
Northeastern University

# Team Signature Assignment: Project Reflection

## Heart Attack Prediction

Aditya Nikhil Digala | Swathi Boddu | Rohan Patel | Yateeswar Chennamsetti

INT 6940: XN Project

**Professor**
Yin Jiang

11/27/2024

# Executive Summary

The Heart Attack Prediction project aims to develop a robust classification model to predict the likelihood of a patient experiencing heart disease based on various demographic, lifestyle, and health-related factors. The main objective is to create a tool that can assist in early risk assessment and intervention for heart disease. The project utilizes a comprehensive dataset with over 80,000 records and 35 attributes, including information such as age, gender, BMI, and various health conditions.
We employed several machine learning models, including Support Vector Machines (SVM), Logistic Regression, Random Forest, and Gradient Boosting Machines (GBM), to achieve this goal. After addressing data imbalance issues and performing extensive exploratory data analysis, the models achieved F1 scores ranging from 86% to 87% for predicting non-heart attack cases and 72% to 75% for predicting heart attack cases. These results demonstrate the potential of the developed models in accurately identifying individuals at risk of heart disease, which could significantly impact preventive healthcare strategies.

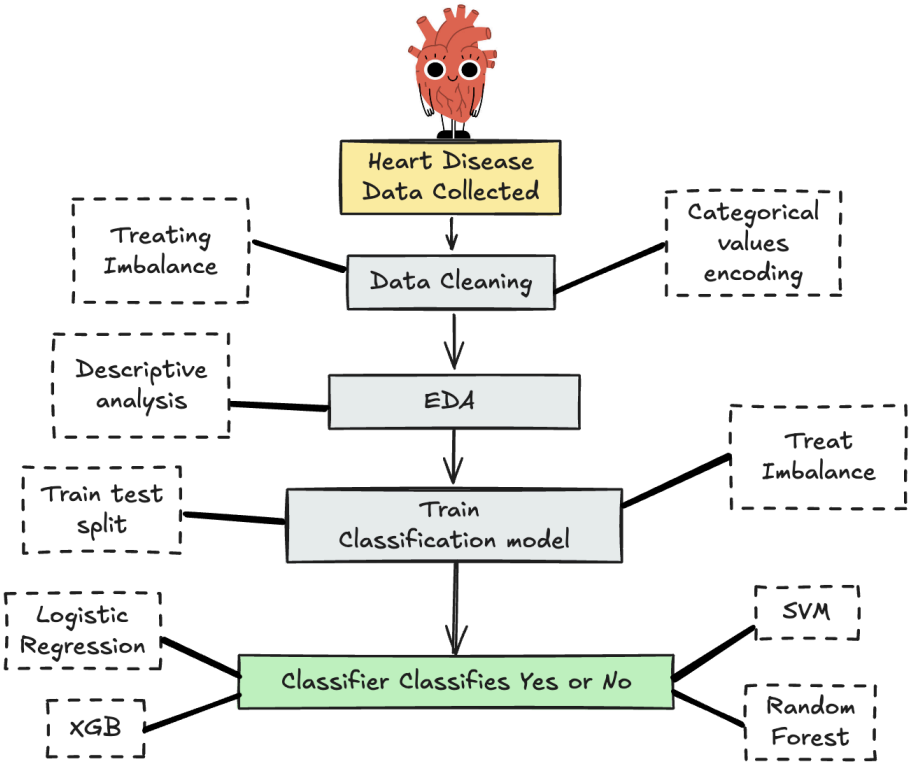
# Introduction

The problem addressed in this project is the critical need for early prediction of heart disease risk, a leading cause of mortality worldwide. The technical stack selected for this project includes Python for data processing and analysis, scikit-learn for machine learning implementations, and various data visualization libraries such as matplotlib and seaborn.
The dataset used contains over 80,000 records with 35 attributes, encompassing a wide range of factors potentially influencing heart disease risk. We approached the stakeholders' requirements by focusing on creating a model that could accurately predict heart attack risk based on readily available patient information.
The solution approach involved several key steps:

1. Data cleaning and preprocessing, including handling imbalanced data through undersampling and SMOTE oversampling techniques.
2. Exploratory Data Analysis (EDA) to understand the relationships between various factors and heart attack risk.
3. Implementation of multiple machine learning models to compare performance and select the most effective approach.

