



THE UNIVERSITY OF  
**SYDNEY**

---

## Project Report for ENGG2112

*Machine Learning-Based Heart Attack Risk Prediction: A  
Data-Driven Method for Early Identification and Prevention*

---

Elvis Nguyen, 510448444, Software Engineering  
Georgina Young, 520485583, Biomedical Engineering  
Kiet Chau, 530525808, Biomedical Engineering  
Alex Unterweger, 530439505, Chemical Engineering

FACULTY OF ENGINEERING

November 3, 2024

## Executive Summary

Heart attacks are a significant global concern as it sits as the leading cause of death worldwide, taking an estimated 17.9 million lives each year. [1] Our report addresses this issue by exploring the implementation of machine learning models to predict the risk of heart attacks. Almost 2In our analysis of two key datasets, the “Heart Attack Analysis Prediction Dataset” and the “Heart Attack Risk Prediction Dataset,” we evaluated a diverse set of diagnostic parameters, ranging from clinical markers like fasting blood sugar and restECG to lifestyle factors such as BMI and smoking habits. We employed data exploration and pre-processing techniques to ensure quality and consistency before applying machine learning models like Logistic Regression, Random Forest, Support Vector Machine (SVM), and Multilayer Perceptron (MLP). These models were tuned to predict the likelihood of heart attacks effectively. Our research found that the clinical dataset produced strong predictive accuracy, with models like KNN and WNN achieving an ROC-AUC score of 0.91, while logistic regression and random forest models followed closely with 0.90. In contrast, the lifestyle dataset performed poorly, with models reaching only around 0.50 ROC-AUC, suggesting near-random performance. This discrepancy highlights the stronger predictive power of straightforward clinical indicators over complex lifestyle data, which often lacks distinct predictive features. While these results are promising, they also point to areas needing further development. To practically implement these models in real-world healthcare, additional fine-tuning and validation on larger, more diverse datasets are required. Future work should also explore combining clinical and lifestyle factors to create a more holistic and accurate model of heart attack risk. The significance of this development is that early detection of the risk of a heart attack can ensure preventive measures are adopted to enhance overall heart health and allow individuals to have better development in terms of long-term survival and effective treatments. By enabling early detection, personalised treatment recommendations, great time and cost efficiency, and higher diagnostic accuracy, these machine learning models have the potential to transform the future of heart attack prevention and treatment research. In summary, this report demonstrates the potential of machine learning to significantly enhance heart attack risk prediction. Our research emphasises early detection, improved patient outcomes, and more efficient healthcare practices, though further refinement is needed to enhance the predictive power of lifestyle-related data.

# Contents

<b>1</b>	<b>Background and Motivation</b>	<b>1</b>
<b>2</b>	<b>Objectives and Problem Statement</b>	<b>1</b>
2.1	Problem Statement . . . . .	1
2.2	Objectives . . . . .	1
<b>3</b>	<b>Methodology</b>	<b>2</b>
3.1	Data Exploration . . . . .	2
3.2	Data Pre-Processing . . . . .	3
3.3	Model Training . . . . .	4
<b>4</b>	<b>Simulation Results</b>	<b>5</b>
4.1	Key Findings and Significance . . . . .	5
4.2	Issues Faced . . . . .	8
<b>5</b>	<b>Potential for Wider Adoption and Practical Applications</b>	<b>8</b>
<b>6</b>	<b>Conclusions</b>	<b>10</b>

# 1 Background and Motivation

This project focuses on the challenge of early detection of heart disease and heart attacks due to the urgent need for enhanced prediction methods which accurately evaluate individuals who are at risk of a heart attack. Supervised learning algorithms have contributed a vital role within the healthcare industry by allowing the development of predictive models to forecast individuals outcomes by using historical data (Varnosfaderani Forouzanfar, 2024) Varnosfaderani and Forouzanfar (2024). Replacing traditional medical diagnostics with Machine Learning methods it provides a more affordable and automated solution for predicting heart attacks

We focused on two datasets, “Heart Attack Analysis Prediction Dataset” - lifestyle dataset, and “Heart Attack Risk Prediction Dataset” - clinical dataset, which gave us a set of diverse parameters to explore with the use of Machine Learning models. The lifestyle dataset provided us with 26 mixed features and 8763 rows and the clinical provided 14 numerical features and 202 rows, both being obtained from Kaggle.com. Throughout our project we used multiple different learning models including Random Forest, K-Nearest Neighbours (KNN), Weighted Nearest Neighbours (WNN), Multi-Layer Perception (MLP), and Support Vector Classifier (SVC). These training methods allowed us to analyse multiple different characteristics of an individuals health profile to forecast the likeliness of a heart attack in order to achieve an elevated diagnostic accuracy and minimise false negatives.

