ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Heart Disease Prediction

Emma Ruiz Serrano

November 7, 2024

Abstract

Heart disease (HD) are among the leading causes of mortality worldwide, and early diagnosis is essential for preventing serious complications and reducing mortality rates. The objective of this paper is to develop a predictive system for HD using machine learning algorithms, enabling the identification of individuals at high risk of experiencing cardiac issues, and identifying the most relevant factors that determine the likelihood of HD. We utilized a cross-sectional dataset from the Behavioral Risk Factor Surveillance System (BRFSS) in 2021, provided by the World Health Organization (WHO), comprising health and lifestyle information from over 300,000 individuals. We applied various machine learning models, including Logistic Regression, Random Forest, K-Nearest Neighbors (KNN) and XGBoosting, apart from using the complex technique of Neural Networks. These algorithms were evaluated in terms of accuracy, sensitivity, and specificity to determine their effectiveness in predicting HD. Our approach not only considers traditional clinical parameters, but also lifestyle factors such as fruit consumption, physical activity or smoking history. Our results indicate that the Neural Network model outperformed other algorithms with a sensitivity of 82% and an overall accuracy of 72%, making it the most suitable model for identifying HD risk. Key factors, such as age, diabetes, smoking history, and general health status, emerged as the most relevant predictors, highlighting the importance of both clinical and lifestyle variables in accurate disease prediction. By integrating these factors, we aim to provide an accessible and precise tool that focuses on individuals' daily lives, ultimately improving the early detection of heart disease and facilitating the implementation of preventive measures that can enhance quality of life and reduce healthcare costs.

1 Introduction

Heart disease (HD)) , now one of the leading causes of death globally, are often linked to factors beyond personal control, such as age, gender, family history, or genetic predispositions. However, modifiable lifestyle choices, including diet, physical activity, and avoidance of harmful substances, also play a significant role in influencing the likelihood of developing these conditions. Heart disease causes over 31 percent of all deaths worldwide, with the World Health Organization reporting approximately 17.9 million deaths each year, and nearly 805,000 heart attacks annually just in the United States -[7][1] . This prevalence makes it critical to improve our methods for diagnosing and predicting heart conditions before they become life-threatening. Early detection could prevent permanent harm, reduce mortality rates, and improve patient outcomes, while also easing the burden on healthcare systems [10].

In response, there has been a growing focus on developing preventive tools capable of identifying heart risk factors before clinical symptoms emerge. However, accurately predicting heart disease remains complex, particularly given the increasing prevalence of lifestyle-related risk factors, including physical inactivity, smoking, and poor dietary habits. Addressing these challenges, Machine Learning (ML) offers a powerful approach to enhance early detection efforts. ML enables the construction of sophisticated predictive models that surpass traditional statistical methods by capturing intricate, nonlinear relationships among diverse risk factors, as underscored in recent findings in Frontiers in Medicine[12].

The aim of this project is to identify the most relevant factors that determine the likelihood of a person developing heart disease, as well as to explore how we can use general health data, lifestyle, and medical history to create an accurate predictive model by using ML. Moreover, this work aspires to enhance the accuracy of heart disease prediction, providing a practical tool that can positively impact the quality of life for individuals through early identification of risks and the timely application of preventive measures.This model not only aids in identifying individuals at higher risk but also serves as a resource for healthcare professionals, allowing them to incorporate personal daily habits into risk assessments and tailored intervention plans. Furthermore, our approach to integrating lifestyle data into heart disease predictions could serve as a foundation for research on other chronic conditions, encouraging a holistic, data-driven perspective on preventive health. By contributing to the growing body of knowledge on machine learning in healthcare, this project aspires to support future studies and practical applications, helping clinicians and researchers alike to harness predictive insights for more effective, real-world health strategies.

This study revealed that integrating clinical and lifestyle factors into predictive models for heart disease can significantly improve accuracy. The Neural Network model demonstrated the highest sensitivity, effectively minimizing the risk of missed diagnoses, which is essential in clinical settings. Important predictors such as age, diabetes, smoking history, and general health were identified as key determinants in heart risk, underscoring the value of a holistic approach. These findings suggest that incorporating diverse health indicators can lead to a more effective assessment of heart risk, supporting early intervention and prevention strategies that could be beneficial in broader applications

1.1 Related Work

Heart disease prediction has been extensively researched using machine learning (ML) techniques due to their ability to handle large datasets and capture complex patterns. Below, we highlight specific approaches from recent studies, emphasizing the methods employed to address challenges in accuracy and data complexity in heart disease prediction.

Mohan et al. (2019) propose a hybrid model that combines Random Forest with a linear model (HRFLM), achieving an accuracy of up to 88.7% by optimizing feature selection. This approach is noteworthy for managing high-dimensional data, though it requires significant manual tuning of hyperparameters[9].

The work of Chiuve et al. (2014) introduces the "Healthy Heart Score," a model that uses lifestyle factors such as exercise, alcohol consumption, body mass index, and smoking status to predict long-term heart risk . This model contributes to primary prevention by identifying at-risk individuals early. However, its accuracy relies on self-reported data, which can introduce bias[3].

In a meta-analysis, Krittanawong et al. (2020) assess various ML algorithms for heart disease prediction, finding that Support Vector Machine (SVM) and boosting algorithms yield high AUC values (approximately 0.90), making them suitable for complex disease prediction tasks . However, this study does not address the impact of lifestyle factors or data complexity on model performance[8].

Aroojthe et al. (2022), implements a deep convolutional neural network (DCNN) as a tool to classify individuals as either healthy or affected by CVD.Through performance metrics such as accuracy, recall, and F1 score, the model achieved a validation accuracy of 91.7%, demonstrating that the DCNN is effective in identifying complex patterns in medical data.[2].