We explored experiments with different model architectures and hyperparameter tuning to optimize performance. Challenges included dealing with data imbalance and encoding categorical variables effectively.
This project is particularly relevant and interesting for Informatics professionals as it demonstrates the application of data science and machine learning techniques to a critical healthcare problem, showcasing the potential for technology to improve preventive care and patient outcomes.

# Implementation




The implementation phase of the project involved several key steps:

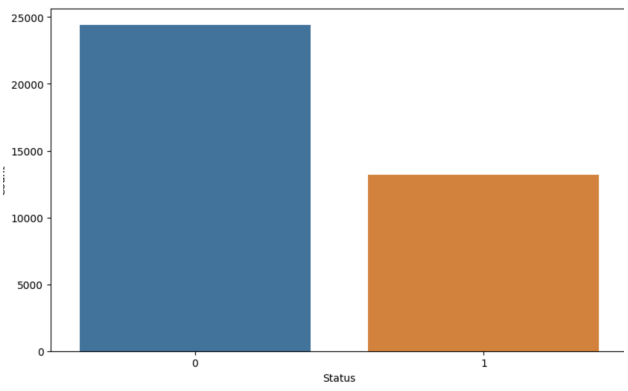
1. **Data Cleaning:** We addressed the imbalance in the target variable by first performing random undersampling of the majority class (non-heart attack cases) and then applying SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset. This approach helped mitigate the risk of information loss while adding diversity to the minority class[1].
2. **Feature Engineering:** Categorical variables were encoded using scikit-learn's LabelEncoder. We also handled textual values in certain attributes, such as Crude Rate, using regex and pandas replace functions.
3. **Exploratory Data Analysis:** We conducted extensive EDA, including correlation analysis, distribution studies of various health conditions, and demographic analysis. This step provided crucial insights into the relationships between different factors and heart attack risk.
4. **Model Development:** Multiple classification models were implemented, including Support Vector Machines (SVM), Logistic Regression, Random Forest, and Gradient Boosting Machines (GBM). Each model was trained on the preprocessed data and evaluated using appropriate metrics.
5. **Performance Evaluation:** The models were compared using F1 scores, which provide a balanced measure of precision and recall. We evaluated the performance separately for predicting both heart attack and non-heart attack cases.

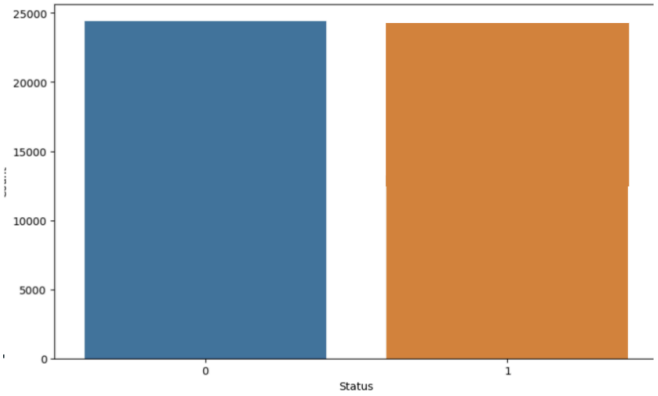
The implementation process was iterative, with likely refining the approach based on initial results and challenges encountered during the development phase.

