Applying Machine Learning Methods to Analyse Cardiovascular Disease.

Machine Learning and Computational Intelligence (ELE00162M) 2023-24.

Introduction

This report investigates machine learning methodologies to analyse a dataset from a clinical study devoted to probing the early identification and handling of cardiovascular disease (CVD), a pivotal health concern worldwide. The primary aim centres on utilising machine learning approaches to predict the presence or absence of CVD based on a curated set of attributes within the dataset. Further to discussing data types, modalities, objectives and success measures, this report aims to provide a holistic understanding by studying the clinical context.

Taking steps to enhance the dataset's efficacy in modelling by exploring data cleaning and encoding strategies, the report navigates through an array of supervised learning methodologies, exploring their purposes and rationale, before delving into discussions surrounding unsupervised learning methods and drawing comparisons.

A conclusion compiles the results from the preceding sections, presenting implementation from both a data-centric and clinical standpoint, further incorporating discussions on applying relevant statistical measures, augmenting the rigour of the findings. The final segment of the conclusion interweaves references to ethicality and sustainability considerations stemming from the methods employed.

Introduction to the Clinical Context

CVD, also called heart and circulatory disease, is an umbrella name for conditions that affect your heart or circulation, including high blood pressure, stroke, and vascular dementia [1]. Heart disease includes conditions that narrow or block blood vessels (coronary heart disease), leading to a heart attack, angina, and some strokes, further covering conditions that affect your heart's muscles or valves or those causing abnormal rhythms (arrhythmias) [1]. Although the exact cause of CVD isn't clear, there are many 'risk factors' that can increase the risk of CVD presence. The risk factors that contribute to CVD include high blood pressure, smoking, high cholesterol, diabetes, inactivity, being overweight or obese, age, gender, and alcohol consumption [2].

The inability to pinpoint the exact cause of CVD in an individual and the presence of many quantifiable risk factors permits the employment of machine learning approaches, deriving the primary objective to develop a robust classifier to predict the presence or absence of the disease based on the provided attributes, measuring the overall system success with the detection accuracy, scoring how well the classifier predicts the presence or absence of CVD. Secondary objectives include:

(a) Identify crucial factors or features strongly associated with CVD.

(b) Explore correlations between lifestyle choices (smoking, alcohol consumption, physical activity) and CVD presence.

(c) Evaluate the impact of physiological factors (age, gender, blood pressure, cholesterol, glucose) on CVD.

The provided dataset contains various risk factor recordings for each patient in the clinical study, with a field containing whether they have CVD. These fields contain values of two modality categories: numerical data, which contains numerical data types with various modalities in terms of the range of values they can take within their respective types, and categorical data, which contains binary or other categorical types, where each value associates to a specific category. Although another modality of binary type exists, one can consider it a subset of the categorical modality type.

(a) Numerical data fields: Age, height, weight, systolic and diastolic blood pressure are numerical data types. Age, height, and systolic and diastolic blood pressure fields all contain integer values measured in days, centimetres, and millimetres of mercury (mmHg), respectively. Weight field containing float values measured in kilograms.

(b) Categorical data fields: Gender, cholesterol, glucose, smoking, alcohol consumption, and physical activity are categorical data types. Gender has two modalities, each representing the male and female genders. Smoking, alcohol consumption, and physical activity are binary fields, with a '1' indicating presence and a '0' indicating absence. Cholesterol and glucose fields contain three modalities, '1' for 'normal', '2' for 'above normal', and '3' for 'well above normal'.

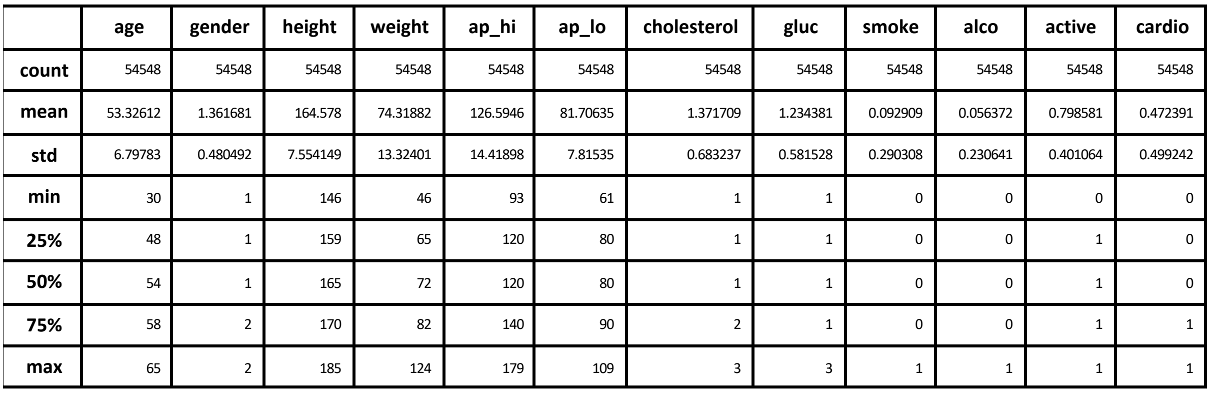
Data Cleaning

The integrity and quality of the dataset are pivotal for robust and insightful analysis. The presence of missing and corrupted values within this dataset has emerged as a critical concern, necessitating an initial focus on data cleansing before delving into subsequent analyses. Although there aren't any missing or null data in any field, there are a couple of fields with erroneous values.

	id	age	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
count	64511	64511	64511	64511	64511	64511	64511	64511	64511	64511	64511	64511
mean	53562.08	19420.4	164.3647	73.9773	128.2011	95.2408	1.353955	1.222768	0.088419	0.053898	0.805258	0.457116
std	27150.76	2473.001	8.215813	14.30391	153.8503	179.1039	0.669552	0.568043	0.283906	0.225818	0.396005	0.498161
min	0	10798	55	10	-150	-70	1	1	0	0	0	0
25%	30880.5	17605	159	65	120	80	1	1	0	0	1	0
50%	53891	19673	165	72	120	80	1	1	0	0	1	0
75%	76850.5	21290	170	82	140	90	1	1	0	0	1	1
max	99999	23713	207	200	16020	11000	3	3	1	1	1	1

Table 1 Pandas description of the original dataset.

