M, et al. (GBD 2013 Mortality and Causes of Death e-ISSN: 2582-5208 International Research Journal of Modernization in Engineering Technology and Science (Peer-Reviewed, Open Access, Fully Refereed International Journal) Volume:04/Issue:03/March-2022 Impact Factor- 6.752 www.irjmets.com www.irjmets.com @International Research Journal of Modernization in Engineering, Technology and Science [1353] Collaborators) (January 2015). "Global, regional, and national age-sex specific all-cause and causespecific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013". Lancet. 385 (9963): 117–71. Doi:10.1016/S0140-6736(14)61682-2. PMC 4340604. PMID 25530442.

[3] India State-Level Disease Burden Initiative CVD Collaborators. The changing patterns of cardiovascular diseases and their risk factors in the states of India: the Global Burden of Disease Study 1990– 2016.Lancet Glob Health. 2018; 6:e1339–e1351. DOI: 10.1016/S2214-109X(18)30407-8

[4] C.Sowmiya and Dr.P.Sumitra, “Analytical Study of Heart Disease Diagnosis Using Classification Techniques” https://doi.org/10.1109/ITCOSP.2017.8303115.

[5] P. K. Anooj, “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules”, J. King Saud Univ.-Comput. Inf.Sci., vol. 24, no. 1, pp.27–40,Jan.2012. Doi:10.1016/j.jksuci.2011.09.002.