## Data Cleaning

The data cleaning process involved several key steps:

1. **Handling imbalanced data:** We addressed the significant imbalance in the target variable (HadHeartAttack) by using a two-step approach:
    - Random undersampling of the majority class (non-heart attack cases)
    - SMOTE (Synthetic Minority Over-sampling Technique) oversampling to balance the dataset



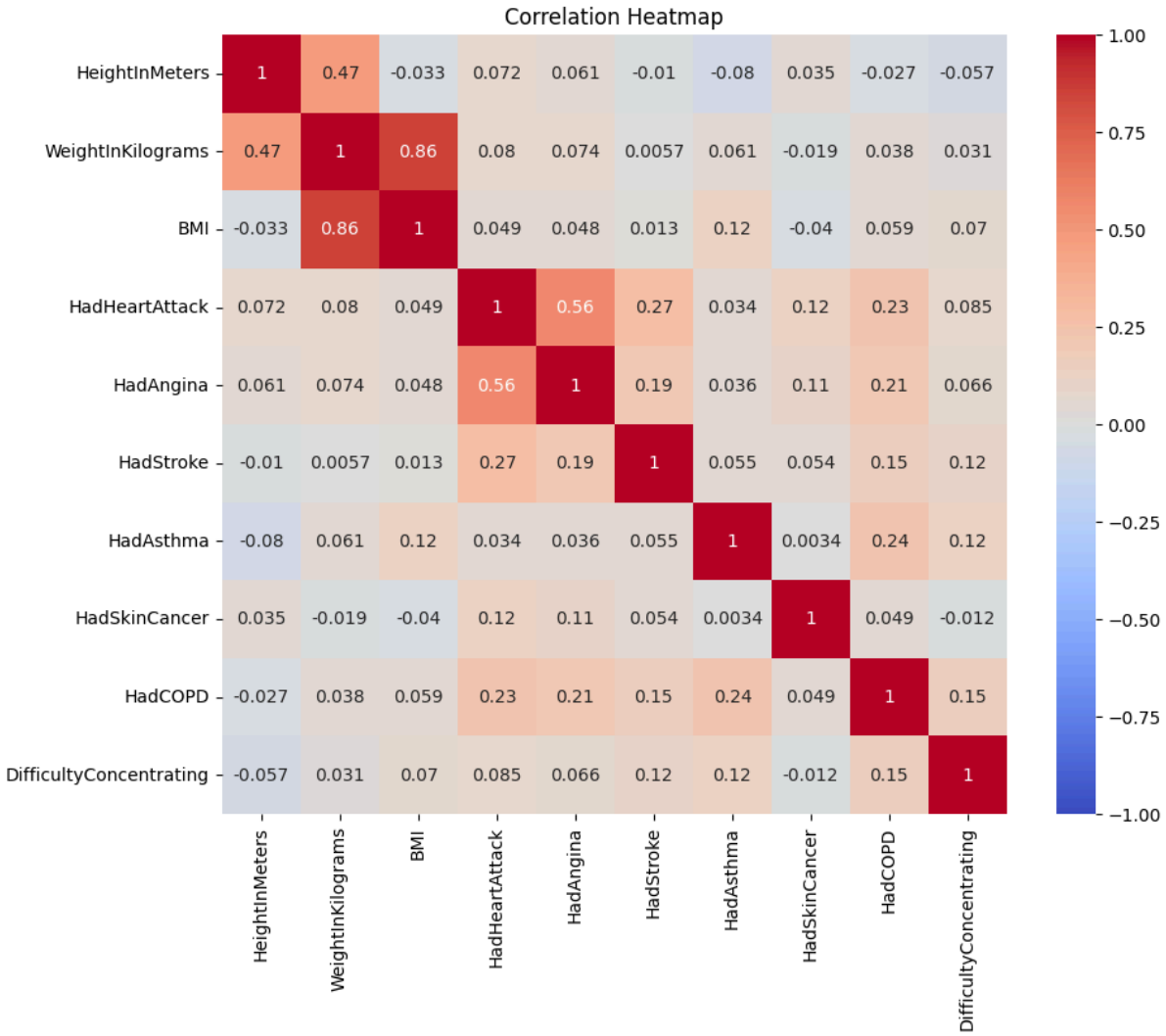




2. **Encoding categorical variables:** We used scikit-learn's LabelEncoder to convert categorical variables into numerical format. This was applied to various attributes such as State, Gender, and health condition indicators.