Pandas data row indexing, which is built-in, makes the included 'id' redundant, hence safe for deletion to improve dataset simplicity and reduce overall size. Converting the 'age' data units from days to years and maintaining integer type rounding aids readability drastically. The 'gender' field, although specified by the attribute sheet to contain two values, '1' for female and '2' for male, contains many non-standardised values such as 'F', 'f', 'm', 'FEMALE', and 'MALE', hence its absence from the Pandas descriptive statistics in Table 1. Systolic pressure is the maximum blood pressure during contraction of the ventricles; diastolic pressure is the minimum pressure recorded just before the next contraction [3], eliminating the possibility of higher diastolic readings than systolic. Rows with higher diastolic pressure readings than systolic lead to the uncovering of further data entry mistakes. Since it is unclear whether it's just swapped data between two rows, entry deletion is the safest choice. The minimum recorded height is 55cm, and the lightest recorded weight is 10kg, both guaranteed erroneous entries given the 10798 to 23713 days (29 to 65 years) patient age range in the study. Further outliers exist in the blood pressure data, with the highest readings at 16020 and 11000 for ‘ap_hi’ and ‘ap_lo’, respectively.



	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
count	54548	54548	54548	54548	54548	54548	54548	54548	54548	54548	54548	54548
mean	53.32612	1.361681	164.578	74.31882	126.5946	81.70635	1.371709	1.234381	0.092909	0.056372	0.798581	0.472391
std	6.79783	0.480492	7.554149	13.32401	14.41898	7.81535	0.683237	0.581528	0.290308	0.230641	0.401064	0.499242
min	30	1	146	46	93	61	1	1	0	0	0	0
25%	48	1	159	65	120	80	1	1	0	0	1	0
50%	54	1	165	72	120	80	1	1	0	0	1	0
75%	58	2	170	82	140	90	2	1	0	0	1	1
max	65	2	185	124	179	109	3	3	1	1	1	1

Table 2 Pandas description of the cleaned dataset.

Defining an interquartile range of 99% enables outlier pruning below 0.5% and above 99.5% thresholds without losing valuable data points. The final stage of dataset cleaning consists of detecting and deleting duplicate values. At this point, 3551 duplicate entries existed in the dataset, all subsequently deleted. Shown in Table 2 is the Pandas data description of the cleansed set with a reduction of approximately 10,000 records.

Data Encoding, Balancing and Feature Extraction

Integrating a body mass index (BMI) column derived from the 'height' and 'weight' columns can enhance the analytical approach by consolidating these individual attributes into a singular metric. The formula for calculating the BMI metric using height (m) and weight (kg) follows:

$$BMI = \frac{weight}{height^2}$$

Similarly, a pulse pressure (PP) column, derived from the systolic and diastolic pressure readings, consolidates both pressure readings into a singular metric [4]. The formula for calculating the PP metric using ‘ap_hi’ (mmHg) and ‘ap_lo’ (mmHg) follows:

$$PP = ap_{hi} - ap_{lo}$$

BMI and PP columns encapsulate a more holistic representation of an individual's body composition, aligning with the intricate relationship between cardiovascular health and overall body mass. This fusion simplifies the data by reducing its dimensionalities, aiding the streamlining of analysis.

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	bmi	pp
count	54548	54548	54548	54548	54548	54548	54548	54548	54548	54548	54548	54548	54548	54548
mean	53.326116	1.361681	164.57795	74.318816	126.594559	81.70635	1.371709	1.234381	0.092909	0.056372	0.798581	0.472391	27.48313	44.888209
std	6.79783	0.480492	7.554149	13.324014	14.418978	7.81535	0.683237	0.581528	0.290308	0.230641	0.401064	0.499242	4.922222	10.520767
min	30	1	146	46	93	61	1	1	0	0	0	0	14.609204	9
25%	48	1	159	65	120	80	1	1	0	0	1	0	23.889463	40
50%	54	1	165	72	120	80	1	1	0	0	1	0	26.54321	40
75%	58	2	170	82	140	90	2	1	0	0	1	1	30.297784	50
max	65	2	185	124	179	109	3	3	1	1	1	1	56.29574	108

Table 3 Pandas description of the engineered dataset.

Table 3 shows the Pandas description of the dataset with the engineered BMI and PP columns. With this in place, feature exploration w.r.t. the risk factors outlined in [2] is possible.