Despite the significant progress, it has been observed that while some studies have considered lifestyle factors, such as the work by Chiuve et al.[3], few models combine clinical, demographic, and behavioral data into a single predictive approach. This study addresses this gap by developing a model that incorporates both traditional medical parameters and behavioral factors, such as physical activity, diet, and smoking history, providing a more comprehensive view of heart disease risk.

Furthermore, in our project, a Deep Learning technique called Neural Network has been implemented. This technique is very useful for predicting CVD because it can analyze complex patterns in large datasets, learning from various input features to identify subtle relationships that traditional models might overlook. This allows for accurate risk prediction and early detection of heart disease. Indeed, it is true that this is not the first time this resource has been used, as seen in Aroojthe et al.[2]. However, it is not very common to find it used alongside other ML techniques and compared to determine the most optimal one.

2 Data analysis

The dataset, sourced from BRFSS of the World Health Organization, comprises rich health and lifestyle information for 308,854 individuals across 19 features, covering general health, lifestyle habits, demographics, and medical conditions. This diversity of features—from general health status and checkup frequency to habits like smoking, diet, and exercise—enables a comprehensive analysis of patterns and relationships, potentially illuminating heart disease risk factors.

Most variables are categorical, capturing various health and lifestyle characteristics. **General_Health** represents self-reported health status with values such as "Poor", "Fair", "Good", "Very Good," and "Excelent". **Checkup** indicates the recency of each individual's last medical checkup, categorized as "Never", "Within the past year", "Within the past 2 year", "Within the past 5 year" or "5 or more years ago" reflecting proactive health practices. **Exercise** is a binary variable showing whether an individual exercises or not, while **Smoking_History** indicates past or current smoking habits. **Alcohol_Consumption** is numerical, measuring the amount of alcohol intake, with an average of 5.1 units and a maximum of 30. Additionally, **Sex** and **Age_Category** provide demographic context, with Sex categorized as "Male" or "Female" and Age Category grouped in ranges such as "60-64," "70-74," etc.

The dataset also includes several health history variables. **Heart_Disease** (the target variable) indicates the presence or absence of heart disease, likely in a binary format ("Yes"/"No"). Other health history variables like **Skin_Cancer**, **Other_Cancer**, **Depression**, **Diabetes**, and **Arthritis** similarly use "Yes"/"No" values, providing context on individual health backgrounds potentially associated with heart disease risk.

Several numerical variables capture physical measurements and dietary habits: **Height** (in centimeters), **Weight** (in kilograms), and Body Mass Index (**BMI**), calculated from height and weight. Unlike many similar studies, this dataset uniquely includes dietary variables such as **Fruit_Consumption**, **Green_Vegetables_Consumption**, and **FriedPotato_Consumption**, which quantitatively represent the intake of fruits, vegetables, and fried potatoes, respectively.
To see the statistics of the variables go to Figure 1 and 2

	Min	Median	Mean	Max
General_Health	1.00	4.00	3.53	5.00
Last_checkup	1.000	5.000	4.618	5.000
Height_cm	91.0	170.0	170.6	241.0
Weight_kg	24.95	81.65	83.59	293.02
BMI	12.02	27.44	28.63	99.33
Alcohol_Consumption	0.00	1.000	5.096	30.000
Fruit_Consumption	0.00	30.00	29.84	120.00
Green_Vegetables_Consumtion	0.00	12.00	15.11	128.00
FriedPotato_Consumption	0.00	4.00	6.297	128.00
Age_numeric	21.00	57.00	54.84	85.00

Figure 1: Min, Median, Mean and Max of numeric variables

	prop 0	prop 1
Exercise	22.4938 %	77.5062 %
Skin_Cancer	90.288615 %	9.711385 %
Other_Cancer	90.326174 %	9.673826 %
Depression	79.95784 %	20.04216 %
Diabetes	86.99353 %	13.00647 %
Diabetes_Pregnancy	99.1432845 %	0.8567155 %
Arthritis	67.27548 %	32.72452 %
Smoking_History	59.44233 %	40.55767 %
Sex_Male	51.86787 %	48.13213 %

Figure 2: Proportion of observations in each class o our categorical variables

To prepare the dataset for analysis, we transformed several categorical and ordinal variables to numerical formats suitable for modeling. We used dummy encoding for binary variables such as Exercise, Heart Disease, Skin Cancer, Other Cancer, Depression, Diabetes, Arthritis, Smoking History, and Sex, removing the first category to prevent multicollinearity. We also transformed ordinal variables General Health and Checkup into ordered numerical scales, ranking health and checkup frequency in descending order for interpretability. Additionally, we created a new Age numeric variable by converting age ranges into representative midpoint values, removing the original categorical Age Category column to streamline the dataset. These transformations enhance model performance and provide a more robust basis for analyzing relationships within the data.

3 Methodology

Machine learning is a branch of artificial intelligence (AI) that enables systems to learn and improve automatically from experience, without explicit programming. The learning process starts with data, examples, or instructions, allowing the system to identify patterns and refine its performance over time as new data becomes available.

3.1 Division in train and test

To effectively train a machine learning model, it is crucial to divide the dataset into two distinct subsets: the training set and the test set. The training set is utilized to fit the model, allowing it to learn the underlying patterns and relationships within the data. Conversely, the test set comprises observations that the model has not encountered during training, facilitating an objective evaluation of the model's performance and its ability to generalize to new, unseen data. This separation is essential for identifying potential issues, such as overfitting, where the model may perform well on training data but poorly on new examples.

In this specific context, significant class imbalance exists in the target variable, with approximately 92% of cases classified as negative and only 8% as positive. This imbalance poses a challenge, as the model may struggle to learn the characteristics of the minority class due to the overwhelming number of negative cases. If left unaddressed, the model could become biased, favoring the majority class and failing to accurately predict positive cases.

To mitigate this issue, downsampling the majority class and training the models with fewer negative cases, so a **balanced training set**, referred to as train_balanced, is created. This set comprises an equal number of observations for both classes—50% positive cases (target value of 1) and 50% negative cases (target value of 0). This approach ensures that the model is exposed to both classes during training, thereby enhancing its ability to recognize and predict the minority class effectively. Implementing this strategy aims to improve the model's overall performance and ensure a more robust understanding of the data.