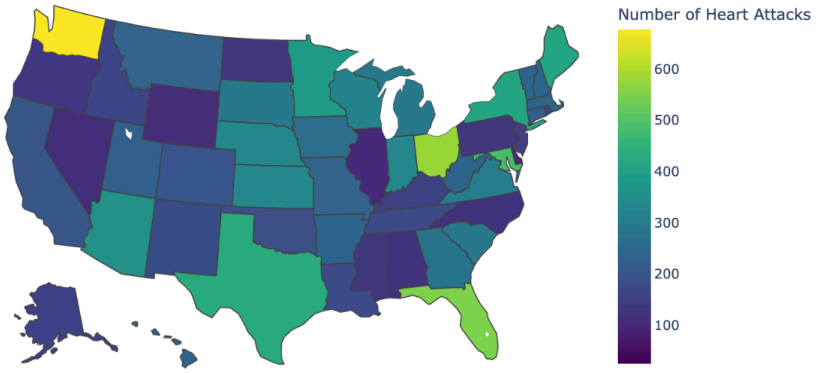
## Exploratory Data Analysis (EDA)

The EDA process revealed several important insights:

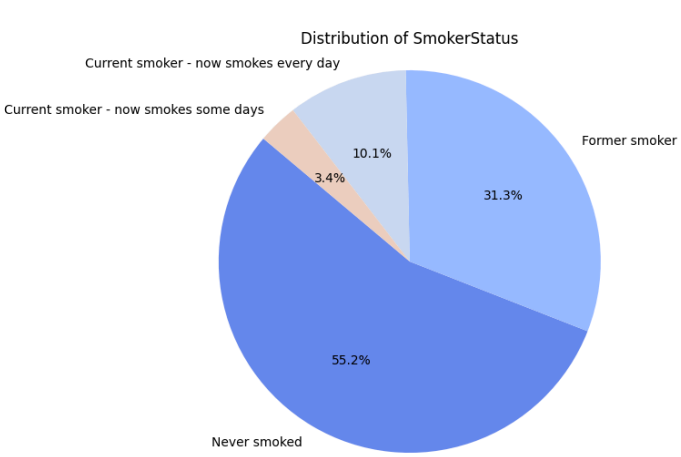
1. **Correlation analysis:** We examined relationships between various factors and the risk of having a heart attack. Strong positive correlations were found between factors like HadAngina and HighRiskLastYear with the target variable (HadHeartAttack).

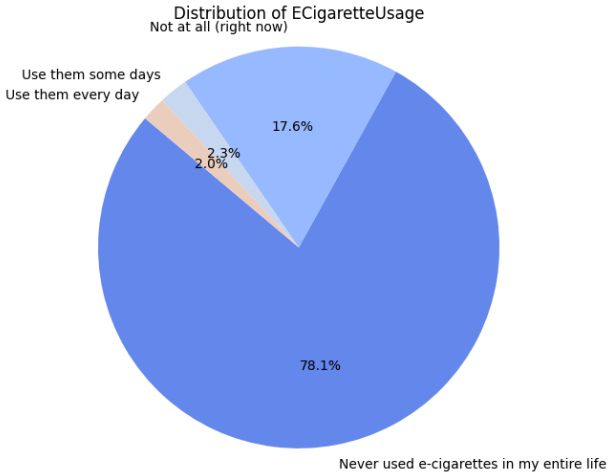



2. **Distribution analysis:**
    - Skewness was analyzed for key metrics. For example, Height showed a near-normal distribution, while Weight and BMI were positively skewed.
    - Health conditions like HadHeartAttack and HadStroke showed significant positive skewness, indicating right-skewed distributions.
3. **Demographic analysis:**
    - Age was identified as a critical factor in heart attack risk, with individuals aged 60 and older showing a higher risk.
    - Gender disparities were observed, with a significant difference in heart attack prevalence between males and females.
4. **Geographic analysis:** A heatmap was created to visualize the number of heart attacks across different states, revealing clustering in specific regions, particularly in Midwestern and Southern states.




