Machine Learning Approaches for Cardiovascular Disease Prediction

Big Data Technologies Coursework Part 1

CS982: Big Data Technologies
University of Strathclyde Glasgow
Group 20
Word count: 2697

Contents

List of Figures

List of Tables

1: Introduction

Cardiovascular diseases stand as a leading contributor to death and illness. The World Health Organization highlights smoking, overconsumption of alcohol, unhealthy diets, and a lack of physical activity as primary contributors to these conditions. Such behaviours potentially result in elevated blood sugar and hypertension, both precursors to cardiovascular issues (WHO, 2021). Therefore, early intervention and preventative measures are crucial in mitigating these risk factors.

By accurately predicting CVDs, healthcare providers can initiate early interventions to stop diseases from progressing and allow for more targeted assessments. This report examines cardiovascular disease data to predict the presence or absence of CVD, considering behavioural and biological risk factors alongside demographic information to enhance accuracy.

Given the prevalence and impact of cardiovascular diseases, accurate prediction models can contribute significantly to early diagnosis and intervention. The dataset serves as a valuable resource for healthcare professionals and data scientists aiming to develop robust predictive models and understand the factors influencing cardiovascular health.

1.1 Key challenges/ objectives

The Cardiovascular Disease Dataset presents a rich source of information for predicting CVDs. Yet, several challenges and problems must be addressed to ensure the accuracy and reliability of the models developed.

One of the primary challenges in the dataset is the potential imbalance in the distribution of the target variable (cardio). When one class significantly outweighs the other, it can detrimentally affect the performance of machine learning models. Addressing this issue is crucial to prevent biased predictions and ensure the model's effectiveness in identifying the presence and absence of cardiovascular disease.

Furthermore, the process of feature selection is paramount. With a diverse set of variables, determining the most relevant features for predicting cardiovascular disease is essential. Feature selection is a critical step in building accurate models, as irrelevant or redundant features can introduce noise and reduce model performance (Bandyopadhyay & Saha, 2013). Identifying the key predictors among demographic, physiological, and lifestyle variables is a crucial problem to address.

Beyond accuracy, understanding the factors contributing to the model's predictions is equally essential. Achieving interpretability in machine learning models is challenging, particularly in complex models. Ensuring that the developed models provide insights into the relationships between input features and cardiovascular disease outcomes is a critical aspect of the analysis.

Lastly, the handling of outliers must be considered. Outliers can indicate variability in the data or errors. They may skew measures of central tendency and summary statistics and can impact the results of machine learning models. It is essential to address significant outliers while ensuring no valuable information is lost.

1.2 Introduction to the dataset

The Cardiovascular Disease Dataset was sourced from two reputable sources: the UCI Machine Learning Repository (Janosi et al., 1988) and Kaggle's Heart Disease Dataset. It contains anonymised patient data to adhere to ethical standards of privacy.

The dataset comprises 68,205 patient records, each with a unique identifier (ID). It includes demographic, lifestyle, and physiological metrics, along with derived variables for in-depth analysis. Refer to Table 1.1 for a detailed overview.

Table 1.1: Description and Dataset variables

Category	Variable	Description	Type
Identification	ID	Unique identifier	Numeric
Demographics	age	Age in days	Numeric
	age_years	Age in years	Derived, Numeric
	gender	Gender (1: Female, 2: Male)	Categorical
	height	Height in centimetres	Numeric
	weight	Weight in kilograms	Numeric
Physiological Measures	ap_hi	Systolic blood pressure	Numeric
	ap_lo	Diastolic blood pressure	Numeric
	cholesterol	Cholesterol levels (1: Normal, 2: Above Normal, 3: Well Above Normal)	Categorical
	gluc	Glucose levels (1: Normal, 2: Above Normal, 3: Well Above Normal)	Categorical

Lifestyle	smoke	Smoking status (0: Non-smoker, 1: Smoker)	Binary
	alco	Alcohol intake (0: Does not consume alcohol, 1: Consumes alcohol)	Binary
	active	Physical activity (0: Not physically active, 1: Physically active)	Binary
Target	cardio	Presence or absence of cardiovascular disease (0: Absence, 1: Presence)	Binary
Derived Measures	bmi	Body Mass Index	Derived, Numeric
	bp_category	Blood pressure category based on ap_hi and ap_lo	Derived, Categorical
	bp_category_encoded	Encoded form of bp_category for machine learning purposes	Derived, Numeric

2: Exploratory Data Analysis

An exploratory analysis of the Cardiovascular Disease Dataset to identify patterns and address problems within the dataset will lay the groundwork for more sophisticated analyses in subsequent chapters.