3.2 Variable selection

Variable selection is essential in machine learning, as it improves model performance and interpretability. Multicollinearity, or high correlation between predictors, can cause redundancy, making it harder to discern each predictor's impact on the target variable.

In addition, effective variable selection helps manage bias and variance. Bias refers to the difference between the model's predictions and the actual values, while variance indicates how much the predictions would change with different training datasets. Reducing bias often increases variance, resulting in a model that fits the training data well but may not generalize effectively to new data—this is known as overfitting. Conversely, lowering variance can increase bias, leading to underfitting.
See Figure 3

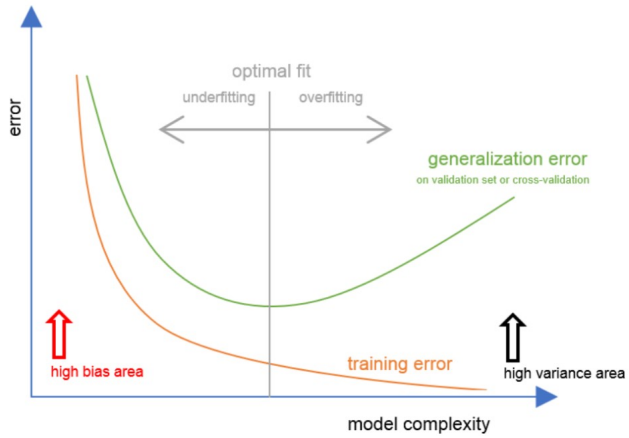


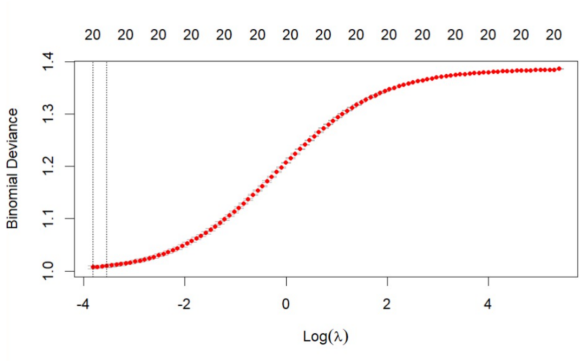
Figure 3: Error - Model Complexity [5]

The **bias-variance trade-off** describes how increasing model complexity initially reduces training error but can lead to overfitting, where the model captures noise rather than true patterns, raising error on new data. The goal is to find a balance that minimizes generalization error. Techniques like Ridge and Lasso regression help by selecting essential features, which reduces variance and increases model robustness with a small bias trade-off. In both ridge and lasso regression, regularization terms are added to the objective function to prevent overfitting by penalizing large coefficients.

Ridge Regression applies a penalty based on the squared magnitude (L2 norm) of the coefficients, constraining them within a circular boundary in parameter space. This circular constraint leads ridge to shrink coefficients towards zero, but it rarely sets any of them exactly to zero. As a result, ridge regression typically retains all predictors in the model, simply reducing their influence when they are less relevant.

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Minimizing the residual sum of squares (RSS) while adding a penalty term that is proportional to the square of the coefficient

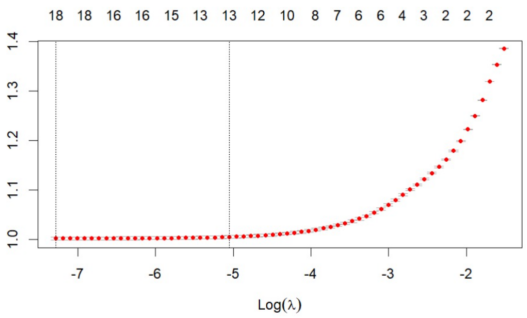


The Ridge regression model selected an optimal lambda of 0.02195, shrinking predictor coefficients without setting any to zero due to the L2 penalty. Thus, all variables remain in the model, though their influence varies. Variables like General_Health, Diabetes, Smoking_History, and Sex_Male have higher coefficients, indicating they are more influential predictors of heart disease in the dataset.

Lasso, applies the L1 penalty, contrasting coefficients within a diamond-shaped boundary that allows some to be reduced exactly to zero. This geometry enables Lasso to perform variable selection, excluding less relevant predictors and making it especially useful for enhancing model interpretability.

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

Similar to Ridge objective function, but with a different penalty



In the adjusted Lasso model, the optimal lambda value was 0.0006861242, allowing for the selection of a subset of relevant variables. Through its penalty, the model effectively eliminated variables like **FriedPotato_Consumption** and **BMI**, and retained those with a stronger influence on predicting heart disease. This feature selection enhances model interpretability, focusing on predictors with significant contributions to cardiovascular risk.

Lasso's interpretation provides a more interpretable model by selecting a subset of predictors, whereas ridge regression includes all predictors in the final model, which can complicate interpretation. This is why, after implementing both methods, we have decided to proceed with the Lasso's one, and delete FriedPotato_Consumption and BMI variables

Cross-validation: In this study, we incorporated cross-validation into our model evaluation process to enhance the robustness and reliability of our results. Specifically, for the logistic regression model, we used 10-fold cross-validation, meaning the dataset was partitioned into 10 subsets (folds), with the model being trained on 9 folds and validated on the remaining fold in each iteration. This process was repeated 10 times, ensuring that each fold served as the validation set once, and the results were averaged to provide a more reliable estimate of model performance. For the random forest model, we employed 5-fold cross-validation, where the dataset was divided into 5 subsets, and the model was trained on 4 folds and validated on the fifth fold. This approach helped mitigate overfitting by ensuring the models were evaluated on different data points, thus offering a more accurate understanding of their ability to generalize to unseen data. Both cross-validation techniques were implemented using the "trainControl" function from the "caret" package, providing a comprehensive and reliable evaluation of the models' performance. This technique helps to mitigate overfitting by ensuring that the model is evaluated on different data points, providing a more reliable estimate of its performance on unseen data. See Figure 4

