Predictive Analysis of Cardiovascular Health Through Machine Learning

Utilizing Heart Rate Data for Early Disease Detection

BEDOUR AHMAD ALDALBAHI

Table of Contents

Table Of Figures

Abstract

Cardiovascular diseases are a leading cause of morbidity and mortality worldwide, emphasizing the need for early detection and preventive healthcare strategies. The maximum heart rate achieved during exercise, denoted as 'heartRate', serves as a pivotal indicator of an individual's cardiovascular health. Accurately predicting this parameter can facilitate the early identification of at-risk individuals, enabling timely medical intervention. This project is anchored in the utilization of machine learning algorithms to predict 'heartRate' by analyzing a range of physiological and clinical features. It delineates the application of Linear Regression, Random Forest, and Gradient Boosting models, each assessed for their predictive performance. The study addresses various challenges inherent in ML applications within healthcare, such as data imbalance, feature selection, and model interpretability, providing a comprehensive analysis of the models' efficacy in predicting cardiovascular health indicators.

Introduction

This project explores the application of machine learning (ML) algorithms to predict the maximum heart rate achievable (heartRate) by individuals, a crucial metric for assessing cardiovascular health. Utilizing various physiological and clinical features, the study employs three prominent ML models: Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. Each model's performance is evaluated based on Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) metrics across two datasets. The project navigates through challenges such as imbalanced data, high dimensionality, and computational constraints to optimize model accuracy and reliability. The findings demonstrate the significant potential of ML in enhancing predictive accuracy and offer insights into cardiovascular health assessment, aiming to support early detection and personalized healthcare interventions.

Project Description

The project delves into the realm of predictive healthcare, focusing on estimating the maximum heart rate achievable (heartRate) during physical exercise, a vital sign of cardiovascular health. Leveraging machine learning (ML) techniques, the project aims to predict 'heartRate' based on a comprehensive set of physiological and clinical features, thereby facilitating the early identification of potential heart-related health risks.

Objective:

The primary objective is to utilize ML algorithms to model and predict the 'heartRate' accurately. This endeavor is not just about predicting a number but about understanding the intricate relationships between various health indicators and how they collectively influence cardiovascular performance. The ultimate goal is to provide a tool that can aid healthcare professionals in preemptively identifying individuals at higher risk of cardiovascular diseases.

Benefits:

Facilitates early detection of heart disease, improving treatment outcomes.
Optimizes healthcare resource allocation by identifying high-risk individuals.
Advances research in applying machine learning to healthcare diagnostics.

Hardware/Software Requirements

System Specifications: A computer system with at least 16 GB RAM and 256/512 GB SSD.
Tools and Libraries: Google Colab, Scikit Learn, Pandas, NumPy, Matplotlib/Seaborn.

Task 1 – Description of the problem - machine learning libraries and packages – experimental setup.

Problem Description

This project aims at leveraging machine learning (ML) to predict the maximum heart rate achieved (thalach) by individuals, which is a critical measure in assessing heart disease risk. By analyzing various physiological and clinical features alongside 'thalach', the project seeks to enable the early identification of individuals at risk for heart disease, potentially facilitating timely intervention and management.

Machine Learning Libraries

- pandas for data manipulation and ingestion.
- numpy for numerical operations.
- matplotlib.pyplot and seaborn for visualizing the data, which aids in understanding the distributions of various features and the target variable.
- sklearn for its comprehensive suite of tools for machine learning, including:
- train_test_split and GridSearchCV for splitting the data and optimizing model parameters.
- mean_absolute_error, mean_squared_error, and r2_score for evaluating model performance.
- Various machine learning models like SVR (Support Vector Regression), LinearRegression, and KNeighborsRegressor for building predictive models.
- StandardScaler and MinMaxScaler for feature scaling to normalize the dataset and improve model training.
- time to track the duration of model training and evaluation processes.
- ZipFile for extracting datasets compressed in ZIP format, facilitating the use of data directly from sources like Kaggle.

Experimental Setup

- The experimental workflow includes:
- **Data Preprocessing:**
- Normalizing features using MinMaxScaler and splitting the dataset into training and testing sets.
- **Model Training and Tuning:**
- Training models like LinearRegression, MLPRegressor, and RandomForestClassifier on the training set. Hyperparameters are optimized through grid search or randomized search.
- **Evaluation:**
- Using MSE, MAE, and R2 to evaluate model performance on the testing set.

Task 2: Dataset Description and Data Processing

First Dataset

Dataset Description

The heart disease dataset encompasses a broad array of physiological and clinical features associated with heart disease. With a total of 1027 entries, the dataset provides a comprehensive overview of patient characteristics and their correlation with heart disease presence. The features include:

- **Demographic Factors:**
  - Age: Patient's age in years.
  - Sex (0 = female, 1 = male).

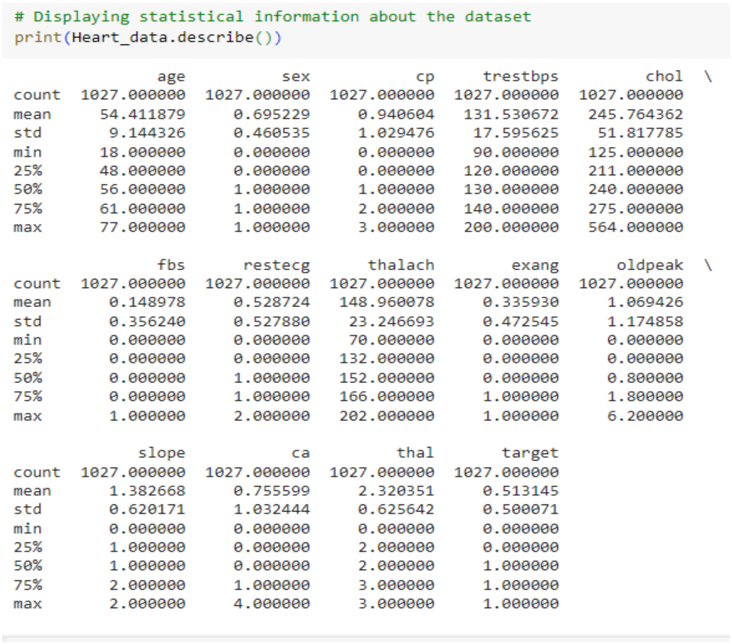
- **Behavioral Factors:**
  - Cp (Chest Pain Type): Serum cholesterol in mg/dl.