2.1 Descriptive Statistics

Descriptive analysis revealed a comprehensive dataset with no missing or duplicate values among its 68,205 entries. The mean age of 53 years, with a standard deviation of 6.77, spanning ages 29 to 64, suggests a dataset that predominantly represents an older demographic, a common characteristic in cardiovascular disease studies. An assessment of BMI indicates a tendency towards being overweight, with extreme values signalling potential data entry errors. Blood pressure measures stay within physiological norms, validating data integrity, though significant variability suggests diverse health statuses among the sample, See Figure 1.2. Histograms and boxplots for these variables (Figures 1.1 and 1.3) illustrate the distributions and outliers.

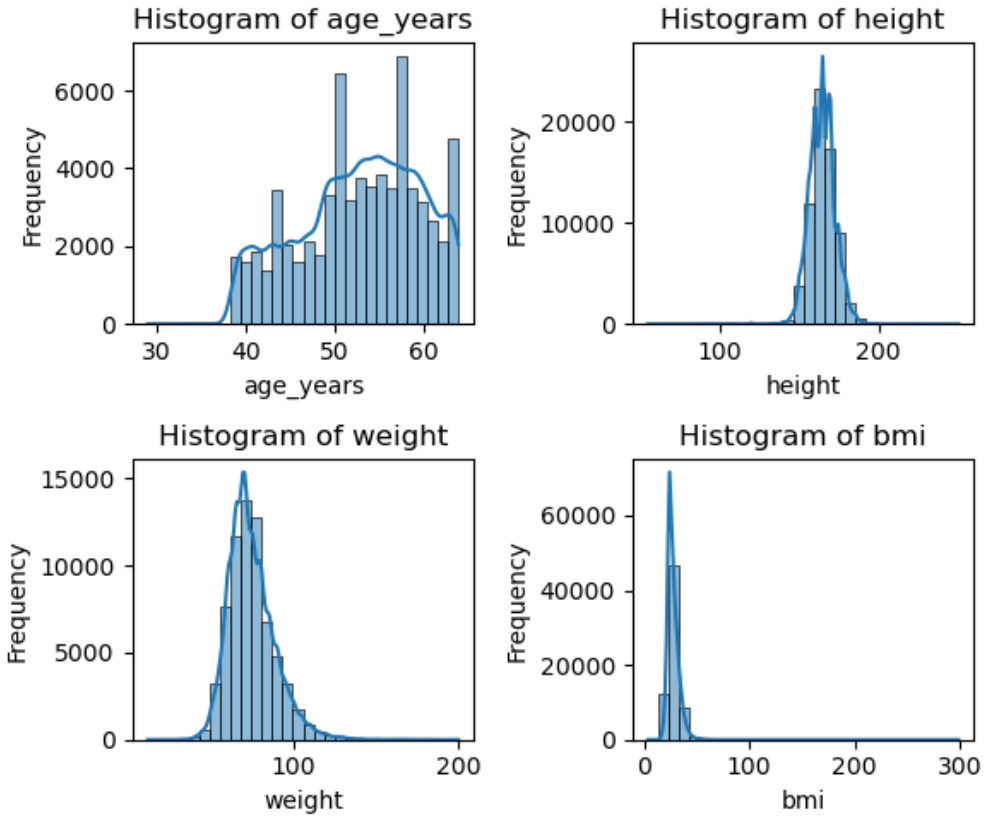

Figure 1.1: Histogram Distributions of Key Variables

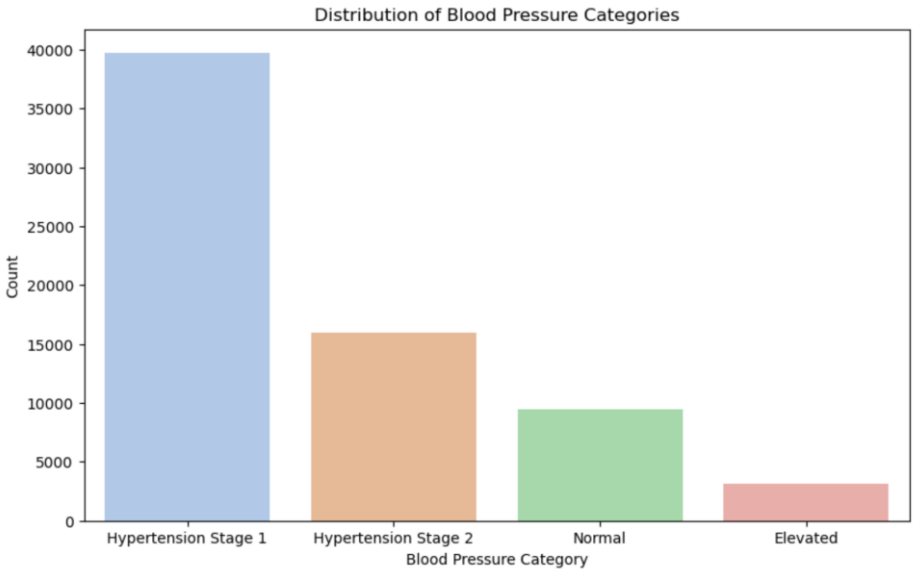

Figure 1.2: Distribution of Blood Pressure Categories

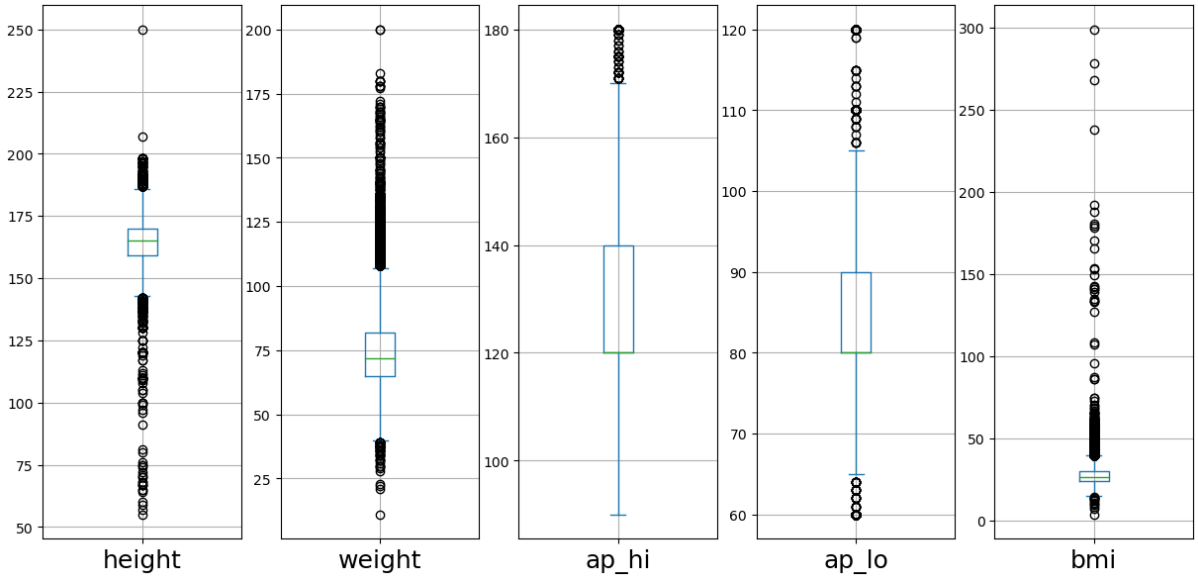

Figure 1.3: Boxplot Distributions of Numeric Variables

2.2 Reflections on Data Quality

The age distribution aligns with expectations, showcasing a dataset that predominantly represents an older demographic, a common characteristic in cardiovascular disease studies. Blood pressure measurements demonstrate a diverse range of cardiovascular health within the population. BMI, height, and weight distributions indicate a potential issue with outliers, warranting data cleaning and normalisation before applying machine learning models.

3: Data Preparation and Cleaning

3.1 Addressing Outliers

Outliers were identified and assessed using a criterion of 3 standard deviations from the mean across continuous variables. Values outside established realistic ranges for BMI, height, weight, and blood pressure were considered errors and removed. This exclusion, amounting to 0.8% of the data, resulted in a refined dataset with 67,634 entries, maintaining data integrity for subsequent analysis.

3.2 Encoding Categorical Data

We categorised BMI and blood pressure according to standard classifications and encoded them for analysis compatibility (Geron, 2022). The redundant 'age in days' feature was removed, enhancing dataset conciseness. Post-cleaning, the dataset features were standardised for machine learning algorithms, culminating in a streamlined dataset of 67,634 entries and an updated variable count reflecting these adjustments. This cleaning process sets the stage for robust and reliable analyses in subsequent chapters.

4: Identifying Correlations

4.1 Visual Analysis of Variables

After cleaning the data and preparing it for further analysis, we plotted some variables to understand the relationships between the variables in the dataset and the target variable. A scatterplot analysis identified a concentration of data in the BMI range of 20 to 40 and age of 45 to 60 years. The distribution of individuals with CVD becomes more dense at higher ages and BMIs, as depicted in Figure 1.4.