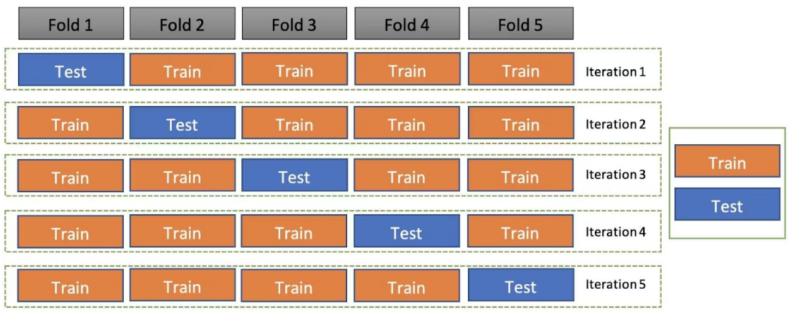


Figure 4: K-fold cross-validation [11]

The image above illustrates a 5-fold cross-validation process, comprising 5 iterations. In each iteration, one fold is designated as the test (or validation) set, while the remaining (k-1) folds (4 folds) serve as the training set. To obtain the final accuracy, the validation accuracies from each of the (k) models are averaged.

Hyperparameter's optimization is a critical process in machine learning that aims to improve model performance by adjusting predefined settings that are not learned from the data. These settings, or hyperparameters, include factors like learning rates, the number of trees in a Random Forest, or regularization strengths in regression models. While machine learning algorithms often have default hyperparameter values, tuning these parameters can lead to significant performance improvements tailored to specific datasets. In this study, we employed grid search for hyperparameter optimization, which exhaustively evaluates all possible combinations of specified hyperparameters. Although computationally intensive, grid search ensures a thorough exploration of parameter configurations, enhancing the model's predictive power.

Classification threshold is a critical parameter in binary classification models that determines the point at which a predicted probability is converted into a class label. Typically, a threshold of 0.5 is used, meaning that if the predicted probability of a positive class exceeds 0.5, the instance is classified as positive; otherwise, it is classified as negative.

Adjusting the decision threshold in a clinical context, particularly when predicting a disease like heart disease, is a critical decision that directly impacts diagnostic outcomes. In our case, where the goal is to accurately predict heart disease, the choice of threshold significantly influences the model's ability to identify patients with the condition. Increasing the threshold reduces sensitivity, meaning some patients with heart disease may be missed (false negatives), but improves specificity, reducing the likelihood of misdiagnosing healthy individuals (false positives). This approach is often used when the priority is to avoid unnecessary follow-up procedures or invasive tests for patients who are not actually at risk.

However, this is not our case. Missing a heart disease diagnosis could have severe consequences, so we prefer to diagnose someone with the disease and later confirm that they do not have it rather than risk overlooking a true positive. Therefore, what we considered was lowering the threshold, which increases sensitivity, enabling the model to identify more patients with heart disease, thereby reducing the chances of missing a diagnosis. Although this lowers specificity and results in more false positives, in clinical settings like heart disease prediction, the risk of missing a diagnosis is far more detrimental than diagnosing someone who may not actually have the disease. Adjusting the threshold in this context requires careful consideration, prioritizing early and accurate disease detection while minimizing the risk of harm from missed diagnoses.

3.3 Machine Learning algorithms

Next, we will emphasize the machine learning methods used in the development of our project:

Logistic Regression is a fundamental method for binary classification that predicts the probability that an observation belongs to a specific class, such as the likelihood of developing heart disease (Heart Disease = 1). This technique utilizes the logistic (or sigmoid) function to transform a linear combination of predictor variables into a probability score. By mapping the linear predictor to a value between 0 and 1, logistic regression allows for the assessment of risk based on an individual's health characteristics and lifestyle factors. Mathematically, the model can be expressed as follows:

$$P(\text{Heart Disease} = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \dots)}}$$

Where $P(\text{Heart Disease} = 1 \mid X)$ is the probability of the outcome, in this case, the probability that a person has heart disease, given their set of characteristics or predictor variables X, such as health information or individual lifestyle. β_0 is the intercept, and $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients, representing the weight or impact of each predictor on the risk of developing the disease. For the predictors $X_1, X_2, \dots, X_n$, we have:X_1 = 'General_Health', X_2 = 'Age_numeric', ..., X_{19} = 'Alcohol_Consumption'.

In this study, Logistic Regression was specifically applied to predict the likelihood of heart disease based on a variety of health and lifestyle-related predictors. Following a comprehensive correlation analysis, it was identified that several variables exhibited significant inter-correlations, particularly between Height and Weight, as well as between Arthritis and Diabetes. Recognizing the importance of these relationships, the model was enhanced by incorporating interaction terms among these correlated variables to address potential non-linearities that could affect predictive accuracy. Through this methodological enhancement, the model not only captures the direct effects of each predictor but also their combined effects, thereby providing a more nuanced understanding of the risk factors associated with heart disease.

Random Forest is an ensemble learning method that enhances model performance by combining multiple decision trees, each trained on a random subset of data through a process known as bagging (bootstrap aggregation). This technique introduces variability among trees and reduces overfitting, as errors from individual trees tend to balance out in the final model. At each tree split, only a random subset of features is considered, which further minimizes correlation among trees, allowing the model to better capture complex, nonlinear relationships in the data.