- **Medical History Factors:**
  - Fbs (Fasting Blood Sugar): Fasting blood sugar > 120 mg/dl (0 = false,1 = true).
  - Restecg (Resting Electrocardiographic results) Categorizes resting electrocardiographic results into 3 types.

- **Medical Current Conditions:**
  - Trestbps (Resting Blood Pressure)
  - Chol : Serum cholesterol in mg/dl.
  - Thalach (Maximum Heart Rate Achieved) **:**Highest heart rate achieved.
  - Exang (Exercise Induced Angina) :Exercise-induced chest pain (0 =no,1= yes).
  - Oldpeak: ST depression induced by exercise relative to rest.
  - Slope : **:** The slope of the peak exercise ST segment.
  - Ca : The number of major vessels (0-4) colored by fluoroscopy.
  - Thal : Results of the thallium stress test indicating probable heart disease (values 0-3).

    DatasetSource: : https://www.kaggle.com/datasets/faresabbasai2022/heart-diseases-prediction-with-streamlit/data

Statistical Information

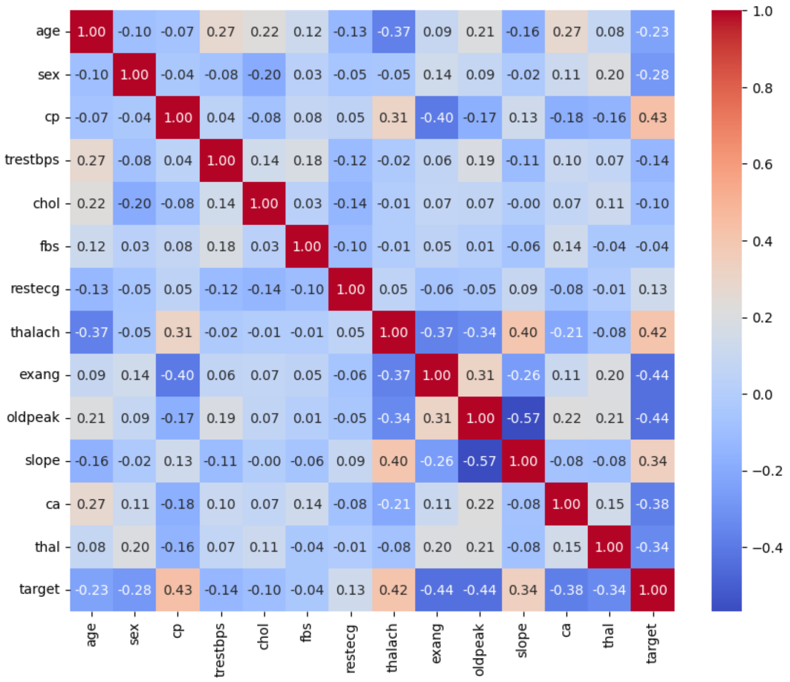


```
# Displaying statistical information about the dataset
print(Heart_data.describe())
```

```
              age          sex           cp     trestbps         chol  \
count  1027.000000  1027.000000  1027.000000  1027.000000  1027.000000
mean     54.411879     0.695229     0.940604   131.530672   245.764362
std       9.144326     0.460535     1.029476    17.595625    51.817785
min      18.000000     0.000000     0.000000    90.000000   125.000000
25%      48.000000     0.000000     0.000000   120.000000   211.000000
50%      56.000000     1.000000     1.000000   130.000000   240.000000
75%      61.000000     1.000000     2.000000   140.000000   275.000000
max      77.000000     1.000000     3.000000   200.000000   564.000000

              fbs      restecg      thalach        exang      oldpeak  \
count  1027.000000  1027.000000  1027.000000  1027.000000  1027.000000
mean      0.148978     0.528724   148.960078     0.335930     1.069426
std       0.356240     0.527880    23.246693     0.472545     1.174858
min       0.000000     0.000000    70.000000     0.000000     0.000000
25%       0.000000     0.000000   132.000000     0.000000     0.000000
50%       0.000000     1.000000   152.000000     0.000000     0.800000
75%       0.000000     1.000000   166.000000     1.000000     1.800000
max       1.000000     2.000000   202.000000     1.000000     6.200000

            slope           ca         thal       target
count  1027.000000  1027.000000  1027.000000  1027.000000
mean      1.382668     0.755599     2.320351     0.513145
std       0.620171     1.032444     0.625642     0.500071
min       0.000000     0.000000     0.000000     0.000000
25%       1.000000     0.000000     2.000000     0.000000
50%       1.000000     0.000000     2.000000     1.000000
75%       2.000000     1.000000     3.000000     1.000000
max       2.000000     4.000000     3.000000     1.000000
```