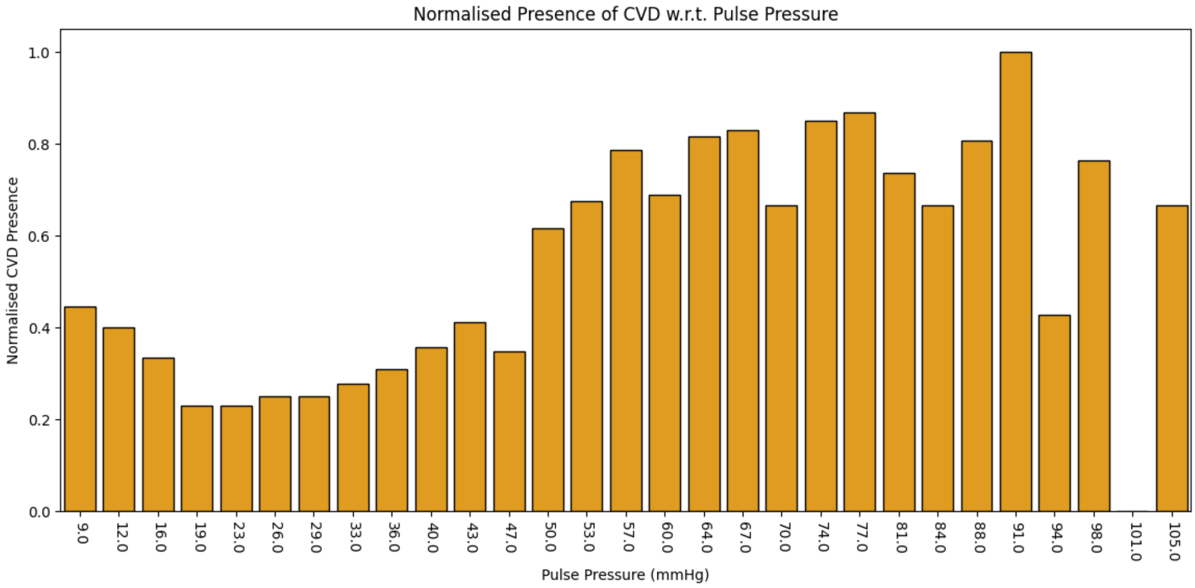


Figure 1 Pulse pressure w.r.t. CVD presence.

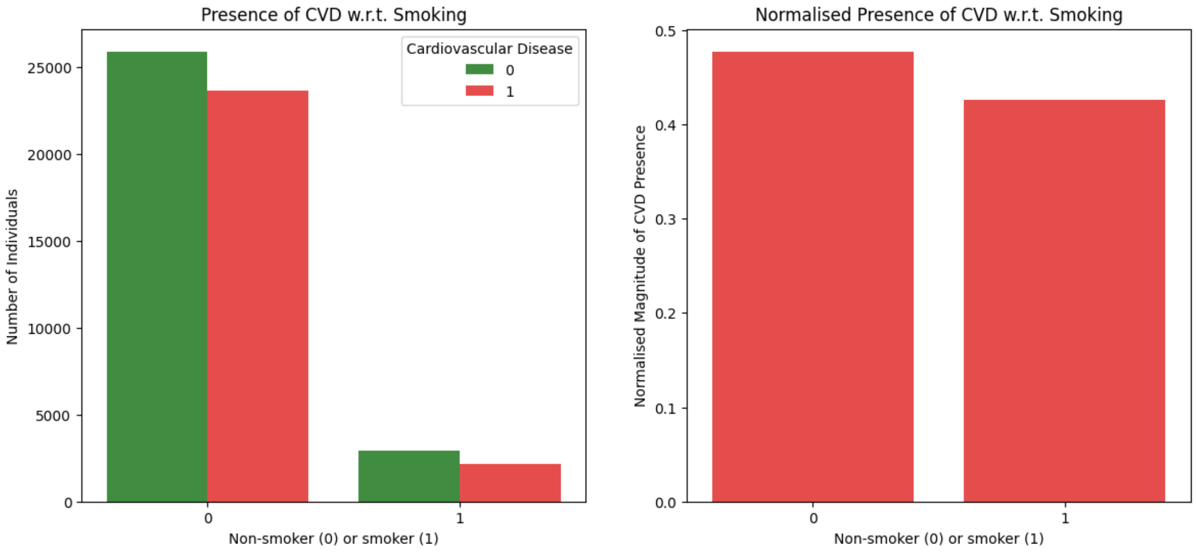


Figure 2 Smoking w.r.t. CVD presence.

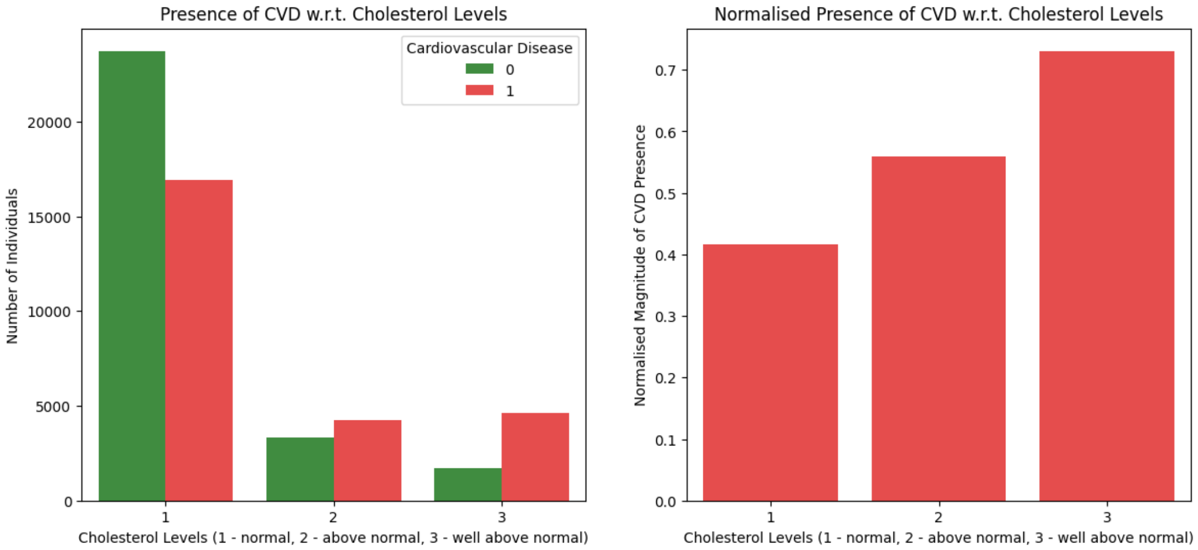


Figure 3 Cholesterol levels w.r.t. CVD presence.