In classification tasks, Random Forest makes its final prediction through majority voting across all trees, which enhances accuracy and robustness.This process enhances the model's ability to recognize nonlinear patterns in risk factors, such as the interaction between diet (Fruit_Consumption, Fried-Potato_Consumtion and Green_Vegetables_Consumption), physical activity, and pre-existing health conditions (like Diabetes or Arthritis), which are often associated with cardiovascular diseases.
See Figure 5

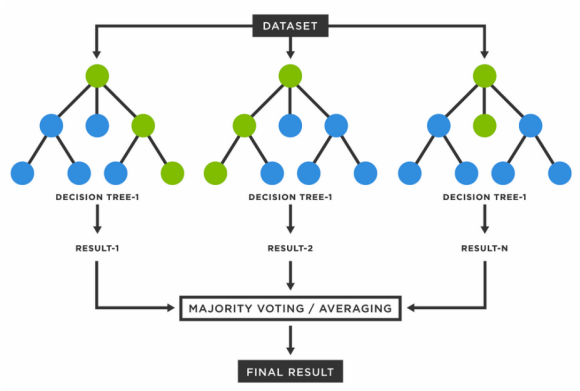


Figure 5: Example Random Forest [4]

K-Nearest Neighbors (KNN): The algorithm is a non-parametric, supervised learning technique commonly used for classification tasks. Since it does not assume an underlying distribution or define parameters, KNN relies directly on the training data to make predictions.

This algorithm identifies the 'k' closest data points in the training set based on a chosen distance metric, typically Euclidean distance. Once these 'k' neighbors are identified, the algorithm assigns the predicted class label to the new data point by majority vote. In the context of heart disease, this means that if the majority of the closest neighbors have heart disease, the new individual is also likely to have heart disease. Conversely, if most neighbors do not have the condition, the new data point will be classified as not having heart disease. In this project, prior to applying the KNN algorithm, we

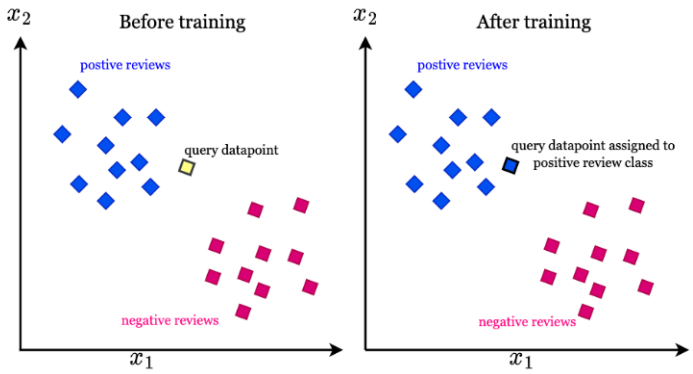


Figure 6: Example Random Forest [6]

normalized both training and testing datasets to ensure that all features had an equal impact on the distance calculations. Then identify the optimal value of k. As the value of k increases, the specificity improves, although sensitivity gradually decreases. K= 9, it has the best balance between sensitivity and specificity, this choice allowed us to maximize specificity while minimizing the loss of sensitivity, thereby enhancing the model's overall performance in predicting heart disease risk.

XGBoosting: Given the large volume of data and the need for high precision, XGBoost is a more suitable choice than Gradient Boosting. The trees are grown sequentially, each tree is grown using information from previously grown trees. It learns slowly but it performs well. It minimizes the predictor error by constructing each new model to correct the residuals, or errors, of the previous models. We can calculate the predictions with:

$$\hat{y} = F(x) = \sum_{t=1}^{T} f_t(x)$$

$\hat{y}$ is the final prediction, F(x) is the accumulated prediction after T iterations, ft(x) is the prediction of the model at iteration t, which in this case is a decision tree and TT is the total number of trees in the model

Predictions were made on the test dataset by converting the predicted probabilities into binary class labels to identify individuals at risk of heart disease. By incorporating XGBoost's advanced regularization techniques, the model's predictive accuracy was significantly improved compared to other boosting methods, while minimizing the risk of overfitting. This ultimately enhanced the model's ability to more reliably assess heart diseases risk.

Support Vector Machines (SVM) is a powerful classification algorithm that aims to find an optimal decision boundary (hyperplane) that best separates the data into two classes, such as "heart disease present" or "heart disease absent" while maximizing the margin between them. The margin is defined as the distance from the hyperplane to the closest points of each class, known as support vectors. By maximizing this margin, SVM enhances the model's generalization capabilities, improving its ability to generalize well to unseen data and reducing the risk of overfitting.

Mathematically, SVM addresses the following optimization problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i(w \cdot x_i + b) \geq 1, \forall i$$

Where w is the weight vector, b is the bias term, and yi represents the class labels.

Neural networks are computational models inspired in the human brain, composed of interconnected layers of nodes, that learn to map inputs to desired outputs through weighted summation and activation functions.

$$y = f\left(\sum_i w_i x_i + b\right)$$

'where xi are input features, wi are weights learned through training, b is a bias term, and f is an activation function'

The network's objective is to minimize the difference between predicted and actual values by adjusting weights and biases through gradient-based optimization methods.

In this study, we developed a neural network model for classifying the likelihood of heart disease. The architecture includes an input layer with 64 neurons, followed by three hidden layers to improve predictive ability: the first has 128 neurons, and the next two layers 64 neurons each.

We useLeakyReLU as the activation function, and dropout layers with rates between 0.2 and 0.3 to mitigate overfitting. The output layer consists of a single neuron with a sigmoid activation function, suitable for binary classification. Additionally, we implemented Early Stopping to halt training when the model's performance ceased to improve, ensuring model generalization on unseen data. This configuration was designed to address class imbalance and improve the sensitivity of heart disease prediction.

This neural network model was implemented using Python due to its comprehensive libraries for deep learning and neural networks, including TensorFlow and Keras.

3.4 Evaluation of the models

The evaluation phase is essential for assessing the performance of learning models. Key evaluation metrics include **accuracy**, **sensitivity** (recall), and **specificity** (precision), all derived from the **confusion matrix**, which details classifier performance on test data. These metrics are calculated using **true positives (TP)**, **true negatives (TN)**, **false positives (FP)**, **and false negatives (FN)**, with values ranging from 0 to 1.