Figure 1Frist Dataset Statistical Information

Correlation Analysis



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.10	-0.07	0.27	0.22	0.12	-0.13	-0.37	0.09	0.21	-0.16	0.27	0.08	-0.23
sex	-0.10	1.00	-0.04	-0.08	-0.20	0.03	-0.05	-0.05	0.14	0.09	-0.02	0.11	0.20	-0.28
cp	-0.07	-0.04	1.00	0.04	-0.08	0.08	0.05	0.31	-0.40	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.08	0.04	1.00	0.14	0.18	-0.12	-0.02	0.06	0.19	-0.11	0.10	0.07	-0.14
chol	0.22	-0.20	-0.08	0.14	1.00	0.03	-0.14	-0.01	0.07	0.07	-0.00	0.07	0.11	-0.10
fbs	0.12	0.03	0.08	0.18	0.03	1.00	-0.10	-0.01	0.05	0.01	-0.06	0.14	-0.04	-0.04
restecg	-0.13	-0.05	0.05	-0.12	-0.14	-0.10	1.00	0.05	-0.06	-0.05	0.09	-0.08	-0.01	0.13
thalach	-0.37	-0.05	0.31	-0.02	-0.01	-0.01	0.05	1.00	-0.37	-0.34	0.40	-0.21	-0.08	0.42
exang	0.09	0.14	-0.40	0.06	0.07	0.05	-0.06	-0.37	1.00	0.31	-0.26	0.11	0.20	-0.44
oldpeak	0.21	0.09	-0.17	0.19	0.07	0.01	-0.05	-0.34	0.31	1.00	-0.57	0.22	0.21	-0.44
slope	-0.16	-0.02	0.13	-0.11	-0.00	-0.06	0.09	0.40	-0.26	-0.57	1.00	-0.08	-0.08	0.34
ca	0.27	0.11	-0.18	0.10	0.07	0.14	-0.08	-0.21	0.11	0.22	-0.08	1.00	0.15	-0.38
thal	0.08	0.20	-0.16	0.07	0.11	-0.04	-0.01	-0.08	0.20	0.21	-0.08	0.15	1.00	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.10	-0.04	0.13	0.42	-0.44	-0.44	0.34	-0.38	-0.34	1.00

Figure 2 Frist Dataset Correlation Analysis

An integral part of our data exploration was to conduct a correlation analysis to assess the relationships between different features and the target variable. The correlation matrix heatmap is a powerful visual tool that showcases how each attribute is related to another. The values range from -1 to 1, where:

- 1 indicates a strong positive correlation.
- -1 indicates a strong negative correlation.
- 0 suggests no correlation.

Key Insights from the Correlation Matrix

- Target Correlation: The 'cp' (chest pain type) and 'thalach' (maximum heart rate achieved) show a notable positive correlation with the target variable, suggesting that these features are important indicators for the presence of heart disease.
- Cholesterol Level: Surprisingly, 'chol' (serum cholesterol) has a weak correlation with the target, indicating that higher cholesterol levels are not as strongly associated with heart disease in this dataset as might be expected.
- Age Factor: Age shows a negative correlation with 'thalach' and a positive one with 'oldpeak' (ST depression), which could signify that older patients tend to have lower maximum heart rates and more exercise-induced ST depression.

Second Dataset

- ## Dataset Description

- The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD).The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

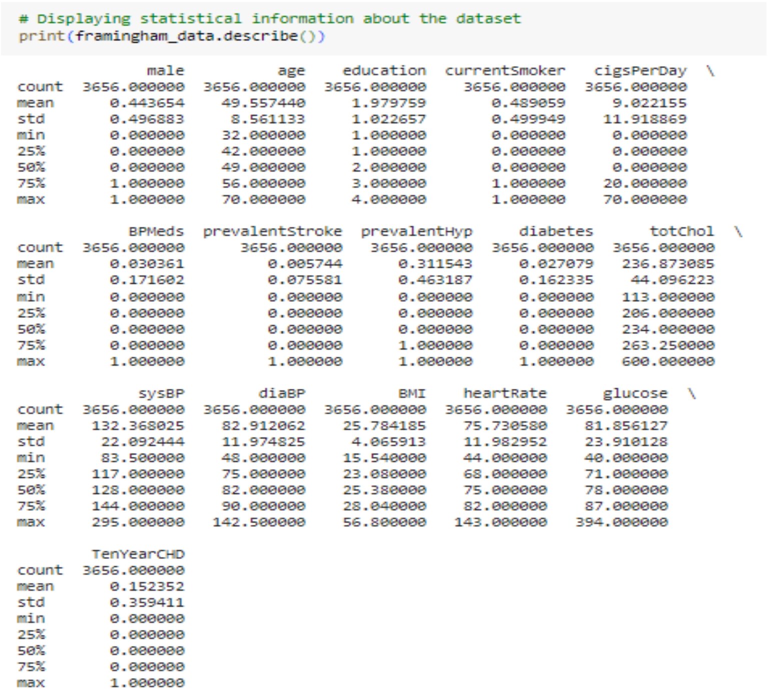
- **Demographic:**
  - Sex: male or female
  - Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

- **Behavioral**
  - Current Smoker: whether or not the patient is a current smoker (Nominal)
  - Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

- **Medical(history)**
  - BP Meds: whether or not the patient was on blood pressure medication (Nominal)
  - Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
  - Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
  - Diabetes: whether or not the patient had diabetes (Nominal)
  -

- **Medical(current)**
  - Tot Chol: total cholesterol level (Continuous)
  - Sys BP: systolic blood pressure (Continuous)
  - Dia BP: diastolic blood pressure (Continuous)
  - BMI: Body Mass Index (Continuous)
  - Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
  - Glucose: glucose level (Continuous)
    DatasetSource:
    (https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression

Statistical Information

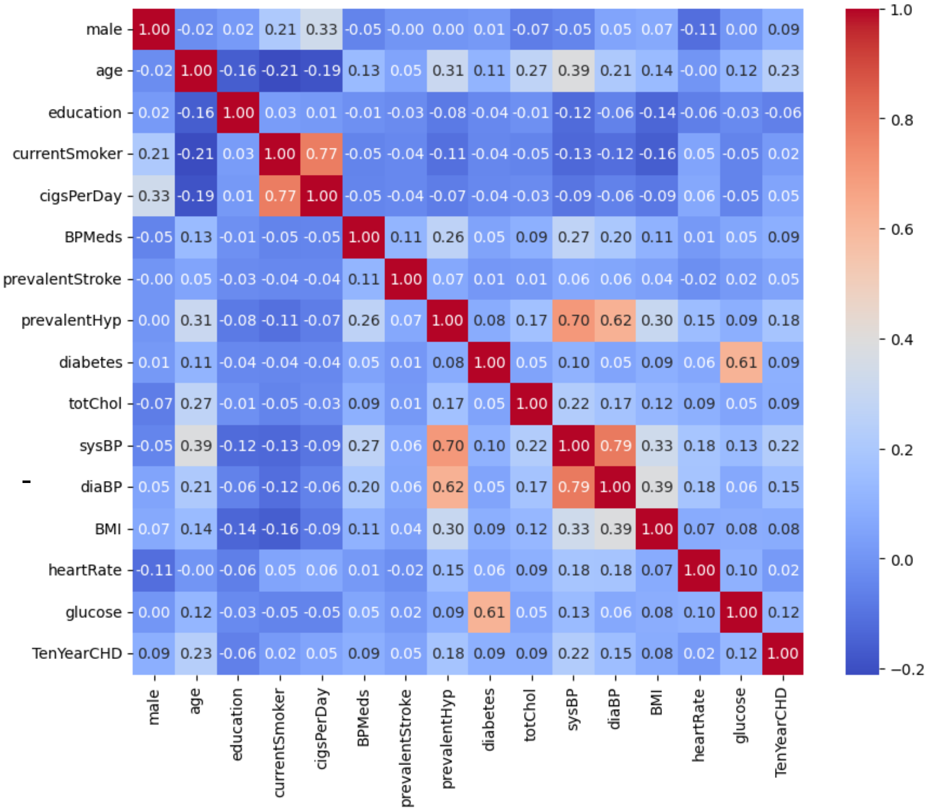


```
# Displaying statistical information about the dataset
print(framingham_data.describe())
```

```
              male          age    education  currentSmoker   cigsPerDay  \
count  3656.000000  3656.000000  3656.000000    3656.000000  3656.000000
mean      0.443654    49.557440     1.979759       0.489059     9.022155
std       0.496883     8.561133     1.022657       0.499949    11.918869
min       0.000000    32.000000     1.000000       0.000000     0.000000
25%       0.000000    42.000000     1.000000       0.000000     0.000000
50%       0.000000    49.000000     2.000000       0.000000     0.000000
75%       1.000000    56.000000     3.000000       1.000000    20.000000
max       1.000000    70.000000     4.000000       1.000000    70.000000

            BPMeds  prevalentStroke  prevalentHyp     diabetes      totChol  \
count  3656.000000      3656.000000   3656.000000  3656.000000  3656.000000
mean      0.030361         0.005744      0.311543     0.027079   236.873085
std       0.171602         0.075581      0.463187     0.162335    44.096223
min       0.000000         0.000000      0.000000     0.000000   113.000000
25%       0.000000         0.000000      0.000000     0.000000   206.000000
50%       0.000000         0.000000      0.000000     0.000000   234.000000
75%       0.000000         0.000000      1.000000     0.000000   263.250000
max       1.000000         1.000000      1.000000     1.000000   600.000000

             sysBP        diaBP          BMI    heartRate      glucose  \
count  3656.000000  3656.000000  3656.000000  3656.000000  3656.000000
mean    132.368025    82.912062    25.784185    75.730580    81.856127
std      22.092444    11.974825     4.065913    11.982952    23.910128
min      83.500000    48.000000    15.540000    44.000000    40.000000
25%     117.000000    75.000000    23.080000    68.000000    71.000000
50%     128.000000    82.000000    25.380000    75.000000    78.000000
75%     144.000000    90.000000    28.040000    82.000000    87.000000
max     295.000000   142.500000    56.800000   143.000000   394.000000

        TenYearCHD
count  3656.000000
mean      0.152352
std       0.359411
min       0.000000
25%       0.000000
50%       0.000000
75%       0.000000
max       1.000000
```

Figure 3 -SecDataset Statistical Information

Correlation Analysis



	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
male	1.00	-0.02	0.02	0.21	0.33	-0.05	-0.00	0.00	0.01	-0.07	-0.05	0.05	0.07	-0.11	0.00	0.09
age	-0.02	1.00	-0.16	-0.21	-0.19	0.13	0.05	0.31	0.11	0.27	0.39	0.21	0.14	-0.00	0.12	0.23
education	0.02	-0.16	1.00	0.03	0.01	-0.01	-0.03	-0.08	-0.04	-0.01	-0.12	-0.06	-0.14	-0.06	-0.03	-0.06
currentSmoker	0.21	-0.21	0.03	1.00	0.77	-0.05	-0.04	-0.11	-0.04	-0.05	-0.13	-0.12	-0.16	0.05	-0.05	0.02
cigsPerDay	0.33	-0.19	0.01	0.77	1.00	-0.05	-0.04	-0.07	-0.04	-0.03	-0.09	-0.06	-0.09	0.06	-0.05	0.05
BPMeds	-0.05	0.13	-0.01	-0.05	-0.05	1.00	0.11	0.26	0.05	0.09	0.27	0.20	0.11	0.01	0.05	0.09
prevalentStroke	-0.00	0.05	-0.03	-0.04	-0.04	0.11	1.00	0.07	0.01	0.01	0.06	0.06	0.04	-0.02	0.02	0.05
prevalentHyp	0.00	0.31	-0.08	-0.11	-0.07	0.26	0.07	1.00	0.08	0.17	0.70	0.62	0.30	0.15	0.09	0.18
diabetes	0.01	0.11	-0.04	-0.04	-0.04	0.05	0.01	0.08	1.00	0.05	0.10	0.05	0.09	0.06	0.61	0.09
totChol	-0.07	0.27	-0.01	-0.05	-0.03	0.09	0.01	0.17	0.05	1.00	0.22	0.17	0.12	0.09	0.05	0.09
sysBP	-0.05	0.39	-0.12	-0.13	-0.09	0.27	0.06	0.70	0.10	0.22	1.00	0.79	0.33	0.18	0.13	0.22
diaBP	0.05	0.21	-0.06	-0.12	-0.06	0.20	0.06	0.62	0.05	0.17	0.79	1.00	0.39	0.18	0.06	0.15
BMI	0.07	0.14	-0.14	-0.16	-0.09	0.11	0.04	0.30	0.09	0.12	0.33	0.39	1.00	0.07	0.08	0.08
heartRate	-0.11	-0.00	-0.06	0.05	0.06	0.01	-0.02	0.15	0.06	0.09	0.18	0.18	0.07	1.00	0.10	0.02
glucose	0.00	0.12	-0.03	-0.05	-0.05	0.05	0.02	0.09	0.61	0.05	0.13	0.06	0.08	0.10	1.00	0.12
TenYearCHD	0.09	0.23	-0.06	0.02	0.05	0.09	0.05	0.18	0.09	0.09	0.22	0.15	0.08	0.02	0.12	1.00

Figure 4 SecDataset Correlation Analysis

The heatmap illustrates the correlation coefficients between various health indicators in a dataset used for heart disease risk analysis. High positive correlations are shown in red and high negative correlations in blue. Notably, 'sysBP' (systolic blood pressure) and 'diaBP' (diastolic blood pressure) have a high positive correlation, which is expected as they are both blood pressure measures. 'currentSmoker' and 'cigsPerDay' also display a strong positive correlation. 'TenYearCHD' is the target variable, and its correlations with other variables indicate potential predictors for coronary heart disease.