## 2 Objectives and Problem Statement

### 2.1 Problem Statement

Heart disease is one of the most common causes of death around the globe, with 702,880 people dying from heart disease in 2022 in the United States (CDC, 2024)for Disease Control and Prevention (2024). The problem addressed in our project discusses the prediction of heart attacks/disease with the use of a machine-learning model. Early detection is vital in preventing these fatal outcomes, however, thousands of individuals at risk stay undiagnosed until it’s too late to prevent it.

### 2.2 Objectives

The goal of this project is to address the issue of heart disease by using machine learning techniques to create a model to predict individuals with the risk of a heart attack. We used two datasets, one being clinical data and one being lifestyle data, in order for the machine learning model to analyse many different variables that may indicate an approaching heart attack. By using such machine learning models this can rapidly improve the speed and accuracy of risk evaluation which gives healthcare professionals a significantly influential tool for early prediction and prevention.

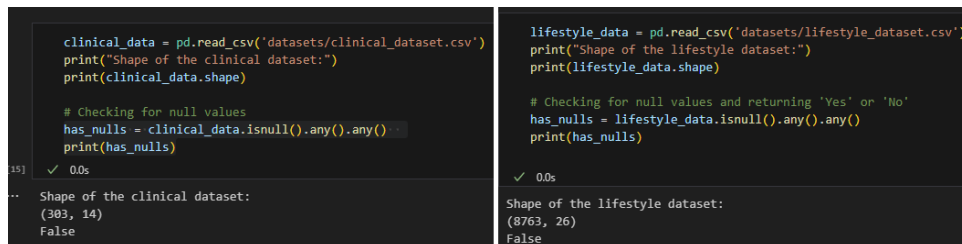
- **Developing a model** The main objective is to develop a machine learning model that provides a binary classification for heart attack risk prediction, (heart attack risk: Yes/No). The aim is to distinguish someone’s likelihood of heaving a heart attack using various parameters, allowing for early and effective medical intervention.

- **Data collection and preprocessing** A core objective is to collect clinical and lifestyle data and conduct pre-processing on these datasets. This requires an understanding of the dataset's structure, addressing missing values, dropping columns, and identifying categorical and numerical columns and transforming any necessary data into a machine learning format. Thorough data preparation was necessary to run our models.
- **Comparing the accuracy** Important once the models were run to compare the accuracies collected from the different machine learning algorithms used, which are: Logistic Regression, K-Nearest Neighbours (KNN), Random Forest, and Support Vector Machine (SVM) and Multilayer Perceptron (MLP). Doing this is crucial to select the most appropriate model; one with the highest accuracy and generalisability.
- **Validating the model** Model evaluation using cross-validation and performance metrics, such as accuracy, precision, and recall. We intend to identify the model producing the highest accuracy and reliability in heart attack risk prediction.
- **Discussing the ethical implications** This is very important as it identifies potential societal impacts of automated health predictions. Our developed model will have direct impacts on people, so ensuring our model is very reliable is a necessity, as well as clearly identifying and outlining their limitations.

## 3 Methodology

### 3.1 Data Exploration

**Dimension analysis** Both datasets had their dimensions checked and scanned for any null values.



```

clinical_data = pd.read_csv('datasets/clinical_dataset.csv')
print("Shape of the clinical dataset:")
print(clinical_data.shape)

# Checking for null values
has_nulls = clinical_data.isnull().any().any()
print(has_nulls)

lifestyle_data = pd.read_csv('datasets/lifestyle_dataset.csv')
print("Shape of the lifestyle dataset:")
print(lifestyle_data.shape)

# Checking for null values and returning 'Yes' or 'No'
has_nulls = lifestyle_data.isnull().any().any()
print(has_nulls)

```

The left screenshot shows the execution of code for the clinical dataset, resulting in a shape of (303, 14) and no null values. The right screenshot shows the execution of code for the lifestyle dataset, resulting in a shape of (8763, 26) and no null values.