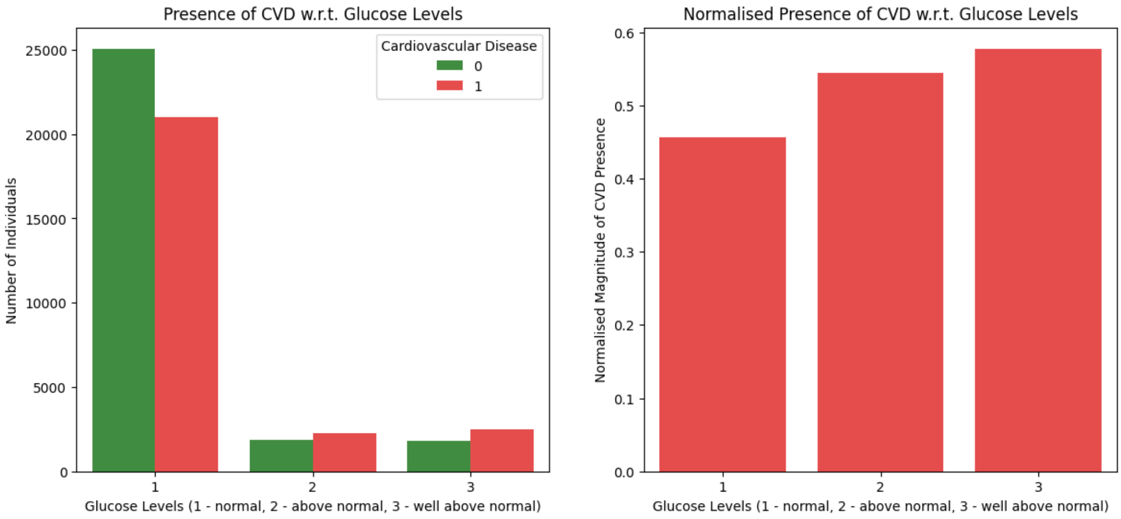


Figure 4 Glucose levels w.r.t. CVD presence.

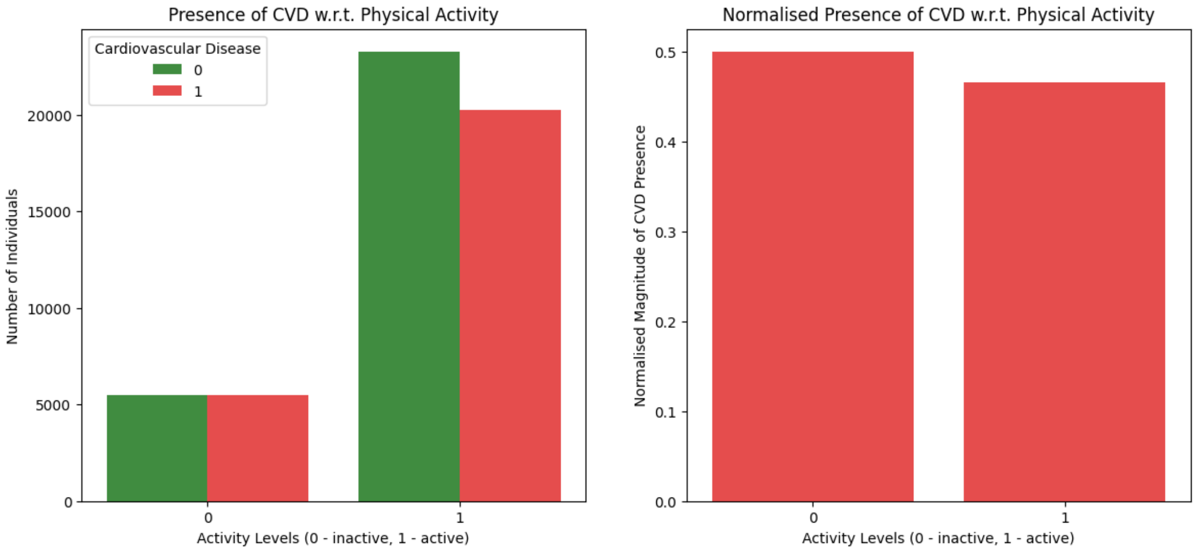


Figure 5 Physical activity w.r.t. CVD presence.