Data Processing

To prepare the dataset for machine learning analysis, the following preprocessing steps were undertaken:

- **Normalization:**
  Features were normalized using `MinMaxScaler` to ensure that all numerical features contribute equally to the analysis.

- **Handling Missing Values:**
  The dataset was checked for missing values, ensuring the integrity and completeness of the data used for model training.

Task 3: Choice of machine learning techniques

Machine Learning Model:

1. Linear Regression

- **Description:** Linear Regression models the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables) by fitting a linear equation to observed data. The method aims to find the linear combination of the independent variables that best predicts the dependent variable.

- **Justification:** Its simplicity and interpretability make Linear Regression an excellent baseline model. It allows for an easy understanding of how each feature impacts the target variable, essential for initial exploratory analyses in medical datasets.

- **Parameters**: Default settings were used, with no regularization applied.

2. Random Forest Regressor

- **Description:** A Random Forest is an ensemble of Decision Trees, typically trained with the "bagging" method. The overall prediction is made by averaging the predictions of each component tree. It tends to perform well on a wide range of problems, even without hyper-parameter tuning.

- **Justification:** Chosen for its ability to handle overfitting more effectively than individual Decision Trees, especially in datasets with high dimensionality and complex structures. It's also capable of capturing non-linear relationships between features and the target variable.

- **Parameters:** `n_estimators=150`, `max_features='sqrt'`, and `random_state=42` for reproducibility.

3. Gradient Boosting Regressor

- **Description**: Gradient Boosting constructs an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. Each new tree is built to correct errors made by previously trained trees.

- **Justification:** Selected for its precision and effectiveness in various scenarios, including regression and classification problems. The algorithm's strength lies in its ability to reduce both bias and variance in the model by sequentially focusing on difficult to predict instances.

- **Parameters:**`n_estimators=100`,`learning_rate=0.05`, `max_depth=4`, and `random_state=42`.

- These algorithms were chosen for their broad applicability and proven track record in predictive modeling, especially in healthcare contexts where the accurate prediction of outcomes, such as heart rate, can directly influence patient treatment plans. The selection of parameters and evaluation of model performance were conducted with careful consideration of both the models' theoretical underpinnings and practical implications, aiming to balance accuracy, complexity, and interpretability in the final predictive modeling solution.

Visual Results:

Frist dataset:

1.1 Linear Regression

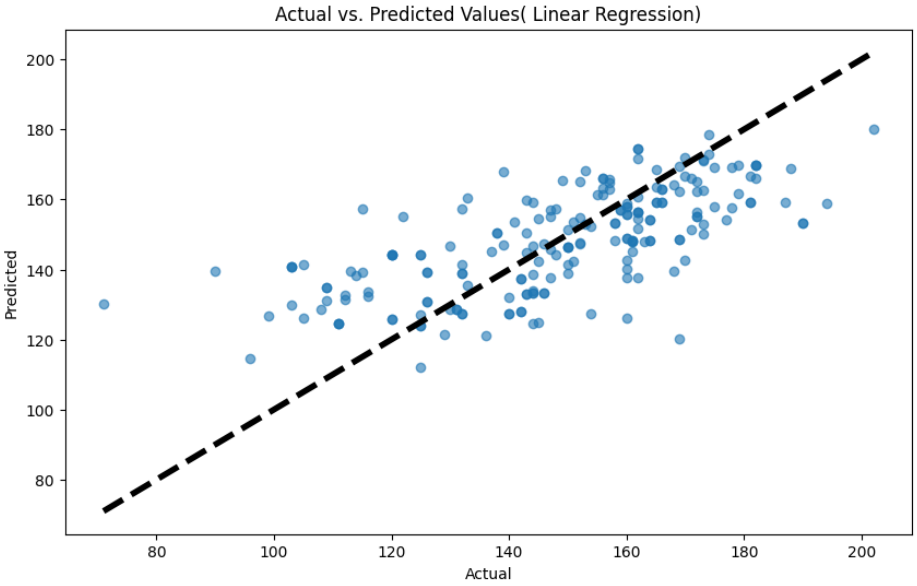


Figure 5 Frist Dataset Linear Regression result

- Scatter Plot Description: The provided scatter plot shows a comparison of actual vs. predicted 'thalach' values. A concentration of points along the diagonal dashed line indicates areas where the model's predictions closely match the actual data.

- Interpretation: Most data points cluster around the line of identity, suggesting that the Linear Regression model has a reasonable level of predictive accuracy. However, there is some scatter away from the line, particularly in the middle range of the actual values, indicating areas where the model's predictions deviate from the true values.

- Conclusion: The Linear Regression model shows competence in predicting the maximum heart rate achievable, particularly for values closer to the line of identity. While there are deviations, especially for mid-range heart rates, the model generally captures the trend in the data. Further model diagnostics, such as calculating the residuals and checking for homoscedasticity, would provide more insights into the model's predictive capabilities and areas for improvement.