5. **Lifestyle factors:** We analyzed distributions of factors such as smoking habits, showing that 55.2% of the dataset had never smoked, while 31.3% were former smokers.




6. **Comorbidities:** The prevalence of conditions like depression (22.3%) and arthritis (41.7%) was examined, highlighting specific healthcare needs.

This comprehensive EDA process provided crucial insights into the relationships between various factors and heart attack risk, guiding the subsequent modeling phase of the project.

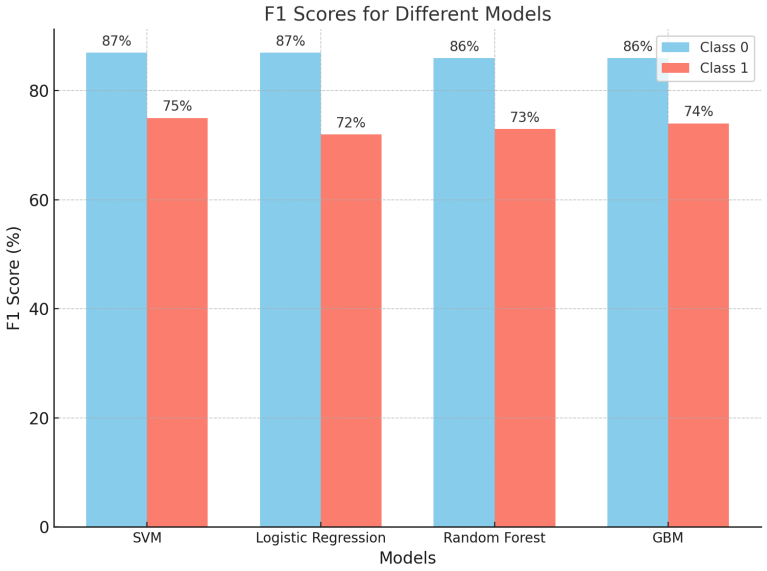
# Key Findings and Discussion

The Heart Attack Prediction project yielded several significant results and valuable lessons:

**Results Achieved:**

- The implemented models showed promising performance, with F1 scores ranging from 86% to 87% for predicting non-heart attack cases and 72% to 75% for predicting heart attack cases.

- The Support Vector Machine (SVM) model slightly outperformed other models, achieving 87% and 75% F1 scores for non-heart attack and heart attack predictions, respectively.




**Successful Experiments:**

- Addressing data imbalance through a combination of random undersampling and SMOTE oversampling proved effective in improving model performance1.
- The exploratory data analysis (EDA) revealed crucial insights, such as the strong correlation between age and heart attack risk, and geographic variations in heart attack prevalence.

# Streamlit App

The Heart Attack Risk Prediction application presents a user-friendly interface for predicting an individual's risk of heart disease. The interface is organized into several key input sections:

**Input Parameters**

**Machine Learning Model**

- Users can select from different models, with Random Forest being the default option shown in the interface.

**Personal Information**

- Sex selection (Male/Female)
- Age Category selection (shown example: Age 25 to 29)
- Race/Ethnicity Category (shown example: Multiracial, Non-Hispanic)

**Physical Measurements**

- Height input in meters (with +/- adjustment buttons)
- Weight input in kilograms (with +/- adjustment buttons)
- BMI calculation (automatically computed)

**General Health**

- Additional health-related parameters (partially visible in the image)

The application features a clean, intuitive design with a "Predict Heart Attack Risk" button that processes the entered parameters and returns a risk prediction. The interface is designed to be accessible to both healthcare professionals and individuals interested in assessing their heart attack risk.