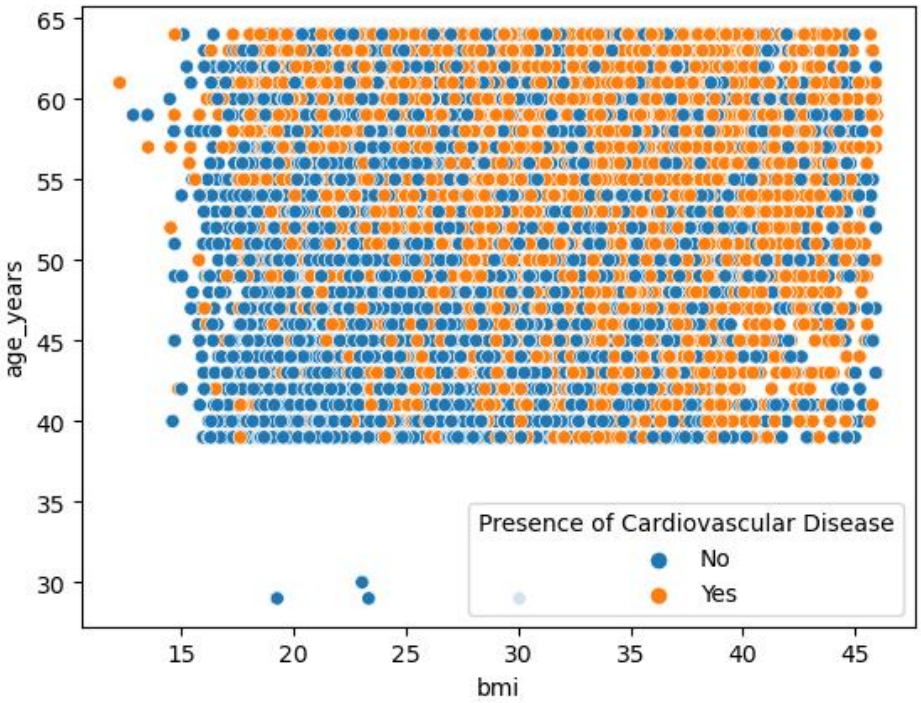

Figure 1.4: Scatterplot Showing the Relationship Between Age, BMI, and CVD

4.2 Correlation Insights

Digging into the data, it's clear there's a significant link between the upper (systolic) and lower (diastolic) numbers in blood pressure readings, with a strong connection at 0.73. There's also a notable, though less intense, relationship between blood sugar and cholesterol levels, showing a moderate correlation of 0.45. Moreover, the incidence of cardiovascular disease (cardio) is moderately correlated with heightened systolic blood pressure, reflected by a correlation coefficient of 0.43.

The heatmap in Figure 1.5 presents a comprehensive view of these relationships.