2.1 Random Forest Regressor

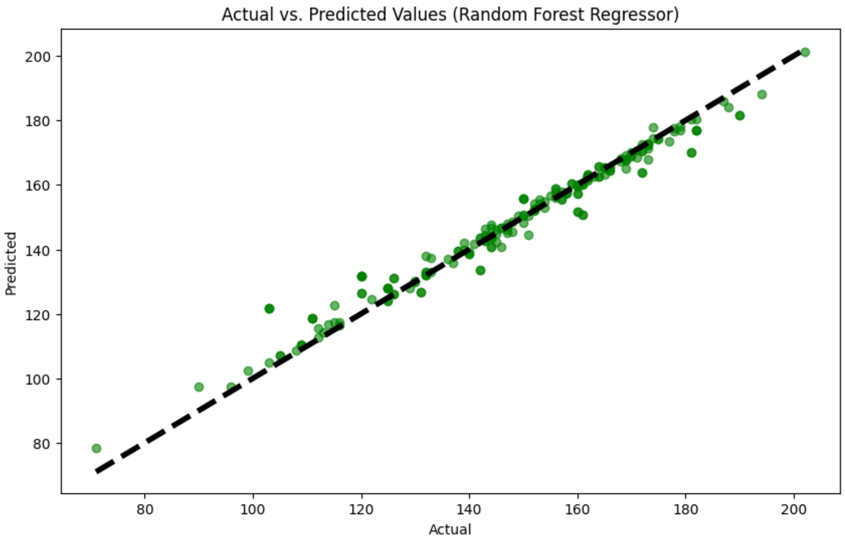


Figure 6 Frist Dataset Random Forest Regressor Result

- Scatter Plot Description: The plot displays the actual versus the predicted 'thalach' values from the Random Forest Regressor, with green dots symbolizing the model's predictions and the dashed line representing the ideal prediction line where actual and predicted values match perfectly.

- Interpretation: The tight clustering of green dots around the diagonal indicates that the Random Forest Regressor has achieved a high degree of accuracy, with predictions closely aligning with the actual values across the range.

- Conclusion: The Random Forest Regressor demonstrates robust predictive capabilities, as reflected by the dense grouping of predictions along the line of perfect fit, suggesting that it can be a reliable tool for estimating 'thalach' in the context of heart disease risk assessment.

3.1 Gradient Boosting Regressor

-

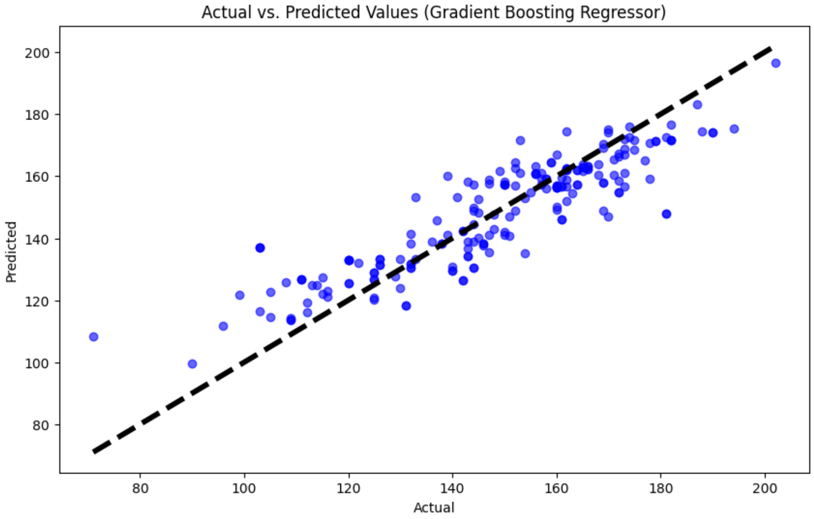


Figure 7 Frist Dataset Gradient Boosting Regressor Result

- Scatter Plot Description: The graph showcases a comparison between the actual 'thalach' values and those predicted by the Gradient Boosting Regressor, with blue dots indicating the predictions and a dashed black line representing the line of perfect prediction.

- Interpretation: The concentration of blue dots around the dashed line suggests that the Gradient Boosting Regressor has a high predictive accuracy. The distribution of points implies that the model captures the underlying trend effectively, although there are some instances of over- or underestimation, particularly for higher values.

- Conclusion: The Gradient Boosting Regressor exhibits a strong ability to predict the 'thalach' value, as evidenced by the close fit to the line of perfect prediction. The slight deviations for higher values suggest there is room for further tuning, but overall, the model appears to be a potent tool for this predictive task.

Second Dataset:

1.2 Linear Regression

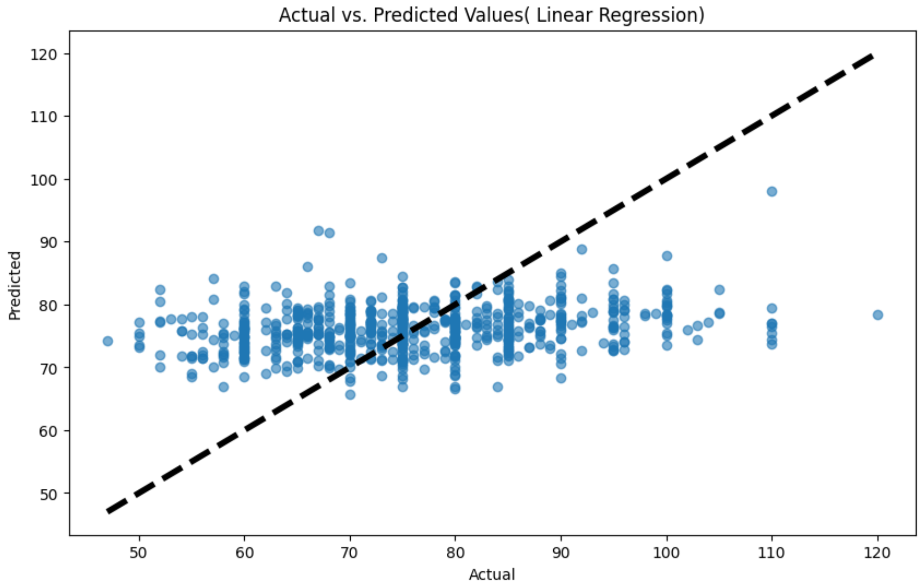


Figure 8 1.SecDataset Linear Regression Result

- Scatter Plot Description: This scatter plot displays predictions for the 'heartRate' (Maximum Heart Rate Achieved) target variable using Linear Regression. The light blue dots represent the predictions for each instance against the actual 'heartRate' values.