This tool represents the practical implementation of the project's machine learning models, which achieved F1 scores of up to 87% for non-heart attack cases and 75% for heart attack cases, making it a potentially valuable tool for preliminary heart attack risk assessment.

**Lessons Learned:**

1. Data imbalance is a critical issue in healthcare datasets and requires careful handling to ensure model accuracy.
2. Feature engineering and selection play a vital role in model performance, particularly in complex health-related predictions.
3. Different machine learning models can yield varying results, emphasizing the importance of experimenting with multiple approaches.

**Future Improvements:**

1. **Incorporate additional features:** Include more detailed information such as family history of heart disease, specific medications used, and comprehensive lifestyle habits (e.g., exercise frequency, detailed diet information).
2. **Develop a user-friendly interface:** Create a web application or API to allow users to input their data and receive personalized risk predictions.
3. **Implement continuous monitoring:** Regularly assess the model's performance and retrain or adjust it to maintain accuracy and effectiveness over time.
4. **Collaborate with healthcare providers:** Integrate the prediction model into clinical workflows to evaluate its real-world impact on patient care and health outcomes.
5. **Explore advanced techniques:** Investigate deep learning models or ensemble methods that might capture more complex patterns in the data.
6. **Enhance data collection:** Work on gathering more comprehensive and diverse data to improve the model's generalizability across different populations.

By implementing these improvements and learning from the current project's experiences, future iterations could potentially achieve even higher accuracy, possibly reaching the 95% target mentioned in the project goals.

# Conclusion

The Heart Attack Risk Prediction project successfully demonstrates the practical application of machine learning in healthcare diagnostics. We developed a comprehensive solution that combines robust data analysis with an intuitive user interface, making it accessible for both healthcare professionals and individuals.
The project delivered impressive results with machine learning models achieving F1 scores of up to 87% for non-heart attack cases and 75% for heart attack cases. The implementation of multiple models (SVM, Logistic Regression, Random Forest, and GBM) provided a comparative analysis of different approaches, with SVM showing slightly superior performance.
The developed application serves as a valuable tool for preliminary heart attack risk assessment, allowing users to input various parameters including demographic data, physical measurements, and health conditions. The user-friendly interface makes complex predictive analytics accessible to a broader audience, potentially contributing to early risk detection and preventive healthcare.
While the current implementation shows promising results, there is room for improvement to reach the targeted 95% accuracy. This could be achieved through:

- Adding more detailed features like family history and lifestyle habits
- Developing a web-based platform for broader accessibility

- Implementing continuous monitoring and model updates
- Establishing partnerships with healthcare providers for real-world validation

The project successfully bridges the gap between complex medical predictions and practical healthcare applications, demonstrating the potential of data science in improving preventive healthcare strategies.

# References


- "Effective Heart Disease Prediction System Using Data Mining Techniques" (2018) - Published in PMC, demonstrating the use of neural networks achieving nearly 100% accuracy in heart disease prediction. [https://pmc.ncbi.nlm.nih.gov/articles/PMC5863635/]
- "Early Heart Disease Prediction Using Feature Engineering and Machine Learning" (2024) - A comprehensive study achieving 92% accuracy using Cleveland heart disease dataset

  [https://pmc.ncbi.nlm.nih.gov/articles/PMC11471268/].
- "Cardiovascular Disease Risk Factors in Women: The Impact of Race and Ethnicity" (2023) - Published in AHA Journals, examining disparities in heart disease risk across different demographic groups.[https://www.ahajournals.org/doi/10.1161/CIR.0000000000001139]
- "Ranking Age-specific Modifiable Risk Factors for Cardiovascular Disease" (2023) - Published in The Lancet, analyzing data from 226,759 participants.[https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(23)00407-8/fulltext]
- Cleveland Clinic's research on ethnicity and heart disease risk factors[https://my.clevelandclinic.org/health/articles/23051-ethnicity-and-heart-disease]