Figure 1: Exploratory Data Analysis (Nguyen, 2024c)

#### Imbalance check:

There was a minor imbalance of classes identified with both datasets, however experimentation with undersampling of the lifestyle dataset yielded very minimal differences in classifier performance.

#### PCA analysis

- Scree plots indicated  $n = 10$  was optimal for the lifestyle dataset
- PCA dimension reduction to 10 principle components was hence considered for models running on the lifestyle dataset due to the large dimensions and long running times.

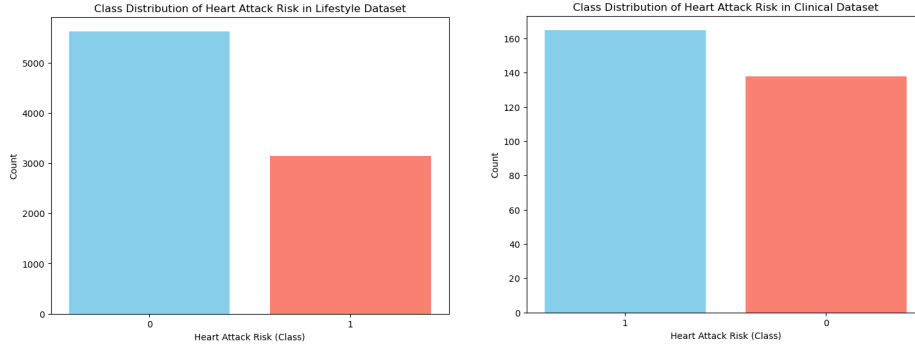


Figure 2: Imbalance check results (Nguyen, 2024d)

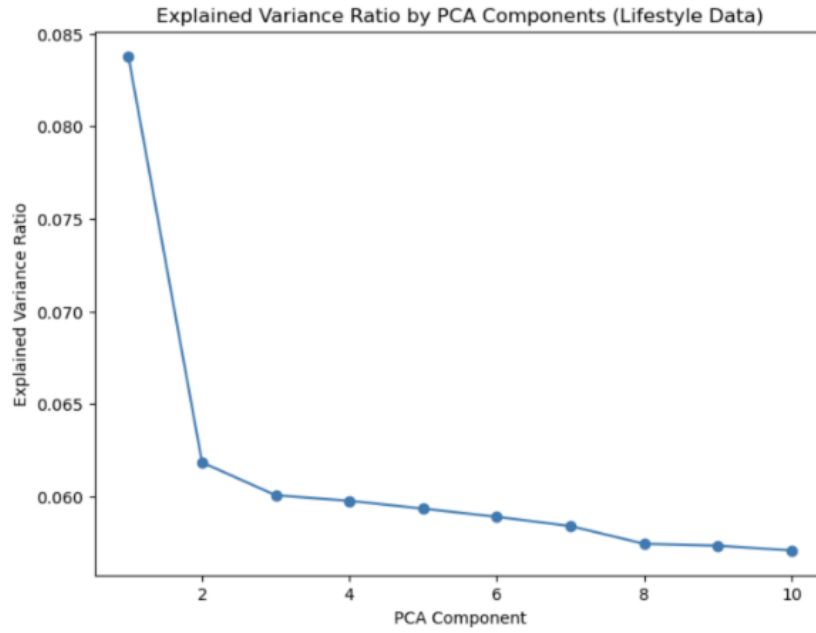


Figure 3: Scree plot for the lifestyle dataset with  $n = 10$  principal components (Young, 2024).

- None of the **clinicalmodels** exhibited overfitting or lengthy execution, thus PCA was not utilised.

### 3.2 Data Pre-Processing

Two datasets were used: a clinical dataset (age, cholesterol, blood pressure, etc.) and a lifestyle dataset (exercise habits, diet, smoking status, etc.). Both datasets were preprocessed separately:

- Clinical Dataset: No preprocessing was required
- Lifestyle Dataset: Blood pressure was split into systolic and diastolic components and categorical variables were one-hot encoded. Finally, numeric features were scaled using StandardScaler to standardise their range.

