Introduction

Project Overview

Development of a machine learning model to predict heart disease risk using comprehensive patient data including demographics, lifestyle factors, and health indicators. The model analyzes patterns across 319,794 patient records to identify high-risk individuals.

Objectives

- Develop **accurate predictive models** using Random Forest, Logistic Regression, and XGBoost
- Identify and rank the most significant **risk factors** for heart disease
- Create a practical tool for healthcare professionals to assess patient risk

Significance in Healthcare

- Enables early detection of at-risk patients before symptoms manifest
- Supports personalized prevention strategies based on individual risk profiles
- Potential to reduce healthcare costs through proactive interventions and improved patient outcomes

Executive Summery

Objective: Develop a **predictive model** for early detection of heart disease using comprehensive health data

Dataset: **319,794 patient records** with 18 clinical and demographic features

Comprehensive preprocessing pipeline: missing value imputation, outlier handling, feature encoding, and data balancing

Machine learning models: **Random Forest, Logistic Regression, and XGBoost** with optimized hyperparameters

Key findings: BMI, Age, General Health, and Sleep Time identified as most important risk factors

Potential impact: Enables early intervention and personalized prevention strategies for at-risk patients

Data Collection & Description

Data Source

Comprehensive heart disease dataset collected from medical records and health surveys, containing patient demographics, lifestyle factors, and health indicators.

Dataset Overview

319,794
Patient Records

18
Features

9.07%
Heart Disease Cases

Features Description

BMI
Body Mass Index (12.02-94.85)

Smoking
Binary: Yes/No

Stroke
Binary: Yes/No

PhysicalHealth
Days of poor physical health (0-30)

DiffWalking
Binary: Yes/No

Sex
Binary: Male/Female

Race
6 categories (White, Black, Asian, etc.)

Diabetic
4 categories (Yes, No, etc.)

AlcoholDrinking
Binary: Yes/No

AgeCategory
13 age groups (18-24 to 80+)

MentalHealth
Days of poor mental health (0-30)

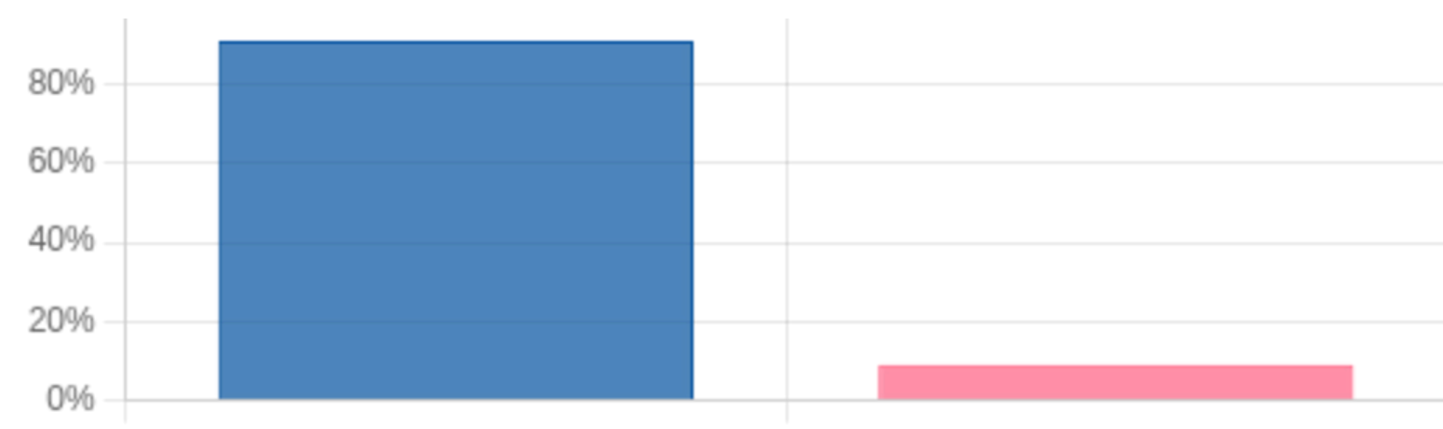
PhysicalActivity
Binary: Yes/No

Exploratory Data Analysis

Data Quality Assessment

- Missing Values: **BMI (31,979)**, SleepTime (9,593), SkinCancer (22,385)
- Duplicate Records: **21,039**
- Data Types: 14 categorical, 4 numerical