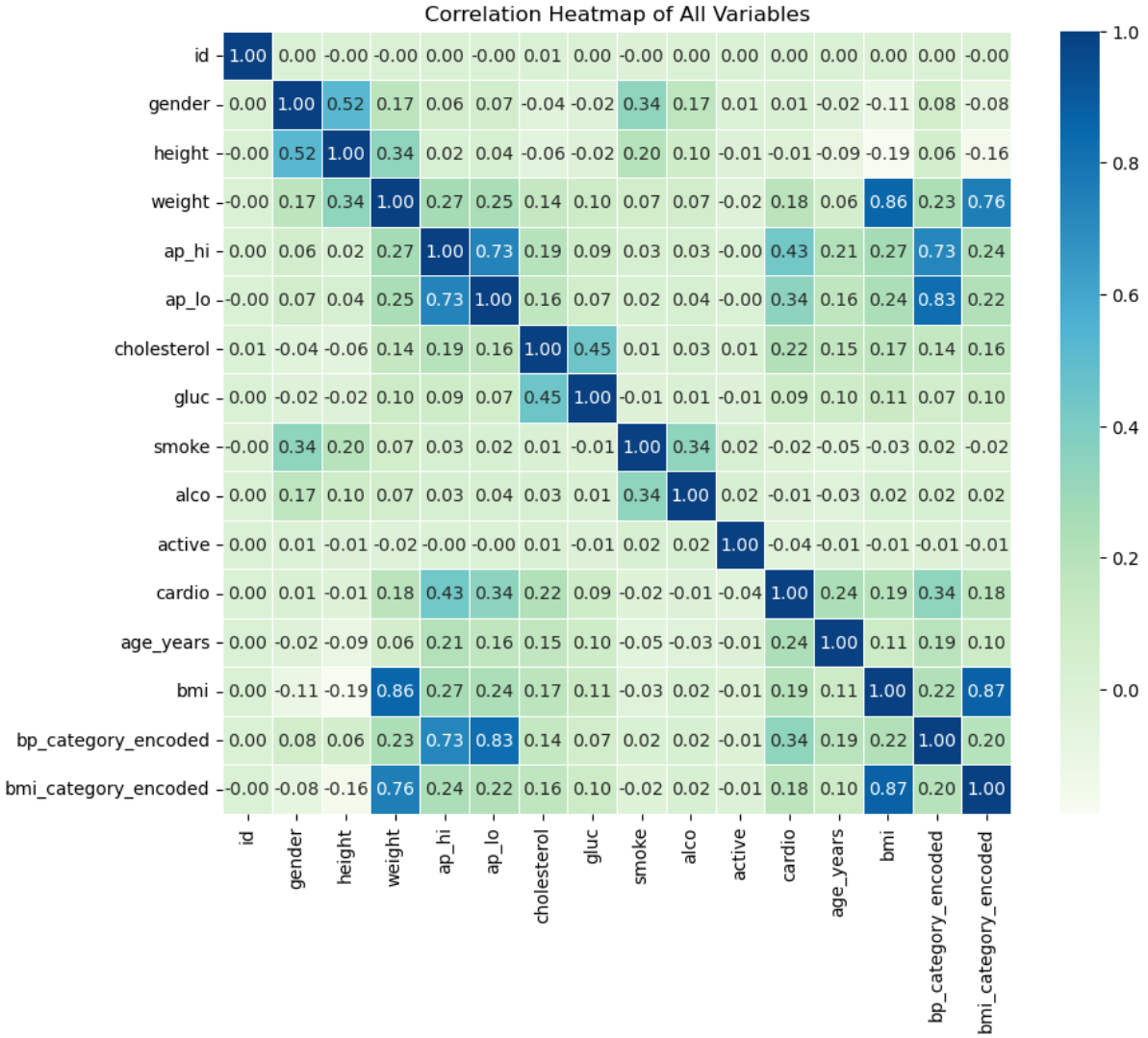

Figure 1.5: Correlation Heatmap of All Variables

5: Unsupervised Analysis

This chapter employs two prominent clustering techniques—K-Means and hierarchical clustering—to discern inherent groupings among individuals based on vital health metrics.

5.1 K-Means Clustering: Determining Optimal k

The Elbow Method determines the optimal number of clusters (k) for K-Means. The Elbow Method graph exhibits a distinct elbow point at k=3, indicating the optimal number of clusters. This choice is pivotal for subsequent clustering analyses. See Figure 1.6.

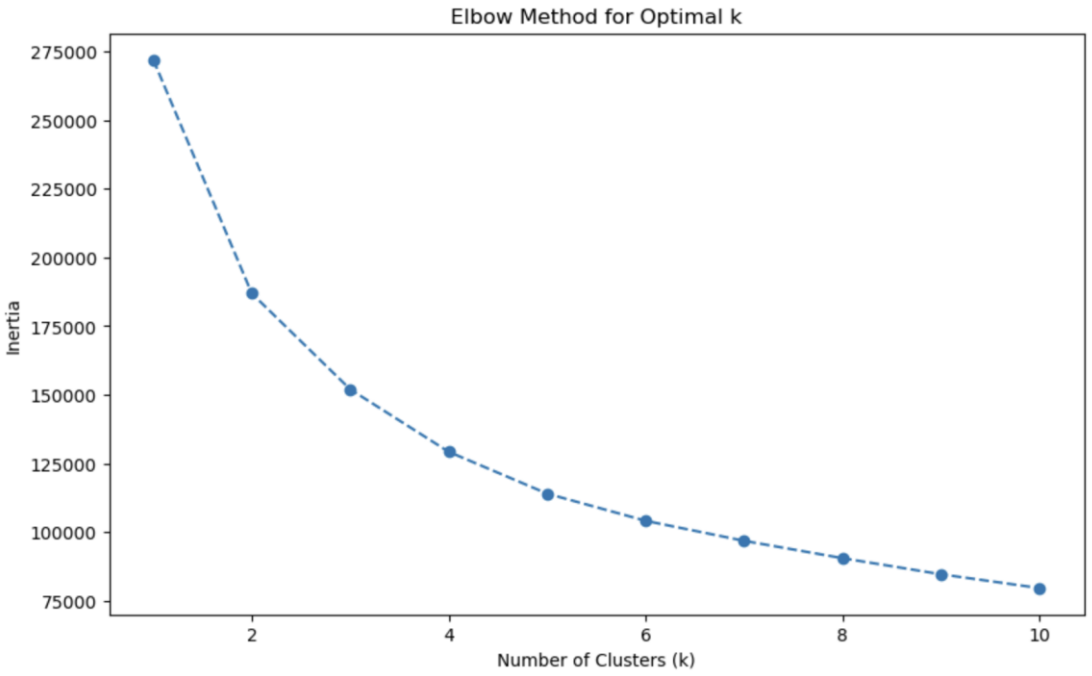

Figure 1.6: Elbow Method Graph

5.2 K-Means Clustering: Applying the Model

K-means clustering is applied to the dataset, segregating individuals into three clusters based on age, BMI, and blood pressure metrics after identifying k=3 as the optimal value. The choice of BMI, blood pressure, and age for K-Means clustering in cardiovascular disease prediction is based on their continuous nature, providing nuanced distinctions. These variables directly reflect cardiovascular health and clustering with these aligns with clinical practices. Conversely,

