**Disease Risk Stratification Using Machine Learning**

**Data 300 Project by Rahma Semma and Nishtha Sharma**

[disease\_prediction\_model.ipynb](https://colab.research.google.com/drive/1d4bm6m2TKdGv8BkRvMq7QT_MUwjCQNGh#scrollTo=SM0wvQpxw2ep)

# **1. Introduction**

In recent years, machine learning has revolutionized predictive analysis in healthcare, particularly in disease risk stratification. Heart disease remains one of the leading causes of mortality worldwide, and its prevention is a critical priority for both public health and individual well-being. Predicting heart disease risk is not only about saving lives but also about reducing the societal and economic burdens caused by late-stage diagnoses and expensive treatments.

Machine learning enables us to analyze large, complex datasets to identify patterns and relationships that traditional statistical models may overlook. These models offer a scalable, data-driven approach to identifying risk factors, enabling early intervention and personalized treatments. Projects like this are particularly important because they demonstrate how machine learning can bridge the gap between data analysis and healthcare solutions. By empowering clinicians with accurate predictions and interpretable insights, machine learning contributes to evidence-based decision-making, improved patient care, and proactive disease prevention.

The primary objective of this project was to build a machine learning pipeline to identify key predictors of heart disease and evaluate models for their predictive performance. We employed Random Forest and LightGBM algorithms alongside feature selection techniques to achieve reliable and interpretable results. The results obtained showcase the potential of machine learning in transforming healthcare by providing insights that guide early diagnosis, risk stratification, and preventive strategies. This work highlights how technology can support health systems in addressing real-world problems and improving patient outcomes.

# **2. Problem Statement and Importance**

Heart disease is a preventable condition, yet it remains a significant public health issue. Early detection of risk factors is crucial to improving treatment outcomes and reducing mortality. Traditional statistical models often fail to capture complex patterns in data, limiting their predictive accuracy.

**Machine learning addresses these limitations by:**

* Identifying complex, non-linear relationships between features and outcomes.
* Ranking key predictors to guide medical decision-making.
* Offering scalable models capable of handling large datasets.

**This project aimed to answer the following questions:**

1. ***What features (e.g., age, cholesterol levels, blood pressure) are most predictive of heart disease?***
2. ***How do machine learning models like Random Forest and LightGBM compare in terms of predictive performance?***
3. ***How can interpretability tools, such as SHAP analysis, help explain model results to medical practitioners?***

Addressing these questions is important because accurate risk stratification enables early interventions, reduces healthcare costs, and saves lives.

# **3. Methods**

We implemented a step-by-step machine learning pipeline to analyze the data and achieve the project goals. The main steps included data preprocessing, feature selection, model training, evaluation, and interpretability analysis.

## **3.1 Data Preprocessing**

The dataset used in this project is a subset of the Framingham Heart Study, a landmark epidemiological study conducted in Framingham, Massachusetts. This long-term prospective study aimed to identify the causes of cardiovascular disease and introduced the concept of risk factors in public health.

The dataset includes 4,240 observations and 16 variables, including:

* **sex**: Gender of the participants, represented as a binary variable named "male" (1 for male, 0 for female).
* **age**: Age of the participant at the time of examination (in years).
* **education**: Educational attainment, categorized as: 1 = Some high school, 2 = High school/GED, 3 = Some college/vocational school, 4 = College.
* **currentSmoker**: Indicates whether the participant is currently smoking (1 = yes, 0 = no).
* **cigsPerDay**: Number of cigarettes smoked per day.
* **BPmeds**: Use of anti-hypertensive medication at the time of the examination (1 = yes, 0 = no).
* **prevalentStroke**: History of stroke (1 = yes, 0 = no).
* **prevalentHyp**: Prevalent hypertension (1 = yes, 0 = no).
* **diabetes**: Indicates whether the participant is diabetic (1 = yes, 0 = no).
* **totChol**: Total cholesterol level (in mg/dL).
* **sysBP**: Systolic blood pressure (in mmHg).
* **diaBP**: Diastolic blood pressure (in mmHg).
* **BMI**: Body mass index, calculated as weight (kg) / height (m)^2.
* **heartRate**: Heart rate (in beats per minute).
* **glucose**: Blood glucose level (in mg/dL).
* **TenYearCHD**: The target variable, indicating the 10-year risk of coronary heart disease (1 = risk, 0 = no risk).