Accuracy measures the proportion of correct predictions out of all predictions, while recall focuses on correctly identifying positive cases, which in our case, refers to correctly predicting the presence of heart disease. **Precision** evaluates how well the model identifies negative cases, in this case, correctly predicting individuals without heart disease.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Recall} = \frac{TP}{TP + FN} \quad \text{Precision} = \frac{TP}{TP + FP}$$

Additionally, the **ROC curve** (Receiver Operating Characteristic curve) is a vital tool in evaluating the diagnostic performance of a model. It provides insight into the trade-offs between sensitivity and specificity at various classification thresholds. The Y-axis represents **Sensitivity** (True Positive Rate), showing the proportion of correctly identified positive cases. The X-axis represents **1 - Specificity** (False Positive Rate), showing the proportion of incorrectly classified negative cases as positive.The area under the ROC curve, known as the AUC, offers a single, comprehensive metric of model performance, where higher values indicate stronger discriminatory power and an improved ability to distinguish between positive and negative cases.

4 Findings

4.1 Correlation analysis

The correlation analysis between predictor variables and the target variable, Heart Disease, elucidates several significant relationships. Among these, **Age_numeric** demonstrates the strongest positive correlation with Heart Disease (0.230), indicating that older age is closely associated with an increased risk of heart conditions. Following this, **Diabetes** (0.181), **Arthritis** (0.154), and **Smoking_History** (0.108) also show noteworthy positive correlations, suggesting that individuals with these conditions may be at heightened risk for heart disease.

Conversely, **General_Health** presents a negative correlation of -0.233, signifying that poorer self-reported health status correlates with a higher incidence of heart disease, suggesting that as general health declines, the likelihood of heart disease increases. This observation underscores the strong association between poor general health and heart conditions, often linked to broader health concerns such as obesity, sedentary lifestyles, and other lifestyle-related risk factors.
See Figure 7

The heat map displayed above provides a comprehensive overview of the correlations between various features in the dataset. Correlation coefficients range from -1 to 1, where -1 signifies a perfect negative correlation, indicating that as one variable increases, the other decreases, 1 denotes a perfect positive correlation, implying that both variables move in the same direction, and 0 represents no correlation, suggesting an absence of a linear relationship between the variables.

Some notable correlations include a strong positive relationship between **Weight_(kg)** and **BMI**, with a correlation coefficient of 0.88. This high correlation is expected, as Body Mass Index (BMI) is a calculation that incorporates both weight and height measurements, reflecting the direct relationship between these two metrics.

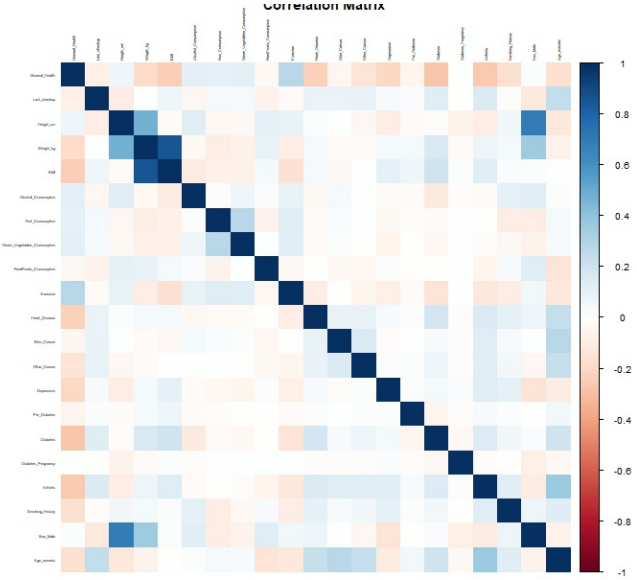


Figure 7: Correlation matrix

Additionally, Age_numeric demonstrates moderate positive correlations with **Heart_Disease** (0.41), **Arthritis** (0.38), and **Diabetes** (0.31), indicating that these conditions tend to be more prevalent in older age groups and suggesting that age is an important risk factor in the development of heart diseases . In contrast, **General_Health** exhibits moderate negative correlations with Heart_Disease (-0.33), Arthritis (-0.32), Diabetes (-0.31), and Age_Category (-0.30), highlighting that poorer self-reported general health is associated with an increased prevalence of these conditions and older age.

It is also important to note the variables **Diabetes_Pregnancy**, **Skin_Cancer**, and **Pre_Diabetes**, which exhibit little to no correlation with the other variables in the dataset. This lack of significant association may suggest that these factors do not share common underlying risk characteristics with the other examined variables.

4.2 Findings of the models

Each model was analyzed for its overall sensitivity, specificity, accuracy and Area Under the Curve (AUC) to determine its strengths and weaknesses in identifying disease presence or absence.

The **Logistic Regression** model achieved a balanced performance, with an AUC of 0.8346, indicating a strong capacity to discriminate between positive and negative cases. It demonstrated a sensitivity of 73.31%, meaning it correctly identified 73.31% of cases where disease is present. The specificity of 79.04% reflects the model's ability to correctly classify 79.04% of true negative cases, or instances without the disease. Together, these metrics suggest the model is reliable in identifying both disease and non-disease cases, making it a useful tool for screening purposes.

For the **Random Forest** model, we applied a 5-fold cross-validation, using four groups for training and one for testing, to find the best configuration. The model performed optimally with an "mtry" parameter set to 3, resulting in an accuracy of approximately 75.75%. To further refine the model, we adjusted the classification threshold between 0.1 and 0.9, observing substantial trade-offs: while lowering the threshold improved sensitivity, it substantially decreased specificity, and vice versa. After evaluating these trade-offs, we determined that a threshold of 0.5 offered the most balanced performance, with a sensitivity of 69.86% and a specificity of 81.63%. The AUC of 0.8285 further highlights the model's robust classification capability, making it an effective tool for clinical contexts that prioritize a balance between correctly identifying positive cases and reducing false positives.