### 3.3 Model Training

Given the two datasets—Clinical (303, 14) with mostly numerical data, and Lifestyle (8763, 26) with mixed data types and high dimensionality—each model was chosen to align with dataset characteristics and optimise heart attack risk prediction.

1. **Random Forest (RF):**

- **Why:** Handles mixed data types and complex feature interactions; ideal for high-dimensional Lifestyle dataset.
- **How:** Configured with 200 trees; max\_features and tree depth tuned via cross-validation. Feature importance analysed for insight into key predictors.
- **Expected Performance:** Strong accuracy and stability on Lifestyle dataset; risk of overfitting on smaller Clinical dataset.

2. **K-Nearest Neighbors (KNN):**

- **Why:** Effective for dense data; suitable for Clinical dataset where risk similarity matters.
- **How:** Used optimal n\_neighbors and MinMax scaling; tested with PCA on Lifestyle dataset to mitigate high-dimensionality issues.
- **Expected Performance:** Performs well on Clinical dataset, leveraging similarity; limited accuracy on Lifestyle due to dimensionality.

3. **Weighted Nearest Neighbors (WNN):**

- **Why:** Enhances KNN by weighting closer neighbours, improving accuracy for Clinical dataset.
- **How:** Applied weights='distance', optimised n\_neighbors, validated improvement over KNN.
- **Expected Performance:** Outperforms KNN on Clinical dataset; marginal gains on Lifestyle due to high-dimensionality.

4. **Multi-Layer Perceptron (MLP):**

- **Why:** Models non-linear relationships; ideal for complex patterns in Lifestyle dataset.
- **How:** Two-layer network, relu activation, adam optimiser, early stopping. Regularisation applied to control overfitting on Clinical.
- **Expected Performance:** Excels on Lifestyle dataset; risks overfitting on Clinical due to limited samples.

5. **Support Vector Classifier (SVC):**

- **Why:** Effective for binary classification; maximises margin on Clinical dataset.
- **How:** RBF kernel, optimised C and gamma; PCA used on Lifestyle for dimensionality management.
- **Expected Performance:** Robust on Clinical dataset; limited by high-dimensionality in Lifestyle despite PCA.

### Evaluation Metrics

For evaluating heart attack risk predictions across our Clinical and Lifestyle datasets, we used 5-fold cross-validation along with targeted metrics to ensure robust model performance:

1. **Accuracy:** Used to provide an overall assessment of model performance.
2. **Precision:** This metric was crucial for minimising false positives, especially for the clinical dataset, where over-diagnosis could lead to unnecessary treatment.
3. **Recall (Sensitivity):** Prioritised for the clinical dataset to ensure that high-risk patients were not missed.

4. **F1 Score:** This balanced metric was particularly valuable for assessing the lifestyle dataset, which had a mix of categorical and numerical data and variable class distributions.
5. **ROC-AUC Score:** Chosen to evaluate how well the models could distinguish between individuals with and without heart attack risk.

## 4 Simulation Results

### 4.1 Key Findings and Significance

#### Clinical dataset results

Table 1: Model Performance Metrics

Model	Mean Accuracy	Mean Precision	Mean Recall	Mean F1 Score
SVC	0.87	0.87	0.87	0.87
WNN	0.89	0.89	0.88	0.88
KNeighbors	0.89	0.89	0.88	0.88
MLP	0.85	0.85	0.85	0.85
Random Forest	0.85	0.85	0.85	0.85
Logistic Regression	0.89	0.89	0.88	0.88

KNN, WNN, and Logistic Regression achieved the highest ROC-AUC scores (0.90 - 0.91) and accuracy (0.89), indicating a reliable ability to distinguish between high and low heart attack risk. These high scores suggest that these simpler models effectively captured the straightforward relationships present in the clinical dataset. In contrast, MLP and SVC performed slightly lower, reflecting limitations when applied to smaller datasets with fewer complex interactions. Overall, the high AUC values for the best-performing models demonstrate that the clinical dataset contains strong, distinct signals related to heart attack risk.

#### Clinical ROC curve analysis

The data shows that the clinical dataset allows for high predictive accuracy in assessing heart attack risk. Using models such as K-Nearest Neighbors (KNN) and Weighted Nearest Neighbors (WNN), we achieved ROC-AUC scores of up to 0.91, while Logistic Regression and Random Forest reached scores of 0.90. These results highlight the reliable accuracy of the clinical dataset. Specific indicators—including fasting blood sugar, cholesterol levels, maximum heart rate achieved, and electrocardiographic results (ECG)—were particularly effective predictors for heart attack risk. These features provide critical insights into cardiovascular health, making them essential components of effective predictive models.