Before training models, we:

* Removed missing values to ensure model robustness (dropped rows).
* Standardized features using StandardScaler to normalize their scales.

**Handling Missing Data:** Although glucose had the highest percentage of missing data, we retained it in the analysis due to its known clinical importance as a predictor of heart disease. Eliminating glucose would have removed valuable information, potentially decreasing the model's accuracy and relevance. Instead, we focused on other strategies like feature selection to ensure the model performed well despite data limitations.

**Importance:** Preprocessing is critical because raw data often contains inconsistencies and missing values that can negatively impact machine learning models. Standardizing data ensures fair comparisons across features.

## **3.2 Feature Selection with Recursive Feature Elimination (RFE)**

To identify the most significant predictors of heart disease, we used **Recursive Feature Elimination (RFE)** with a Random Forest classifier. RFE iteratively removes less important features based on their importance scores until the desired number of top features is selected.

**The top 5 features identified were:**

* *Age*
* *Total Cholesterol (totChol)*
* *Cigarettes Per Day (cigsPerDay)*
* *Glucose Levels*
* *Systolic Blood Pressure (sysBP)*

**Importance:** Feature selection reduces model complexity, improves computational efficiency, and enhances interpretability by focusing on the most relevant predictors.

## **3.3 Model Training**

We compared two machine learning algorithms:

1. **Random Forest**: A robust ensemble method that builds multiple decision trees and averages their results.
2. **LightGBM (Light Gradient Boosting Machine)**: A fast, high-performance gradient boosting algorithm optimized for large datasets.

We trained both models on the selected features (from RFE) and evaluated their performance using the AUC-ROC score, a widely used metric for binary classification tasks.

**Importance:** Comparing models helps identify the most accurate and efficient algorithm for disease prediction. LightGBM is particularly useful due to its speed and ability to handle imbalanced data.

## **3.4 Evaluation and Visualization**

* **ROC Curves**: We plotted ROC curves for both Random Forest and LightGBM to compare their ability to distinguish between patients at high and low risk of heart disease.
* **SHAP Analysis**: We used SHAP (SHapley Additive exPlanations) to interpret the LightGBM model. SHAP values explain the contribution of each feature to the model's predictions, providing transparency.

**Importance:** Model evaluation ensures reliability and trustworthiness, while interpretability tools like SHAP make machine learning models understandable to healthcare professionals.

# **4. Results**

The following results were obtained from the analysis:

## **4.1 Feature Selection Results**

RFE identified the following top 5 features as the most predictive of heart disease:

1. *Age*
2. *Total Cholesterol (totChol)*
3. *Cigarettes Per Day (cigsPerDay)*
4. *Glucose Levels*
5. *Systolic Blood Pressure (sysBP)*

These results align with medical literature, where factors like age and cholesterol are known risk indicators for heart disease.

## **4.2 Model Performance**

* **Random Forest** achieved an AUC-ROC score of **0.85**.
* **LightGBM** achieved a slightly higher AUC-ROC score of **0.88**.

The ROC curves showed that both models performed well, but LightGBM outperformed Random Forest slightly in terms of predictive accuracy.

**What This Means:**

* **AUC-ROC** measures how well the model differentiates between positive and negative cases.
* A score of 0.88 indicates that LightGBM has strong predictive performance, with a good balance of sensitivity and specificity.

## **4.3 SHAP Analysis**

SHAP analysis revealed the following:

* **Age** and **Systolic Blood Pressure** had the highest SHAP values, indicating they were the most influential features in predicting heart disease.
* Features like **Total Cholesterol** and **Glucose Levels** also significantly contributed to the model's predictions.

The SHAP summary plot visually confirmed the importance of these features and highlighted their impact on the model's decisions.

**Why It Matters:**

* SHAP values provide transparency, explaining how each feature impacts predictions. This makes machine learning models trustworthy and interpretable for healthcare practitioners.