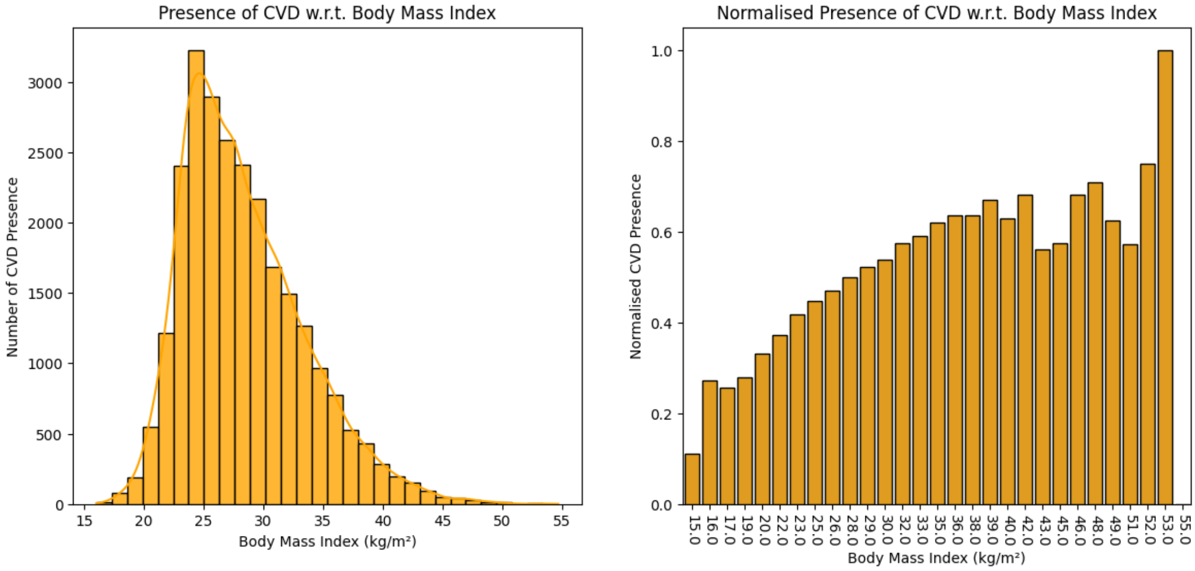


Figure 6 BMI w.r.t. CVD presence.

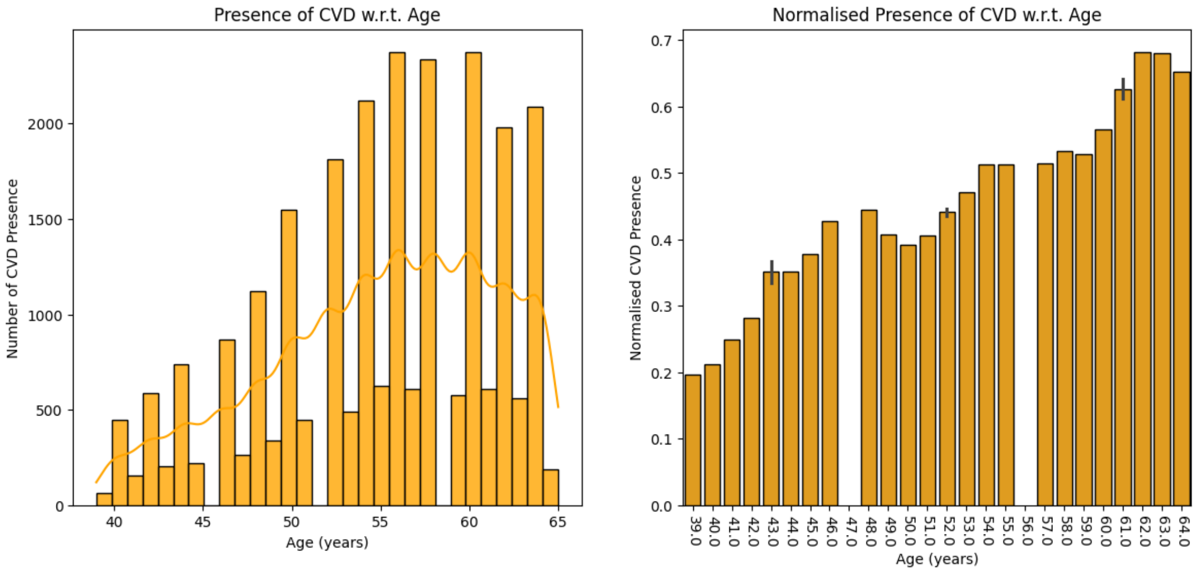


Figure 7 Age w.r.t. CVD presence.

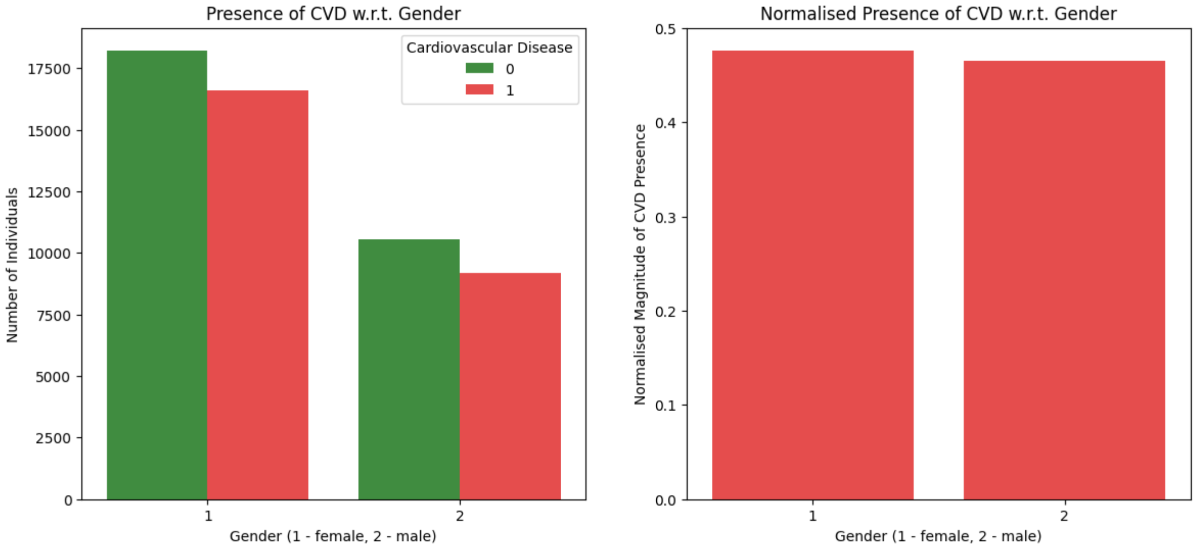


Figure 8 Gender w.r.t. CVD presence.