#### Significance of the clinical simulations

The clinical dataset results demonstrate the feasibility of using simpler models to predict heart attack risk based on straightforward clinical data, whereas the lifestyle dataset highlights



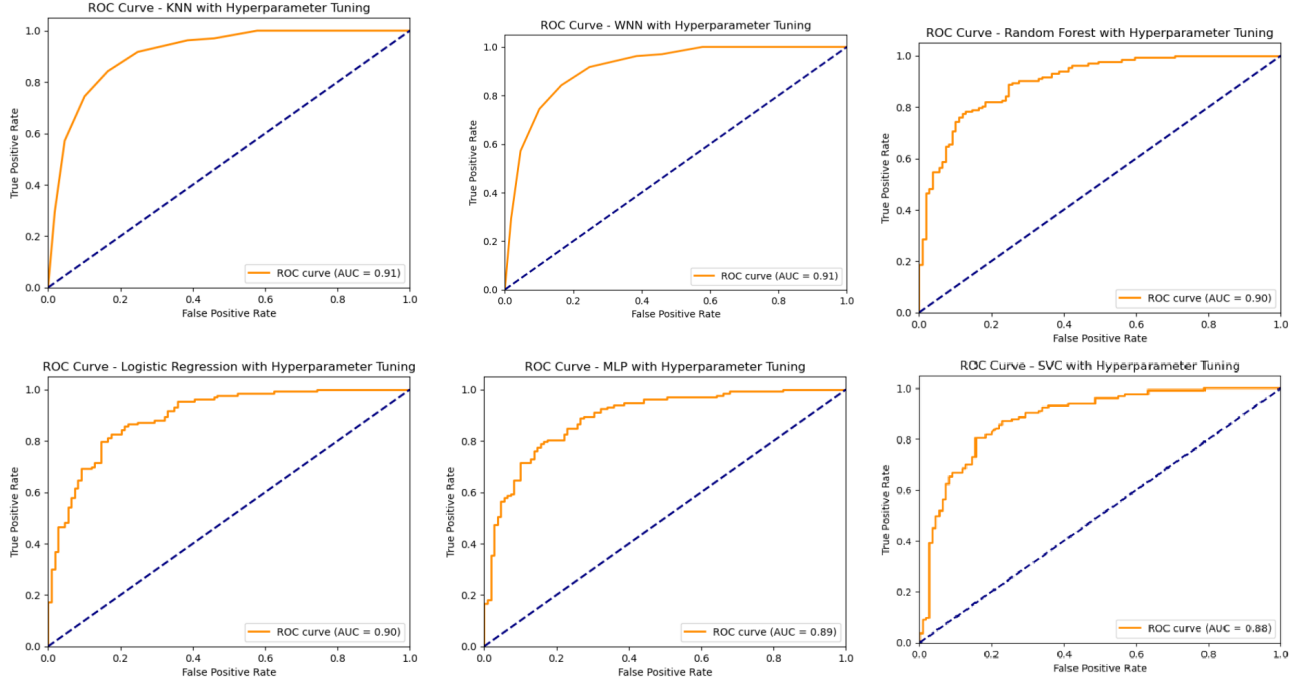


Figure 4: Clinical ROC curves  
(Nguyen, 2024b).

the challenges of dealing with complex, high-dimensional data that contains potential noise. The clinical model’s ability to accurately predict heart attack risk at an early stage has significant implications for healthcare. By predicting and identifying high-risk patients before symptoms arise, healthcare providers can implement preventative measures such as lifestyle modification, medication, or other interventions in advance. This approach reduces the likelihood of heart attacks and contributes to a healthier quality of life for at-risk individuals (Pearson, 2010). Moreover, early preventive strategies are often more cost-effective compared to emergency treatment, as they reduce the need for intensive care and long-term hospital stays (Arena and Myers, 2014).