In the case of **XGBoost**, the model achieved a moderate accuracy of 75.41% with an AUC of 0.82526, a sensitivity of 71.11% and a specificity of 79.64%, indicating a reliable performance in recognizing negative cases but with room for improvement in positive case identification. For clinical applications, this model would benefit from further optimization to enhance sensitivity, ensuring that potential disease cases are less likely to be missed.

The **SVM** model similarly demonstrated a balanced but moderate classification capability, achieving an accuracy of 75.75% and an AUC of 0.8241. Its specificity (80.94%) indicates strong identification of negative cases and a sensitivity of70.65%.

Finally, with **K-Nearest Neighbors** (KNN), we selected the optimal value for k by testing a range of odd values between 1 and 20. As k increases, specificity generally improves, while sensitivity slightly decreases. To achieve the best balance, we selected k = 9, which maximizes specificity without substantial loss of sensitivity. The model achieved an overall accuracy of 73.58%, with a sensitivity of 69.38% and a specificity of 77.78%, indicating a reasonable ability to correctly identify non-disease cases.

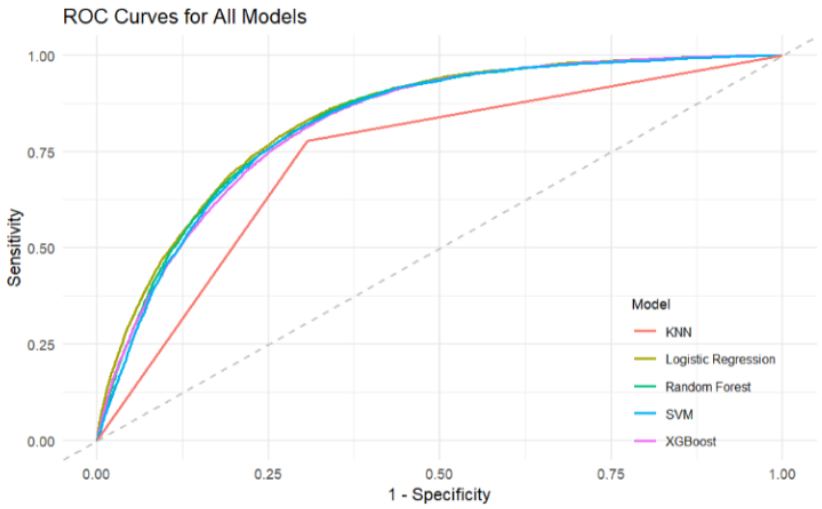
Finally, the **Neural Network** model developed for this study showed promising classification performance, with a sensitivity of 82%, specificity of 71%, and accuracy of 72%. The model's sensitivity is notably high compared to the other machine learning algorithms, underscoring its strength in identifying cases of cardiovascular disease effectively and reducing the likelihood of missed diagnoses. Its specificity of 71%, while slightly lower than other models, indicates reasonable performance in recognizing non-disease cases. Overall, the model's accuracy of 72% positions it as a viable alternative for disease prediction in clinical contexts, particularly in cases where sensitivity is prioritized. These results indicate that, although additional refinement would be beneficial, the neural network could serve as a valuable supplement to traditional methods and improve decision-making support in cardiovascular diagnostics.

MODEL	SENSITIVITY	SPECIFICITY	ACCURACY	AUC
Logistic regression	0.7323	0.7906	0.7614	0.8346094
Random Forest	0.6986	0.8163	0.7575	0.8285761
Gradient Boosting	0.7118	0.7964	0.7541	0.82526
SVM	0.7065	0.8094	0.7579	0.8240967
KNN	0.6938	0.7778	0.7358	0.7357894
Neural Network	0.82	0.71	0.72	0.84

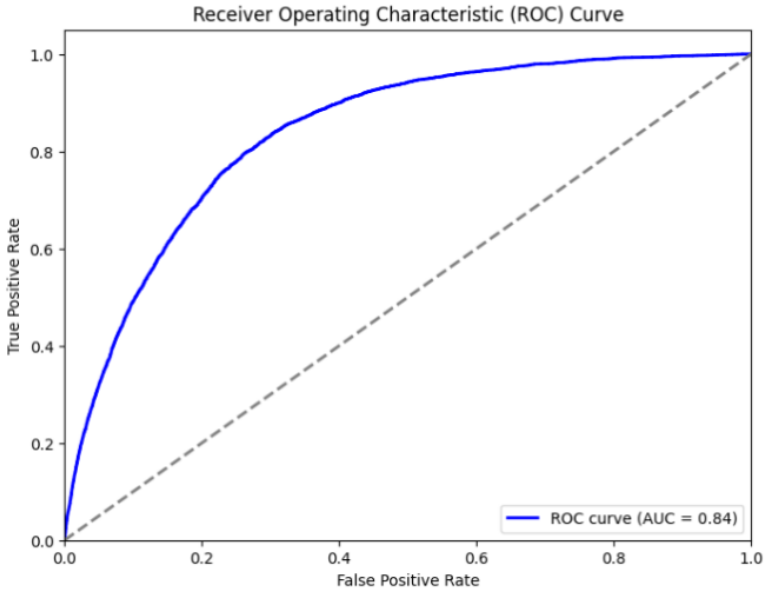
Figure 8: Comparison of metrics performance for various models

The Neural Network model is the most suitable choice for this clinical case, achieving the highest sensitivity (0.82), correctly identifying 82% of the patients with heart disease. Although its specificity is 0.71, sensitivity is more critical in this case, as missing a positive diagnosis could have serious clinical consequences. With an overall accuracy of 72%, the model demonstrates strong performance in distinguishing between disease and non-disease patients. Given the importance of reducing false negatives in a clinical setting, the model's high sensitivity makes it an optimal choice for detecting heart disease.