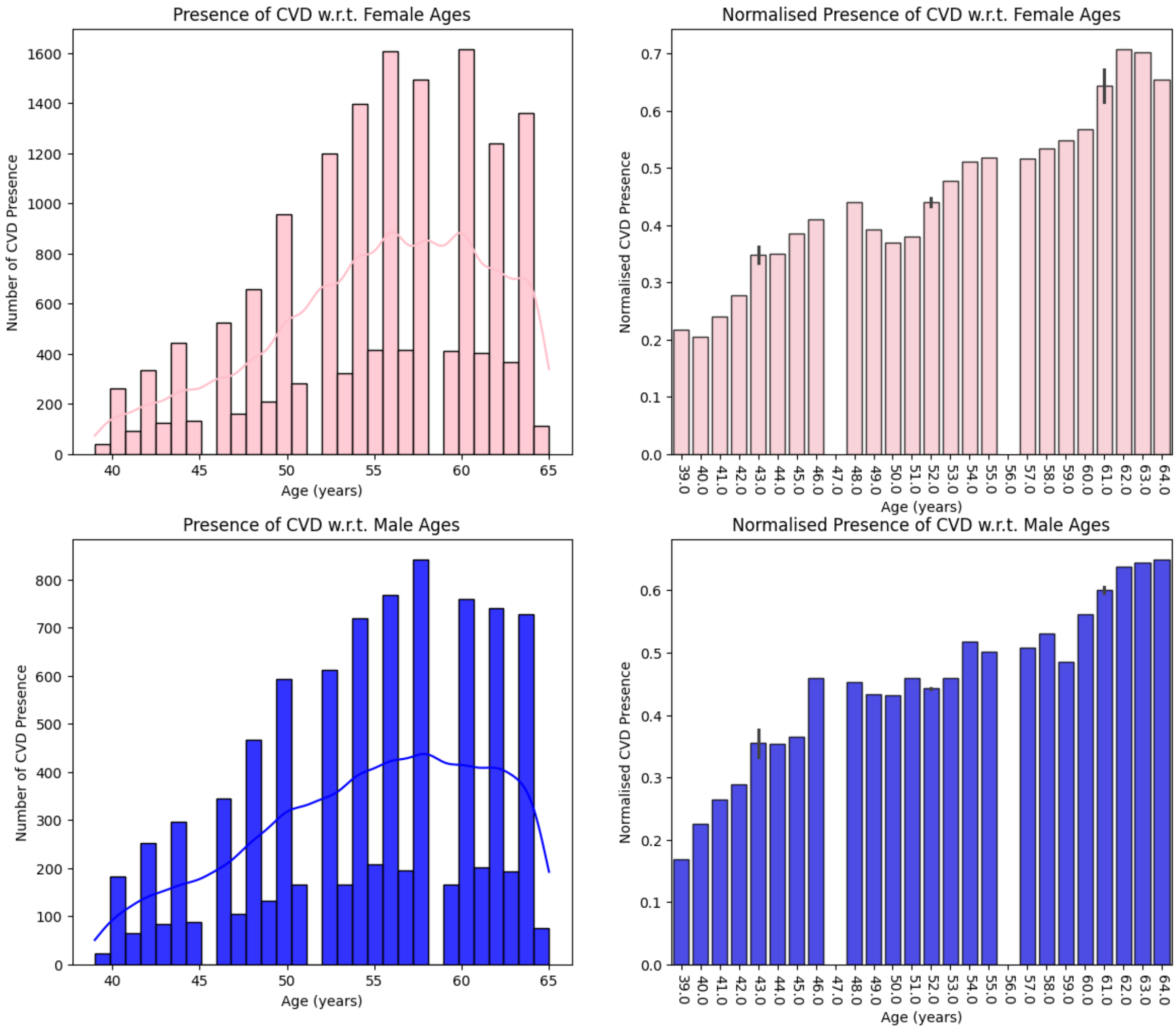


Figure 9 Age w.r.t. gender w.r.t. CVD presence.