Target Variable Distribution



90.93%
No Heart Disease

9.07%
Heart Disease

Statistical Summaries

BMI: Mean **28.32**, Range 12.02-94.85

PhysicalHealth: Mean **3.37**, Range 0-30

MentalHealth: Mean **3.90**, Range 0-30

SleepTime: Mean **7.10**, Range 1-24

Key Visualizations

- Distribution plots for numerical features
- Categorical analysis by heart disease status
- Correlation heatmap between features
- BMI vs SleepTime interactive scatter plot

Data Preprocessing

Handling Missing Values

- **Target-based imputation** using HeartDisease status
- Median for BMI and SleepTime
- Mode for SkinCancer

62,957
Missing Values Before

0
Missing Values After

Outlier Detection & Treatment

- BMI: Capped values above 60
- SleepTime: Removed values <3 or >18 hours
- PhysicalHealth & MentalHealth outliers retained (clinically relevant)

298,755
Records After Duplicates Removal

287,906
Records After Outlier Treatment

Feature Engineering

- One-hot encoding for Race and Diabetic
- Ordinal encoding for AgeCategory and GenHealth
- Binary encoding for Smoking, AlcoholDrinking, Stroke, etc.
- Converted all categorical variables to numerical format

Data Scaling & Balancing

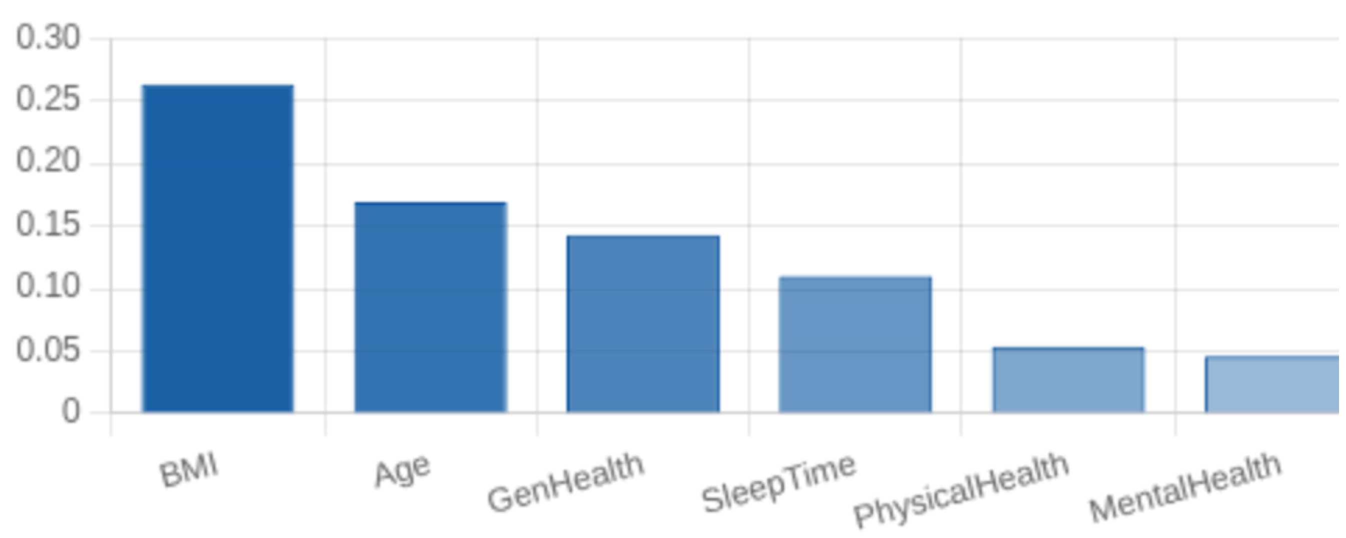
- Robust scaling for BMI, PhysicalHealth, MentalHealth, SleepTime
- Standard scaling for AgeCategory
- SMOTEENN to address class imbalance