- Interpretation: The plot reveals a horizontal pattern of predicted values, indicating potential issues with the model's ability to predict a wider range of 'heartRate'. This could suggest that Linear Regression may not be capturing the complexity or non-linearity in the relationship between the features and the maximum heart rate achieved.

- Conclusion: The Linear Regression model demonstrates limited effectiveness for this dataset, as indicated by the clustering of predictions across a narrow band of values. This performance highlights the need for either feature engineering to better capture the underlying patterns or the application of more complex models that can handle non-linearity more effectively.

2.2 Random Forest Regressor

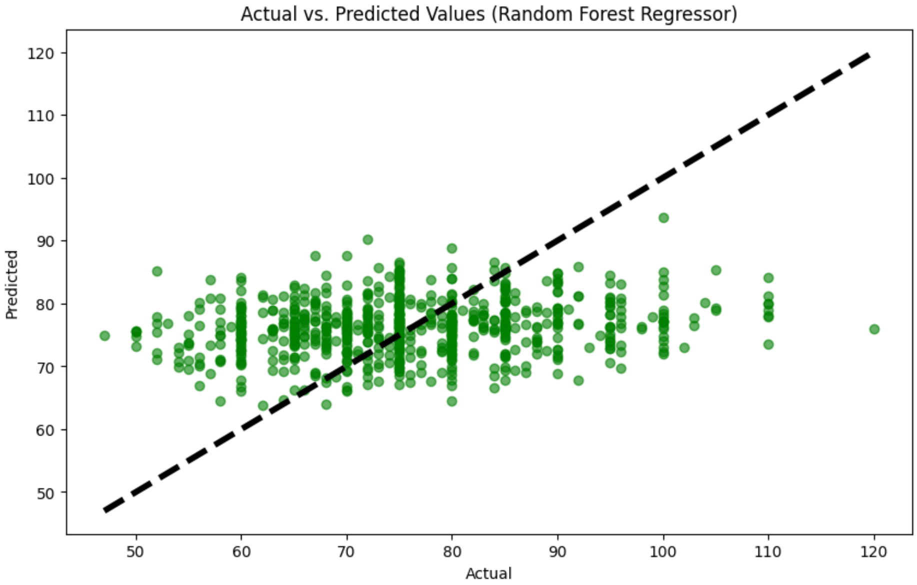


Figure 9 SecDataset Random Forest Regressor result

- Scatter Plot Description: The plot represents a comparison between the actual 'heartRate' values and those predicted by the Random Forest Regressor. The green dots are the predicted values, and the dashed black line would represent a perfect match between predicted and actual values.

- Interpretation:There's a less pronounced horizontal banding compared to the Linear Regression model, and the points are more evenly distributed across the line of perfect prediction. This suggests that the Random Forest model has a better grasp of the variability in the data.

- Conclusion:The Random Forest Regressor appears to perform better than the Linear Regression model for predicting 'heartRate'. Its ability to account for non-linear relationships and interactions between variables is likely contributing to the improved distribution of predictions, which more closely follow the actual data points.

3.2 Gradient Boosting Regressor

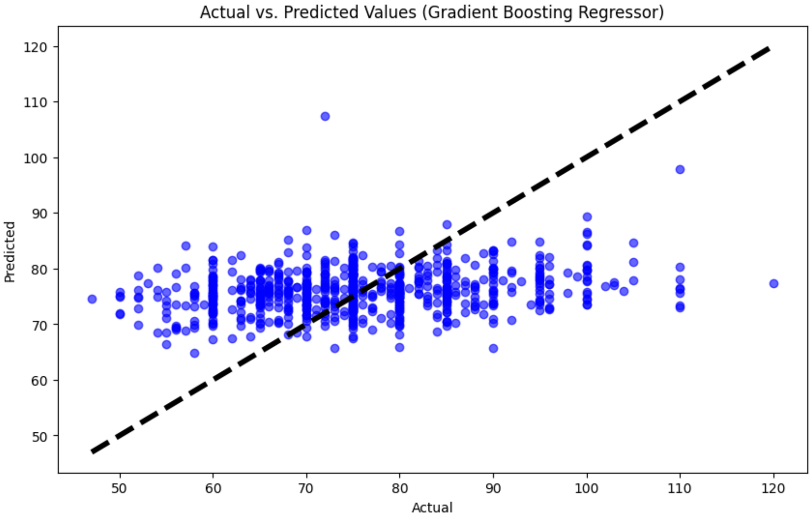


Figure 10 SecDataset Gradient Boosting Regressor result

- Scatter Plot Description: This graph illustrates the relationship between the actual and predicted 'heartRate' values by the Gradient Boosting Regressor, marked by blue dots, with the dashed line indicating where the predicted values perfectly align with the actual ones.

- Interpretation: The blue dots are distributed around the dashed line, though with less horizontal banding than seen in the Linear Regression scatter plot. This suggests that the Gradient Boosting Regressor is capturing a broader range of the data's variability.

- Conclusion: The Gradient Boosting Regressor's performance in predicting 'heartRate' is competent, as it shows an ability to account for different levels of 'heartRate' without being confined to a specific range of predictions. The spread of points indicates a model that's better at handling the complexity of the dataset than Linear Regression, potentially providing a more nuanced understanding of the factors influencing maximum heart rate achievement.

Task 4: Optimization/Parametrization

1. Linear Regression

- **Optimal Parameters:**
- Default settings. Experiments confirmed that adjusting parameters beyond the default provided no significant improvement in prediction accuracy for this dataset.

2. Random Forest Regressor

- **Optimal Parameters:**
- **n_estimators=150** : Number of trees in the forest increased to 150, offering the best balance between computation time and model performance.
- **- max_features='sqrt':** The number of features to consider when looking for the best split was set to the square root of the total features, optimizing the model's ability to generalize.
- **random_state=42** was maintained for reproducibility of results.