## Lifestyle dataset results

### Validation metrics for lifestyle based models

Table 2: Model Performance Metrics

Model	Mean Accuracy	Mean Precision	Mean Recall	Mean F1 Score
SVC	0.641789	0	0	0
WNN	0.582904	0.380088	0.260592	0.308901
KNeighbors	0.583247	0.380909	0.261229	0.309639
MLP	0.582219	0.350656	0.196198	0.227193
Random Forest	0.620679	0.301412	0.043007	0.075157
Logistic Regression	0.641789	0	0	0

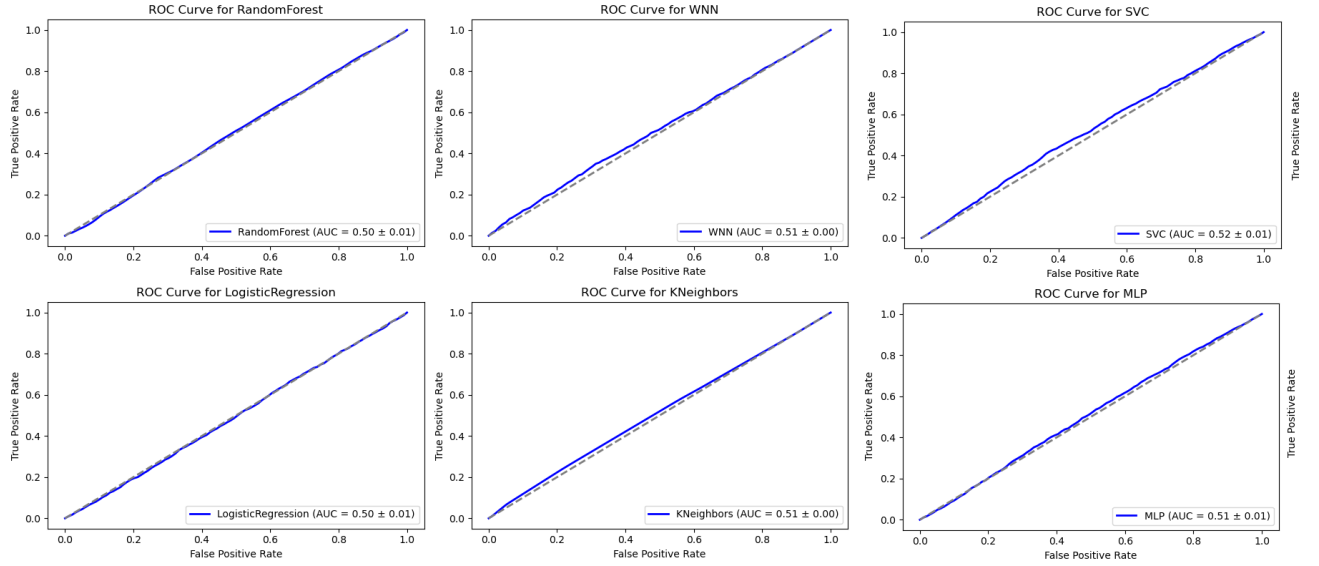


Figure 5: Lifestyle ROC curves  
(Nguyen, 2024a).

### Lifestyle metrics analysis

The evaluation metrics indicate that the overall predictive performance of the models on the lifestyle dataset was limited, with most models achieving modest accuracy scores ranging from 0.53 to 0.64. The Support Vector Classifier (SVC) and Logistic Regression produced mean accuracy scores of 0.6417, but both had zero precision, recall, and F1 scores, indicating a complete failure to effectively classify positive cases. Models like WNN, KNeighbors, and MLP demonstrated slightly better precision values (around 0.33 - 0.38), but their low recall (around 0.19 - 0.26) reveals difficulty in capturing true positive instances of heart attack risk. The Random Forest model also struggled, with an accuracy of 0.6206 and very low recall of 0.043, emphasising its inability to identify high-risk patients.

### Lifestyle ROC Curve analysis

The ROC-AUC scores for all models ranged between 0.50 to 0.52, indicating performance close to random guessing. These low AUC scores highlight that the models are unable to effectively distinguish between patients with high or low risk of heart attack using the lifestyle data alone. This suggests that the dataset's high feature variability and noise may obscure meaningful patterns, even after techniques like PCA were applied to reduce dimensionality. The poor AUC and F1 scores (mostly below 0.31) demonstrate the challenge of balancing precision and recall using indirect features like physical activity, diet, and sleep in isolation.