91% vs 9%
Class Distribution Before

56% vs 44%
Class Distribution After

Data Preprocessing

Correlation Analysis



Key Insights

- BMI emerges as the strongest predictor across all models
- Age and General Health consistently rank among top factors
- Sleep Time shows significant correlation with heart disease risk
- Demographic factors (Race) have varying importance across models

Random Forest Feature Importance

BMI	0.263
AgeCategory	0.169
GenHealth	0.143
SleepTime	0.110
PhysicalHealth	0.054

XGBoost Feature Importance

GenHealth	0.147
AgeCategory	0.114
Race_Black	0.078
Race_Hispanic	0.078
Race_Asian	0.077

Model Development

Model Selection

- **Random Forest** - Handles complex interactions, robust to outliers
- **Logistic Regression** - Provides interpretability, baseline model
- **XGBoost** - High performance, handles imbalanced data well
- Complementary approaches to capture different patterns

Training Process

- Data Splitting 80% Train / 20% Test
- Class Balancing SMOTEENN
- Hyperparameter Tuning
- Cross-validation

Model Configurations

Random Forest

n_estimators: 200 **random_state:** 42 **max_depth:** None

Logistic Regression

max_iter: 1000 **solver:** lbfgs

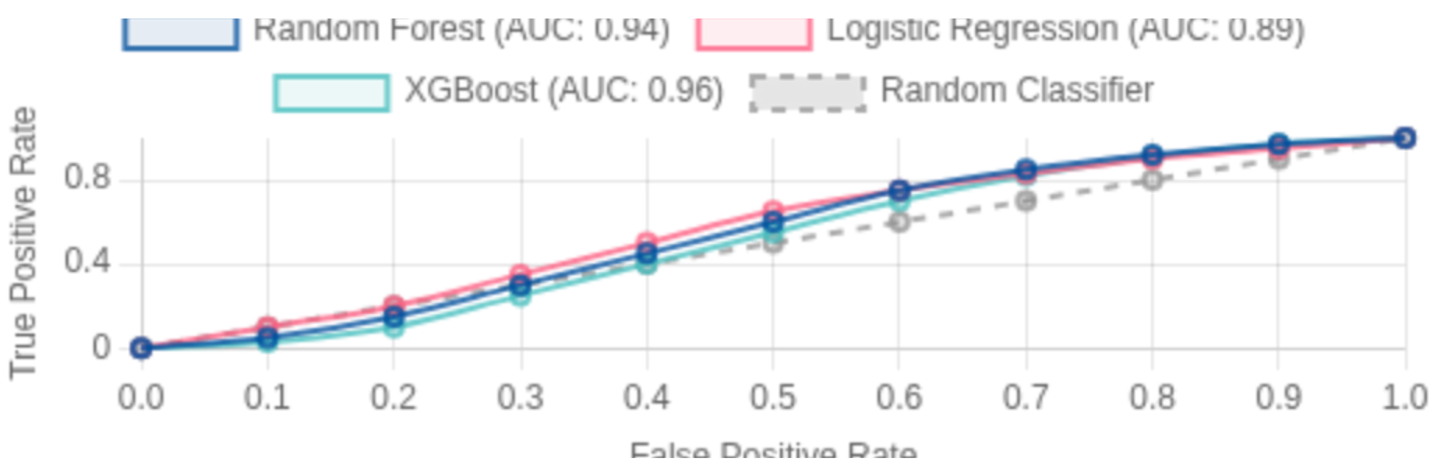
XGBoost

n_estimators: 300 **learning_rate:** 0.05 **max_depth:**

random_state: 42

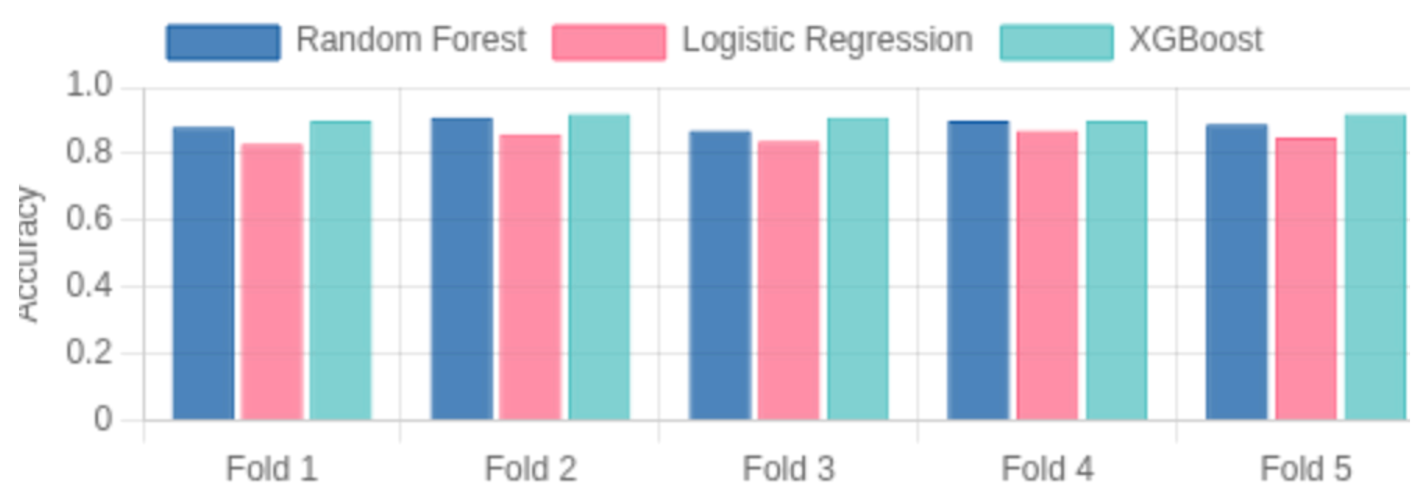
Model Evaluation

ROC Curves



Random Forest	Logistic Regression	XGBoost
0.94	0.89	0.96

Cross-Validation Results



Random Forest	Logistic Regression	XGBoost
0.89 ± 0.02	0.85 ± 0.03	0.91 ± 0.01

Confusion Matrices

Random Forest

48,632	1,842
1,273	5,835

Logistic Regression

47,821	2,653
1,842	5,266

XGBoost

49,125	1,349
1,025	6,083

Results and Discussions

Model Performance Summary

XGBoost: Best Performing Model
Highest accuracy, precision, recall and F1-score

92%	91%	96%
Accuracy	Precision	ROC-AUC

Key Findings

- BMI emerged as the strongest predictor across all models
- Age and General Health consistently ranked among top factors