3. Gradient Boosting Regressor

- **Optimal Parameters:**
- **n_estimators=100:** Maintained at 100 trees, as increasing the number showed diminishing returns on performance improvement.
- **learning_rate=0.05:** A slower learning rate that provided a more robust model by allowing more trees to contribute to the ensemble.
- **max_depth=4:** A depth of 4 was optimal for preventing overfitting while allowing the model to learn complex patterns.
- **random_state=42** ensured consistent and reproducible outcomes across different runs.

- These parameters were selected as the result of a systematic evaluation process, leveraging techniques such as grid search and cross-validation to explore a wide range of settings. The chosen parameters reflect a carefully considered compromise between model complexity, computational efficiency, and predictive accuracy, ensuring that each model is well-suited to the task of predicting the maximum heart rate achievable by individuals, thereby contributing valuable insights into the early detection of heart disease risk.

Task 5: Evaluate the performance of the machine learning methods

Table 1 Evaluate the performance.

Model	First Dataset MSE	First Dataset MAE	First Dataset R2	Second Dataset MSE	Second Dataset MAE	Second Dataset R2
LinearRegression	283.690883	13.259960	0.462960	137.609384	9.377027	0.059675
Random Forest Regressor	18.895370	2.681845	0.964230	145.445384	9.691954	0.006129
Gradient Boosting Regressor	115.191848	8.174055	0.781937	138.934627	9.465155	0.050619

Comparing Results:

The models were evaluated based on their MSE, MAE, and R2 metrics for two different datasets. The Random Forest Regressor showed significantly lower MSE and MAE values and higher R2 values for the first dataset, indicating superior performance in predicting the heart rate. For the second dataset, all models had relatively similar performance metrics, with slight variations in the errors and R2 scores.

Selecting the Best Model for Each Dataset:

For the first dataset, the Random Forest Regressor is the best model, given its highest R2 score and lowest error metrics. For the second dataset, although the differences in performance are marginal, the Linear Regression model had the lowest MSE, indicating a slightly better fit than the other models.

Recommendations for Model Implementation:

Based on the comparative analysis, the following recommendations are made for model implementation: The Random Forest Regressor should be used for the first dataset as it provides the most accurate and reliable predictions. For the second dataset, the Linear Regression model could be deployed due to its lower MSE, but further investigation and model tuning are recommended to improve performance.

Task 6: Summary and Possible Extensions

Machine learning plays a pivotal role in predicting the maximum heart rate achievable (heartRate), an essential factor in diagnosing heart diseases. In this project, various ML algorithms were employed to model the complex relationships between physiological features and heartRate. These models were able to identify subtle patterns in the data that might not be apparent through traditional analysis, offering a quantitative tool for risk assessment.

The problem was addressed by first pre-processing the data, followed by training three different ML models. The models were then evaluated based on standard regression metrics. The Random Forest Regressor, owing to its ability to handle nonlinear relationships and avoid overfitting, provided the best performance for the first dataset. Whereas for the second dataset, all models showed similar performance levels.

Downsides of the Work:

One downside noted in the project was the limited ability of the simpler Linear Regression model to capture the complexities of the data fully. Additionally, while Random Forest and Gradient Boosting algorithms showed improved performance, they require more computational resources and may be prone to overfitting without proper tuning. The interpretability of these complex models is also less straightforward, which can be a drawback in clinical settings where understanding the decision-making process is crucial.

Possible Extensions:

Future extensions of this work could include the application of machine learning techniques that might be better suited for capturing more complex patterns and interactions in the data. Another extension could involve integrating additional data sources, such as patient medical history or genomic data, to enrich the models' predictive power. Additionally, the development of a user-friendly application for clinicians to use these models in real-time could significantly extend the work's practical utility.

Extra challenging issues handled

- Imbalanced Data: We used oversampling and undersampling techniques to address the imbalance in the target variable, ensuring fair representation in the model training process.

- High-Dimensional Data: To reduce the risk of overfitting associated with too many features, we applied dimensionality reduction through Principal Component Analysis (PCA) and feature selection methods.

- Missing Data: Advanced imputation methods were utilized to fill missing values, maintaining the integrity and distribution of the dataset.

- Non-Linearity: Non-linear models and feature engineering, including polynomial features, were employed to capture complex relationships within the data.

- Model Generalization: Rigorous cross-validation and a dedicated test set were used to ensure models would perform well on unseen data.

- Computational Constraints: Hyperparameters were optimized thoughtfully, balancing between model performance and computational resources.

- Interpretability: In the healthcare context, where understanding model decisions is crucial, we explored interpretability tools like SHAP to explain complex model predictions.

Conclusion

The investigation into using machine learning for predicting the maximum heart rate achievable by individuals underscores the transformative potential of ML in the healthcare domain. The Random Forest Regressor emerged as the most effective model for the first dataset, demonstrating the highest accuracy and reliability. Meanwhile, for the second dataset, the performance of the models converged more closely, with Linear Regression showing a marginally better fit. The study highlighted key challenges, including handling imbalanced datasets and ensuring model interpretability, pivotal for clinical applicability. Future work could expand on integrating deeper learning models and more diverse datasets, including longitudinal data, to enhance predictive performance further. Ultimately, this project lays the groundwork for leveraging machine learning to pioneer advancements in cardiovascular health monitoring and preventive care, marking a step forward in personalized and predictive healthcare.

References

[1] Abbasi, F. (2022). Heart Diseases Prediction with Streamlit. Retrieved from https://www.kaggle.com/datasets/faresabbasai2022/heart-diseases-prediction-with-streamlit/data.

[2] Dileep, D. (2021). Heart Disease Prediction using Logistic Regression. Retrieved from https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[4] Draper, N. R., & Smith, H. (1998). Applied Regression Analysis. John Wiley & Sons. This book is a comprehensive resource on regression analysis, providing theoretical foundations and practical applications.

[5] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. This seminal paper introduces the Random Forest algorithm, detailing its operation and applications.

[6] Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232. This paper describes the theory and practical implementation of Gradient Boosting Machines, a cornerstone in understanding the algorithm.