## **4.4 Real-World Insights**

* **Young Adults at Risk**: Among patients under 45 years of age, approximately 6.35% showed a significant risk of heart disease. This highlights that while heart disease is often perceived as an issue for older adults, younger populations are not immune. Factors like unhealthy diets, smoking, stress, and sedentary lifestyles may contribute to this trend. The findings emphasize the urgent need for targeted awareness campaigns and interventions among young adults to encourage healthier habits and routine health check-ups.
* **Preventability**: Early detection of high-risk individuals allows for lifestyle changes, medical treatments, and preventive care to reduce the incidence of heart disease. For example, patients with elevated blood pressure or cholesterol levels can be advised on dietary modifications, physical activity, and, if necessary, medications. Preventive measures, when implemented early, can significantly reduce the burden on healthcare systems and improve long-term patient outcomes.
* **Feature Importance and Real-Life Impact**: The identification of key predictors such as age, systolic blood pressure, and cholesterol levels aligns with real-world medical findings. These results can help healthcare professionals prioritize screenings and treatments for patients exhibiting these risk factors. Additionally, tools like SHAP provide insights into the relative contribution of each predictor, enabling doctors to explain and communicate risks to patients more effectively.
* **Personalized Medicine**: The findings pave the way for personalized interventions, where treatments and recommendations can be tailored based on an individual’s risk profile. For instance, a patient identified as high risk due to their smoking history and high glucose levels can be targeted with specific smoking cessation programs and blood sugar management plans.

# **5. Significance and Impact**

The findings from this project have several significant implications:

1. **Improved Predictive Models**: Machine learning algorithms like LightGBM can accurately predict heart disease risk, providing a reliable alternative to traditional methods.
2. **Key Risk Factors Identified**: The project highlighted key features (e.g., age, cholesterol, blood pressure) that healthcare providers can monitor for early detection.
3. **Interpretability with SHAP**: By making models interpretable, SHAP analysis bridges the gap between machine learning and clinical decision-making, increasing trust in AI-based predictions.
4. **Young Adult Insights**: Early identification of at-risk young adults emphasizes the need for education on stress management, smoking cessation, and regular health monitoring.

This project showcases the power of machine learning in solving real-world problems, offering insights that can improve public health outcomes.

# **6. Conclusion**

In this project, we implemented and evaluated machine learning models to predict heart disease risk. Key steps included data preprocessing, feature selection using RFE, training Random Forest and LightGBM models, and interpreting results using SHAP analysis. The results showed that LightGBM achieved strong performance with an AUC-ROC score of **0.88**.

Projects like this are particularly important because they demonstrate how machine learning can bridge the gap between data analysis and healthcare solutions. For doctors, such tools are helpful as they provide not only accurate predictions but also interpretable results that can easily be integrated into clinical workflows. By highlighting key risk factors like age and blood pressure, the results allow doctors to focus on high-risk patients, prioritize interventions, and provide tailored recommendations.

The ability to explain predictions using SHAP analysis ensures that these models are not black-box solutions but transparent tools that clinicians can trust. By identifying critical risk factors and leveraging interpretable machine learning models, this project demonstrates a practical approach to disease risk stratification. Such methods can guide early interventions, reduce healthcare costs, and ultimately save lives.

### **Collaboration and Contribution Statement**

We worked collaboratively throughout this project, ensuring equal participation in all stages of the work.

1. **Coding and Analysis:**
   * We used **Google Colab** to write, execute, and debug the code together. This platform allowed us to collaborate in real-time, share ideas, and contribute equally to data preprocessing, model building, evaluation, and interpretability analysis. All decisions, including feature selection, model choices, and SHAP analysis, were discussed and agreed upon as a team.
2. **Presentation Slides and Poster Design:**
   * We created PowerPoint **slides** and a **poster** using shared tools like **Slides** and shared documents. Both of us contributed to designing the layout, adding content, refining visuals, and ensuring that the final materials were clear, cohesive, and professional.
3. **Presentation Delivery:**
   * We presented the project equally, dividing the speaking roles to explain the methods, results, and conclusions. Each of us took responsibility for specific sections, ensuring a balanced and organized delivery.
4. **Report Preparation:**
   * The report was drafted collaboratively, with both of us contributing to writing, editing, and formatting. We reviewed and finalized all sections together to ensure the content was consistent and well-structured.