To further explore model performance, we conducted ROC curve analysis across all models:



The ROC curve analysis revealed that **Logistic Regression** and **Random Forest** consistently outperformed the other models, demonstrating superior discriminatory power. Their curves were consistently closer to the top-left corner of the plot, indicating high sensitivity and specificity. **SVM** and **XGBoost** also exhibited good performance, although slightly below the top-performing models. KNN showed the weakest performance, with its curve being farthest from the ideal point. Overall, these findings suggest that **Logistic Regression** and **Random Forest** are the most suitable algorithms for this particular dataset and classification task. The effectiveness of Logistic Regression, a linear model, implies that our dataset is primarily linear in nature. Any non-linear relationships present have been captured by incorporating interaction terms between variables, allowing the model to better handle complex associations and enhance predictive performance.



In addition, we present the ROC curve of the Neural Network model (in a separate plot, as it was implemented in a Python environment). This curve demonstrates slightly enhanced performance, highlighting the Neural Network's ability to accurately classify patients. This combination of high sensitivity, accuracy and robust ROC performance confirms the Neural Network's value in our clinical setting, where early and precise heart disease detection is crucial.

5 Conclusion

In this study, we investigated key risk factors and various machine learning models for predicting the presence of heart disease. Among the identified correlations, age, diabetes, arthritis, and smoking history showed positive associations with heart disease presence, underscoring the impact of pre-existing health conditions and lifestyle factors on cardiovascular risk. Conversely, General Health exhibited a strong negative correlation with heart disease, indicating that poorer overall health increases the likelihood of developing heart disease. These findings suggest that combining clinical data with lifestyle factors provides a comprehensive view of risk and is crucial for building effective predictive models.

Of the evaluated models, the Neural Network demonstrated the highest performance in identifying heart disease, achieving a sensitivity of 82%, a specificity of 71%, and an overall accuracy of 72%. This high sensitivity is particularly relevant in a clinical context, where minimizing false negatives is essential to avoid missed diagnoses that could have serious consequences. Although the model's specificity is moderate, it remains effective for detecting at-risk heart disease patients, making it a viable option for heart disease prediction. Other models, such as Logistic Regression and Random Forest, also showed robust performance, especially in terms of specificity, making them suitable for applications where reducing false positives is a priority.

Despite the promising results, the overall evaluation of the models reveals limitations in prediction accuracy. This is understandable, given that heart diseases are complex and their development depends on multiple clinical, demographic, and lifestyle factors, some of which are not included in our dataset. The inclusion of additional variables could significantly improve model performance. Specifically, adding genetic markers, specific biomarkers (such as cholesterol levels and blood pressure), and detailed family medical history data could provide a more nuanced view of heart disease risk. Additionally, psychological or quality-of-life variables, such as stress levels or sleep patterns, could also yield relevant insights.

This analysis demonstrates that machine learning has great potential for heart disease prediction, but also highlights the need for continuous improvement. Future research could benefit from incorporating a wider range of medical and behavioral health data, which could allow models not only to be more accurate but also more useful in clinical practice, helping to identify individuals at risk and facilitating early, personalized interventions to improve heart health on a population level

References

[1] Keith B Allen, James E Alexander, Joshua N Liberman, and Susan Gabriel. Correction to: Implications of payment for acute myocardial infarctions as a 90-day bundled single episode of care: A cost of illness analysis. *PharmacoEconomics-Open*, 6(6):897–897, 2022.

[2] Sadia Arooj, Saif ur Rehman, Azhar Imran, Abdullah Almuhaimeed, A Khuzaim Alzahrani, and Abdulkareem Alzahrani. A deep convolutional neural network for the early detection of heart disease. *Biomedicines*, 10(11):2796, 2022.

[3] Stephanie E Chiuve, Nancy R Cook, Christina M Shay, Kathryn M Rexrode, Christine M Albert, JoAnn E Manson, Walter C Willett, and Eric B Rimm. Lifestyle-based prediction model for the prevention of cvd: The healthy heart score. *Journal of the American Heart Association*, 3(6):e000954, 2014.

[4] Andrea D'Agostino. Algoritmi di machine learning: Guida introduttiva per comprendere i principi e le applicazioni. https://www.diariodiunanalista.it/posts/algoritmi-di-machine-learning-guida-introduttiva-per-comprendere-i-principi-e-le-applicazioni/. Sep 30, 2023.

[5] griko. Answer to a question. https://datascience.stackexchange.com/questions/117189/relation-between-underfitting-vs-high-bias-and-low-variance. Accessed: Dec 20, 2022.

[6] Roshna S H. K-nearest neighbors algorithm. https://intuitivetutorial.com/2023/04/07/k-nearest-neighbors-algorithm/.

[7] Nikolai Khaltaev and Svetlana Axelrod. Countrywide cardiovascular disease prevention and control in 49 countries with different socio-economic status. *Chronic Diseases and Translational Medicine*, 8(04):296–304, 2022.

[8] Chayakrit Krittanawong, Hafeez Ul Hassan Virk, Sripal Bangalore, Zhen Wang, Kipp W Johnson, Rachel Pinotti, HongJu Zhang, Scott Kaplin, Bharat Narasimhan, Takeshi Kitai, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific reports*, 10(1):16057, 2020.

[9] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7:81542–81554, 2019.

[10] Medline Plus. Cómo prevenir las enfermedades del corazón. https://medlineplus.gov/spanish/howtopreventheartdisease.html. Accessed: Jun 19, 2024.

[11] Gopal Krishna Ranjan. data analysis, data preprocessing, data science. https://sqlrelease.com/introduction-to-k-fold-cross-validation-in-python. Jul 12, 2021.

[12] Sivakannan Subramani, Neeraj Varshney, M Vijay Anand, Manzoore Elahi M Soudagar, Lamya Ahmed Al-Keridis, Tarun Kumar Upadhyay, Nawaf Alshammari, Mohd Saeed, Kumaran Subramanian, Krishnan Anbarasu, et al. Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in medicine*, 10:1150933, 2023.