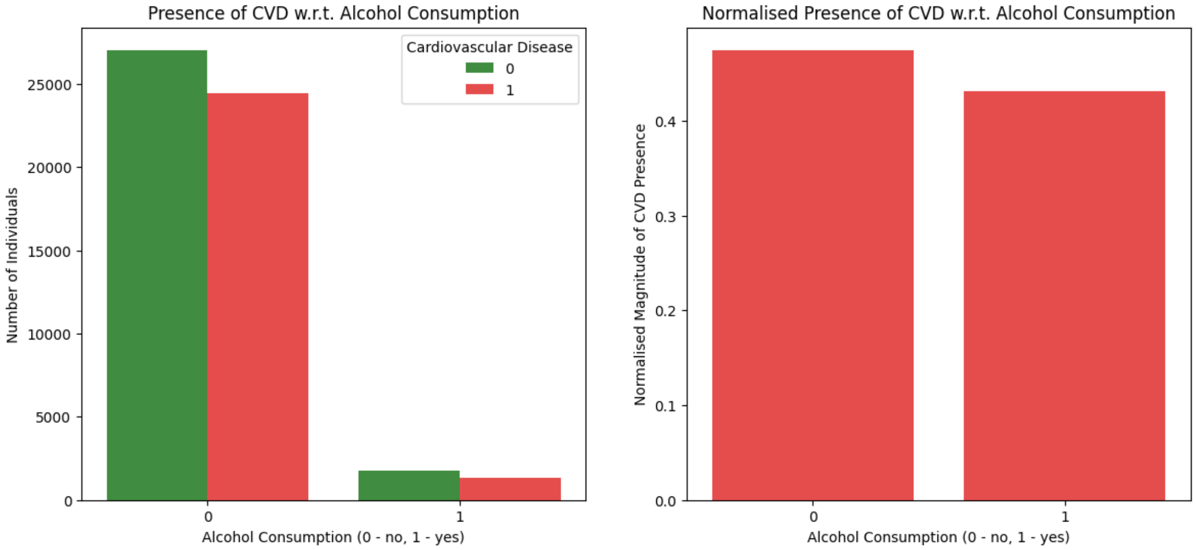


Figure 10 Alcohol consumption w.r.t. CVD presence.

Figure 1 shows the plot of CVD presence w.r.t. pulse pressure. It is clear to see a trend of higher CVD presence at the higher pulse pressure readings with an evident left-skew. Furthermore, the normalised histograms prove a positive correlation for CVD presence as pulse pressure readings grow, albeit with exceptions on either extremity. Figure 2 outlines the relationship between smoking status and the presence of CVD, uncovering an astronomical data imbalance between recorded smokers and non-smokers. The normalised box plot, rather wrongfully, shows a higher CVD presence in non-smokers, a direct repercussion of the data imbalance causing a higher number of collected samples for non-smokers compared to smokers. Figure 3 shows the relationship between cholesterol levels and CVD. Similar to the smoking data, there is a significant data imbalance with a higher sample of individuals with normal cholesterol levels. Despite the imbalance, the normalised box plot shows a precise trend of positive correlation, proving higher levels of cholesterol result in higher CVD presence. Figure 4 shows glucose levels w.r.t. CVD, and just like the cholesterol levels w.r.t. CVD, there is a monumental imbalance. However, showing a trend, albeit not as clear as the cholesterol data, in positive correlation. Like the cholesterol levels plot, the normalised box plot proves higher glucose levels result in higher CVD presence. The graph of physical activity w.r.t. CVD presence in Figure 5 shows a massive data imbalance similar to the indicated instances above. The normalised box plots unveil a trend, albeit faintly, in negative correlation, proving that physical activity impedes CVD presence. Plotting a histogram of CVD-present individuals w.r.t. BMI in Figure 6 shows a predominant left-skew, primarily caused by data imbalances in various histogram data bins. Normalising this data to plot another histogram shows an unmistakable positive correlation trend, proving that higher BMIs heighten the presence of CVD. Figure 7 demonstrates the rise in CVD with age, results as expected with a clear positive trend proving the fact older age results in more elevated CVD presence. Figure 8 shows another instance of immense data imbalance due to more female than male records. The normalised box plot shows a microscopic edge in CVD presence for females than males. Figure 9 shows the relationship between age and gender for CVD presence. Although the sample size of females is much higher, the overall result of CVD presence in males at younger ages is evident in the normalised plot. Figure 10 shows the relationship between alcohol consumption and CVD, there is, again, a colossal imbalance between the two categories, thus deriving an odd result within the normalised box plots, showing the reduced presence of CVD in alcohol drinkers.