### Significance of the lifestyle simulations

The insights from our analysis underline several important implications for building predictive models for heart attack risk. The low ROC-AUC, precision, and recall scores—especially for models like SVC and Logistic Regression—strongly suggest that lifestyle factors alone are insufficient to accurately predict heart attack risk. These findings emphasise the necessity of incorporating more direct clinical indicators, such as cholesterol levels, blood pressure, and other biomarkers, to complement lifestyle data and enhance predictive power. Notably, these results align with findings from other studies on Kaggle, where similar models trained on lifestyle data alone achieved f1 scores around 0.6, further supporting the notion that lifestyle factors by them-

selves do not provide enough distinctiveness for robust predictions.

Out[64]:

	LogisticRegression	DecisionTreeClassifier	RandomForestClassifier	GaussianNB	KNeighborsClassifier
Accuracy	72.130000	60.360000	70.710000	70.890000	63.560000
F-1 Score	60.640000	61.080000	63.610000	61.450000	68.270000
Precision Score	100.000000	59.220000	82.170000	88.620000	59.840000
Recall Score	43.510000	63.060000	51.890000	47.030000	79.460000

(Riaz, 2024)

	Models	Accuracy	Precision	Recall	F1score
0	Logistic Regression	0.841757	0.500000	0.011146	0.021807
2	RF	0.627496	0.365591	0.054140	0.094313
1	DT	0.582484	0.394856	0.415805	0.404985

(Abodegham, 2024)

Nonetheless, the model presents significant potential for highly personalised applications when used alongside clinical data. By integrating diverse factors—ranging from diet, cholesterol, blood pressure to physical activity—it becomes possible to deliver more tailored health recommendations. This integration can lead to a more comprehensive risk profile, improving preventive care and ensuring that interventions are closely aligned with each patient’s unique health circumstances. Moreover, the insights from these predictive models can contribute to public health initiatives by informing targeted campaigns and preventive screening efforts for high-risk populations, thereby potentially reducing the overall incidence of heart attacks at a larger scale (Ingram and Vane, 2022).

## 4.2 Issues Faced

- **Long Training Times for Lifestyle Dataset:** Advanced models on the lifestyle dataset required extensive runtime due to high dimensionality. This was mitigated with PCA (n=10), which reduced training time while retaining essential information.
- **Challenges with Feature Selection:** Deciding whether to include categorical variables like “Hemisphere” in the lifestyle dataset was difficult, as some columns added complexity without apparent predictive value.
- **Poor Model Performance on Lifestyle Dataset:** ROC-AUC scores near 0.50 and low accuracy indicate that the lifestyle dataset features may lack the distinctiveness needed for effective prediction, complicating model optimisation.

## 5 Potential for Wider Adoption and Practical Applications

The high performance of models on the clinical dataset demonstrates significant potential for real-world healthcare applications, particularly where clinical data with strong signals can yield effective predictive models for heart attack risk. A promising application involves the development of digital health tools, such as web or mobile apps and wearables, that integrate clinical

data (e.g., ECG readings obtained from healthcare providers) with user-updated lifestyle data (e.g., exercise habits, stress levels, sleep quality, diet, and location). This combination allows for continuous health tracking and real-time alerts for users and healthcare providers when risk levels are detected. Integrating both clinical and lifestyle data enables these tools to provide adaptive, personalised insights into heart health, supporting timely interventions and reducing health risks, especially outside clinical settings (Lavie et al., 2013).

Embedding this technology within Electronic Health Record (EHR) systems can allow seamless integration of clinical data from healthcare providers and lifestyle data from users, facilitating real-time risk assessment during patient visits and improving patient outcomes through comprehensive data insights. Similarly, integrating these models into telemedicine platforms can enhance remote consultations by providing risk assessments that include both clinical and lifestyle information. This combination can significantly enhance patient evaluations and improve the quality of virtual care by giving healthcare providers access to up-to-date, holistic patient profiles

Moreover, this technology holds potential for broader commercialisation. It could be developed into a standalone software platform or mobile application designed for accessibility across various sectors, including healthcare, government, and private industry. Such an application would enable individuals to monitor their health continuously, make informed lifestyle choices, and receive timely alerts for preventive interventions, ultimately contributing to better health management and reduced emergency healthcare costs.