categorical variables like cholesterol and binary variables like smoking, alcohol, and physical activity were omitted due to their limited variability and potential for oversimplification in cluster analysis. While crucial for health assessment, these variables may be more suitable for other analytical approaches.

5.2.1 Cluster Profiles (K-Means Clustering)

Analysing the cluster profiles reveals distinct characteristics for each group. The mean age values, BMI, and blood pressure metrics show each cluster's composition. This information enables the characterisation of clusters, aiding in the interpretation of potential risk factors and health patterns.

The cluster profiles derived from the K-Means clustering method provide a detailed overview of the distinct characteristics exhibited by each identified group. These profiles, based on age, BMI, and blood pressure metrics, offer valuable insights into potential risk factors and health patterns within the Cardiovascular Disease Dataset. See Figure 1.7.

Cluster Profiles (K-Means):

cluster_kmeans	age_years	bmi	ap_hi	ap_lo
0	45.692722	25.479340	115.957278	75.745605
1	57.266433	26.772638	121.009071	78.404623
2	54.349412	30.059305	145.554429	91.349882

Figure 1.7: Cluster profiles

Cluster 0 (Low Risk): Younger age, moderate BMI, normal blood pressure.

Cluster 1 (Moderate Risk): Middle-aged, higher BMI and blood pressure than Cluster 0.

Cluster 2 (High Risk): Oldest age group, highest BMI, and hypertension levels.

5.2.2 Interpretation and Implications:

The cluster profiles affirm the stratification of individuals into low, moderate, and high-risk categories based on age, BMI, and blood pressure. The delineation of risk factors across clusters aligns with established cardiovascular health principles. These findings hold potential clinical implications for targeted interventions and risk assessments.

5.2.3 Limitations and Considerations:

While these cluster profiles offer valuable insights, there are some limitations of the K-Means clustering approach. The assumption of spherical clusters and equal variance across dimensions may only partially capture the complex relationships within the dataset. Additionally, the interpretation of clusters relies on the selected features and other relevant factors may contribute to cardiovascular health.

In the subsequent chapters, these cluster profiles will be further explored and integrated into developing predictive models for cardiovascular disease, contributing to a more comprehensive understanding of health outcomes.

5.3 Hierarchical Clustering: Constructing the Dendrogram

The output of the hierarchical clustering dendrogram provides a visual representation of the hierarchical relationships between data points and the formation of clusters. The dendrogram has vertical lines representing individual data points, horizontal lines indicating cluster mergers. The higher these join points are, the more different the clusters are. Truncating the dendrogram simplifies its interpretation, revealing the number of main clusters and the dissimilarity measures used to form them. By examining a dendrogram, we can visually deduce the most appropriate cluster count, and for our dataset, three distinct clusters are suggested.

The truncated hierarchical clustering dendrogram provides a simplified visualisation of data relationships, revealing branches that merge at varying heights, which reflect cluster dissimilarities. Truncated to show the most significant structures (p=3), it suggests three main clusters, with the choice of 'ward' linkage and 'Euclidean' distance informing the clustering process. The x-axis shows the indicators of the samples **i**n the dataset. By visually inspecting the dendrogram, we can make decisions about the optimal number of clusters based on the differences between clusters. In this case**,** three clusters may be appropriate. See Figure 1.8