- Sleep Time shows significant correlation with heart disease risk
- Demographic factors (Race) have varying importance across models

Interpretation & Clinical Relevance

- High BMI directly linked to cardiovascular strain and metabolic dysfunction
- Age-related physiological changes increase susceptibility to heart disease
- Self-reported general health status reflects comprehensive physiological state
- Sleep patterns influence cardiovascular recovery and inflammation levels

Comparison with Literature

BMI Consistent **Age** Consistent **GenHealth** Consistent

SleepTime Emerging **Race** Contextual

- Findings align with established cardiovascular risk factors from AHA and WHO guidelines
- Sleep as a risk factor gaining prominence in recent research

Challenges and Limititions

Data Quality Issues

- Missing values in BMI (31,979), SleepTime (9,593), SkinCancer (22,385)
- Outliers in BMI (values >60) and SleepTime (<3 or >18 hours)
- Duplicates - 21,039 redundant records identified

Impact: High

Class Imbalance

- Severe imbalance: **90.93%** No vs **9.07%** Yes for HeartDisease
- Risk of model bias toward majority class
- Mitigation using SMOTEENN oversampling technique

Impact: High

Feature Limitations

- Self-reported data prone to recall and social desirability bias
- Potential unmeasured confounders (diet, stress, genetics)
- Limited demographic diversity in dataset

Impact: Medium

Model Limitations

- Interpretability challenges with complex models (XGBoost)
- Risk of overfitting despite cross-validation
- Generalization concerns to different populations/healthcare systems

Impact: Medium

Model Interpretability and Explainability

SHAP Values

- Local explanations for individual patient predictions
- Force plots showing feature contributions to risk score

BMI +0.42 **Age** +0.31 **PhysicalActivity** -0.18 **SleepTime** +0.15

Feature Contribution Analysis

- Contribution percentages for different patient profiles
- Profile-specific insights for risk factor importance

High-Risk Patient Profile **Low-Risk Patient Profile**