## **Impact of development**

The potential of this development is demonstrated by its capability to predict heart attack risk through a comprehensive analysis of both a patient’s historical clinical data and real-time lifestyle information. Traditional methods often fall short due to high costs, inefficiency, and the risk of human error. This innovative approach overcomes these challenges by automating the process, ensuring consistency, accuracy, and enabling early-stage detection while significantly reducing processing times. This leads to more informed decision-making and better resource allocation within healthcare.

## **Efficiency and Accessibility**

Automating core aspects of data processing and analysis significantly reduces time and costs associated with traditional methods. This reduction in complexity allows organisations to allocate resources more effectively and helps make advanced technology more accessible, especially for smaller organisations that previously found it financially restrictive.

## **Ethical and Privacy Considerations**

An important consideration for broader implementation is the risk of biases within training data, which could skew results if not addressed properly. Ensuring diverse and representative datasets will be essential for producing fair and ethical outcomes. Additionally, maintaining data privacy and security is crucial within healthcare. Future iterations of this technology should prioritise patient data security by ensuring secure storage solutions, and compliance with privacy regulations to safeguard user data and build trust.

## 6 Conclusions

Our project highlights the value of machine learning in predicting heart attack risk, particularly with clinical datasets. Models like K-Nearest Neighbors (KNN), Weighted Nearest Neighbors (WNN), and Logistic Regression showed high predictive accuracy (ROC-AUC up to 0.91), demonstrating their reliability with structured clinical data. In contrast, the lifestyle dataset presented challenges, reflected in low ROC-AUC scores (0.50), indicating limited predictive power due to complexity and variability. Enhancing feature engineering and integrating lifestyle data with clinical parameters could improve model performance. Early detection through these models supports preventive care, improves patient outcomes, and reduces emergency healthcare costs. Integrating them into digital health tools, such as apps and wearables, can enable continuous monitoring and timely intervention. Further research should validate these models on larger, diverse datasets and address ethical considerations like data bias and privacy to ensure fairness and accuracy. In conclusion, while clinical data models show strong potential for heart attack risk prediction, lifestyle data requires further refinement. Integrating these models into healthcare can support early intervention and improved patient outcomes.

## References

- Nada Abodegham. Heart attack. <https://www.kaggle.com/code/nadaabodegham/haert-attack/noteb>, 2024. Accessed: 2024-11-03.
- R. Arena and J. Myers. Cardiovascular preventive care: Enhancing quality and reducing costs. *Journal of Cardiovascular Medicine*, 11(2):115–123, February 2014.
- Centers for Disease Control and Prevention. Heart disease facts, April 2024. Retrieved from Heart Disease website: <https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html>.
- S. E. Ingram and J. R. Vane. Cardiovascular disease in the modern era: Challenges and future directions. *Cardiovascular Research*, 118(10):2253–2264, December 2022.
- C. J. Lavie, R. Arena, and J. Myers. Exercise and the cardiovascular system: Clinical science and cardiovascular outcomes. *Sports Medicine*, 43(2):117–130, February 2013.
- Elvis Nguyen. Roc curve analysis for lifestyle models, 2024a. Created on 22/10/2024, University of Sydney, on local VS Code IDE.
- Elvis Nguyen. Roc curve analysis for clinical models, 2024b. Created on 22/10/2024, University of Sydney, on local VS Code IDE.
- Elvis Nguyen. Dimension analysis, 2024c. Created on 22/10/2024, University of Sydney, on local VS Code IDE.
- Elvis Nguyen. Imbalance analysis, 2024d. Created on 22/10/2024, University of Sydney, on local VS Code IDE.
- T. A. Pearson. Early intervention in cardiovascular disease: Reducing risk through preventive care. *American Heart Journal*, 160(3):1–9, September 2010.
- Samra Riaz. Heart attack risk prediction. <https://www.kaggle.com/code/samraariaz/heart-attack-risk-prediction>, 2024. Accessed: 2024-11-03.

S. M. Varnosfaderani and M. Forouzanfar. The role of ai in hospitals and clinics: Transforming healthcare in the 21st century. *Bioengineering*, 11(4):337, 2024. doi: 10.3390/bioengineering11040337.

Georgina Young. Scree plot for lifestyle models, 2024. Created on 22/10/2024, University of Sydney, on local VS Code IDE.