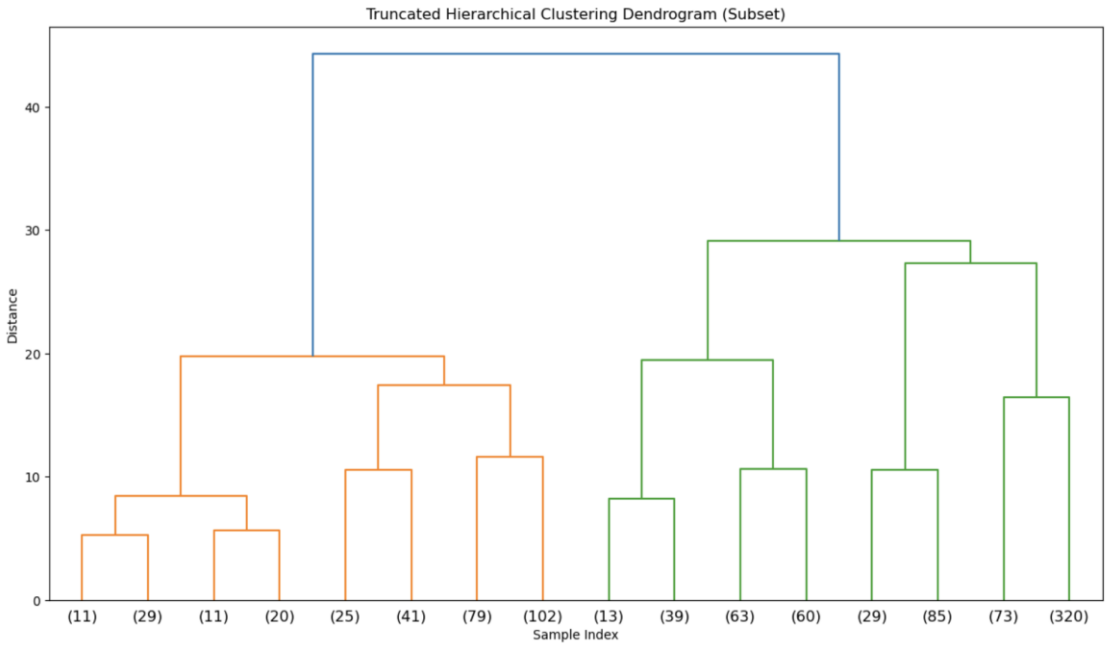

Figure 1.8: Hierarchical Clustering Dendrogram

5.4 Hierarchical Clustering: Applying the Model

In this section, we apply hierarchical clustering to the standardised dataset and determine an optimal number of clusters through visual inspection of the dendrogram. Examining the cluster profiles for hierarchical clustering provides an additional layer of insight into the inherent structures within the dataset.

These cluster profiles offer valuable insights into the characteristics of each group. Cluster 2 represents individuals with lower average age, BMI, and blood pressure metrics. In contrast, Cluster 1 and Cluster 0 exhibit higher values across these parameters, with Cluster 1 particularly standing out for elevated BMI and blood pressure. See figure 1.9.

Cluster Profiles (Hierarchical):

cluster_hierarchical	age_years	bmi	ap_hi	ap_lo
0.0	55.171216	26.888870	120.908189	80.158809
1.0	54.248428	28.591660	144.949686	91.034591
2.0	48.297491	26.076166	113.698925	71.885305

Figure 1.9: Hierarchical Cluster Profiles

5.5 Reflection on clustering analysis:

This chapter's clustering analysis, integrating K-Means and hierarchical methods, has revealed complex cardiovascular health patterns. Cluster profiles have deepened the understanding of the groups formed. K-Means has efficiently categorised the data into distinct groups by similarities in age, BMI, and blood pressure, while hierarchical clustering has added depth by illustrating the data's layered structure and varying risk factors.

Comparatively, K-Means is adept at forming clear clusters for large datasets, and hierarchical clustering offers detailed relationship insights. Both emphasise the significance of age, BMI, and blood pressure in cardiovascular risk assessment, enhancing the analysis narrative. Although hierarchical clustering provides detailed perspectives, K-Means remains crucial for its simplicity and efficiency, especially in determining the number of clusters, a pivotal element in the clustering process.

6: Supervised Analysis

This chapter focuses on developing predictive models for cardiovascular disease. Two fundamental supervised learning techniques will be employed to understand the relationship between key health metrics, identified clusters, and the presence or absence of cardiovascular disease.

1. Linear Regression
2. Logistic Regression

In Supervised Analysis for predictive modelling, a pivotal step involves implementing a 70/30 train-test split, allocating 70% of the dataset for model training and reserving the remaining 30% for evaluating the model's efficacy on new, unseen data. This division ensures a comprehensive assessment of the model's generalisation capabilities and accuracy in real-world scenarios.

6.1 Linear Regression

Linear Regression is a versatile method used for predicting a continuous outcome. In the context of cardiovascular disease prediction, we can utilise Linear Regression to model the relationship between numerical predictors (e.g., age, BMI, blood pressure) and a continuous target variable.

The Mean Squared Error (MSE) measures the average squared difference between the predicted and actual values. In this context, an MSE of 0.1957 indicates a relatively low level of prediction error. The R-squared value of 0.2172 suggests that the linear model accounts for 21.44% of the variability in the target variable. While this indicates a moderate fit, it emphasises the need to explore Logistic Regression for a more appropriate modelling of the binary outcome. See Figure 1.9

```
Linear Regression Evaluation:
Mean Squared Error: 0.19562009903050853
R-squared: 0.2172319134501366
```

Figure 2.0: Linear Regression Evaluation

6.2 Logistic Regression

Logistic Regression is a binary classification algorithm that predicts the probability of an event, such as the likelihood of cardiovascular disease from health metrics and identified clusters. The Logistic Regression model exhibits an accuracy of 0.7166, indicating that it correctly predicts cardiovascular disease presence or absence in approximately 71.66% of cases. The confusion matrix in Figure 2.0 reveals the distribution of true positive, true negative, false positive, and false negative predictions. Precision, recall, and F1-score provide a detailed breakdown of the model's performance for each class (0 and 1). The model's ability to balance precision and recall is crucial for accurate predictions, and further optimisation can be explored in subsequent chapters.

```
Logistic Regression Evaluation:
Accuracy: 0.7166231334089005
Confusion Matrix:
 [[8121 2219]
 [3531 6420]]
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.79      0.74     10340
           1       0.74      0.65      0.69      9951

    accuracy                           0.72     20291
   macro avg       0.72      0.72      0.71     20291
weighted avg       0.72      0.72      0.72     20291
```

Figure 2.1: Logistic Regression Evaluation

6.3 Comparative Analysis:

Comparing the results of both Linear Regression and Logistic Regression models, it's evident that Logistic Regression, specifically designed for binary classification tasks, outperforms Linear Regression in this context. The accuracy of 71.66% suggests that the model effectively predicts the presence or absence of cardiovascular disease based on the selected features and cluster information.

7: Reflection, Conclusion, and Future Exploration

7.1 Reflection on Analytical Approaches

Our analysis of the Cardiovascular Disease Dataset integrated both unsupervised and supervised techniques to uncover underlying patterns and predict outcomes. The unsupervised phase utilised K-means clustering to discern underlying patterns, while supervised analysis employed Linear and Logistic Regression for predictive modelling.

The analysis has highlighted some strengths and potential weaknesses of the approaches taken, particularly in how K-means clustering assumes spherical cluster distribution, which may only sometimes apply. Addressing the class imbalance through advanced balancing techniques, such as oversampling and under-sampling, is suggested for future research to mitigate class imbalance effects.

K-means clustering illuminated distinct data patterns, aiding in profiling different health segments. Logistic Regression demonstrated robust accuracy in predicting cardiovascular disease, proving its applicability to classification tasks. We also propose that future research investigate ensemble methods like Random Forest Gradient Boosting, which could capture intricate data relationships—coupled with innovative feature engineering thorough hyperparameter tuning to boost the predictive performance of our models.

7.3 Conclusion

In conclusion, our analysis has explored essential facets of cardiovascular health alongside the nuanced effectiveness of each analytical method. While machine learning models are successful in predicting CVDs. Future efforts should refine models, address imbalances, and incorporate advanced methodologies to elevate our understanding of CVD data.

Bibliography

Aurélien Géron (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 'O'Reilly Media, Inc.'

Bandyopadhyay, S. and Saha, S. (2013). *Unsupervised Classification Similarity Measures, Classical and Metaheuristic Approaches, and Applications*. Berlin, Heidelberg Springer.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, [online] 9(3), pp.90–95. doi:https://doi.org/10.1109/mcse.2007.55.

Janosi, A., Steinbrunn, W., Pfisterer, M. and Detrano, R. (1989). *UCI Machine Learning Repository*. [online] archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/dataset/45/heart+disease.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 445. doi:https://doi.org/10.25080/majora-92bf1922-00a.

Pedregosa, F., Fabian Pedregosa@inria Fr, Gael Varoquaux@normalesup Org, Alexandre Gramfort@inria Fr, Michel, V., Thirion, B., Bertrand Thirion@inria Fr and Grisel, O. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Alexandre Gramfort Vincent Dubourg Alexandre Passos Matthieu Perrot. *Journal of Machine Learning Research*, [online] 12, pp.2825–2830. Available at: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

WHO (2021). *Cardiovascular diseases (CVDs)*. [online] www.who.int. Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=The%20most%20important%20behavioural%20risk [Accessed 31 Oct. 2023].

www.kaggle.com. (n.d.). *Cardiovascular Disease*. [online] Available at: https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease/data [Accessed 3 Nov. 2023].

Appendix A

Development Environment and Software Details

Python Version:
Python 3.11.4

Integrated Development Environment (IDE):
Jupyter Notebook Version 6.5.4

Packages:
The following list of packages and versions were used:
matplotlib ~ 3.7.1
numpy ~ 1.24.3
pandas ~ 1.5.3
requests ~ 2.31.0
scikit-learn ~ 1.3.0
scipy ~ 1.10.1
seaborn ~ 0.12.2