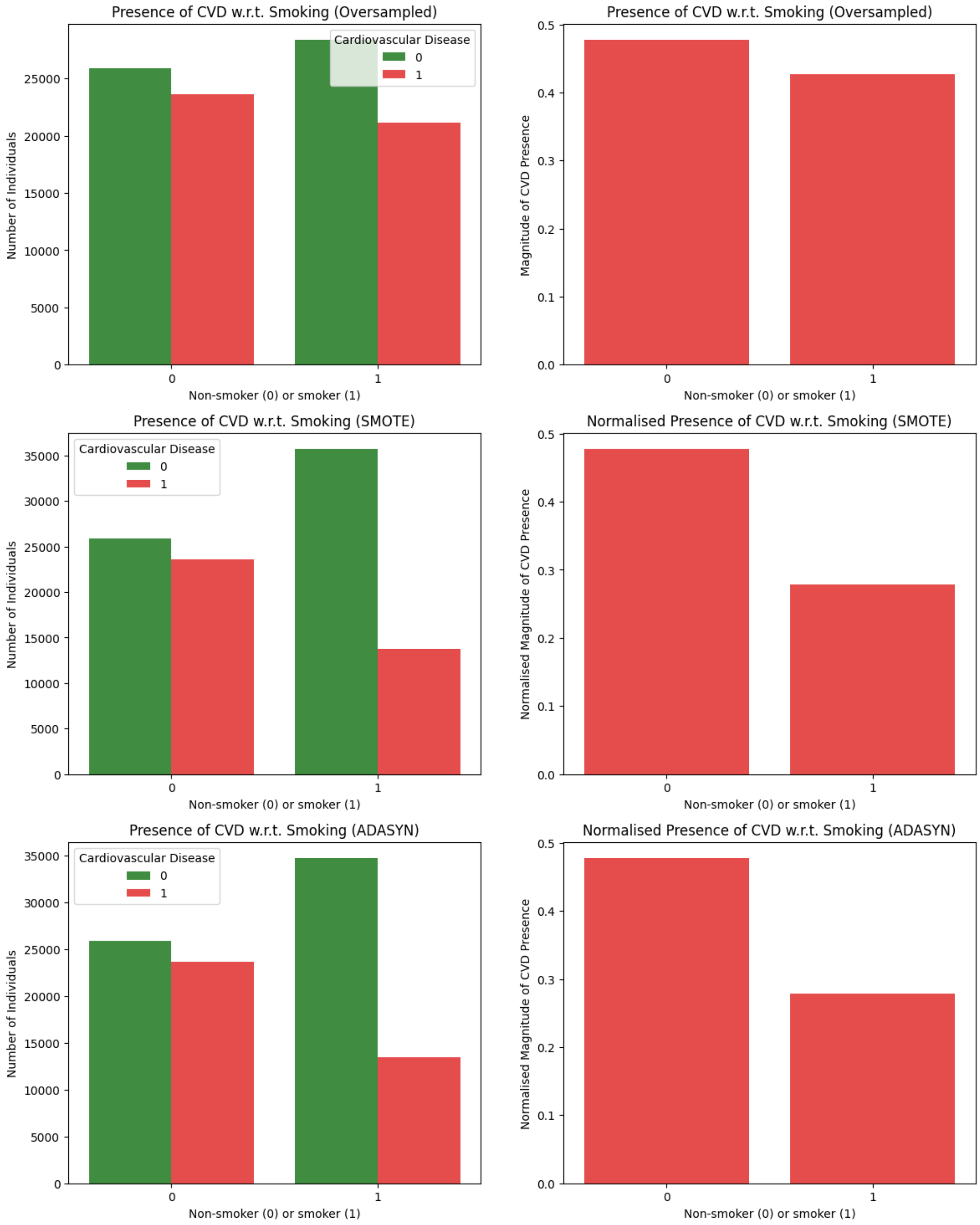


Figure 11 Resampled smoker data.

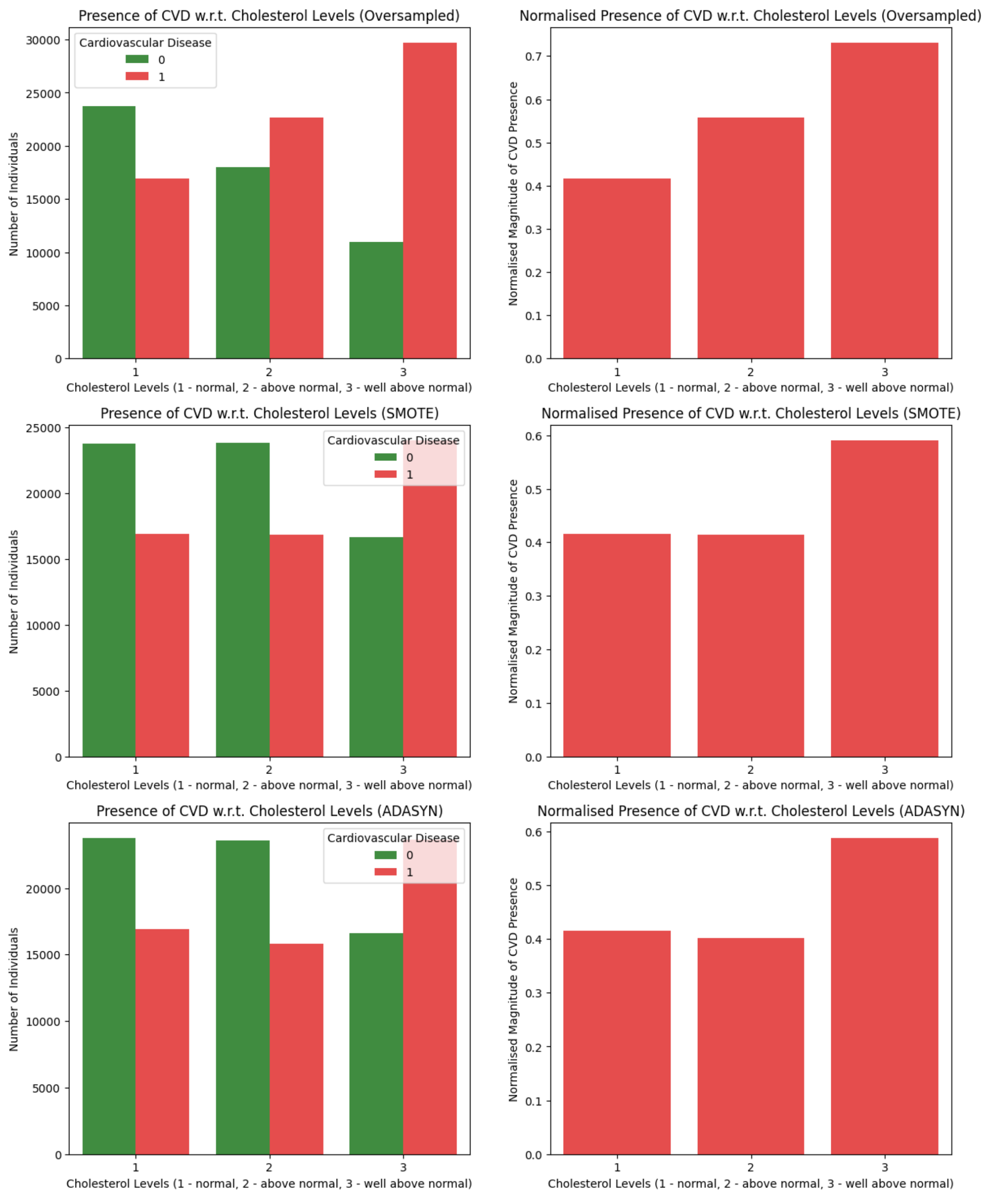


Figure 12 Resampled cholesterol data.

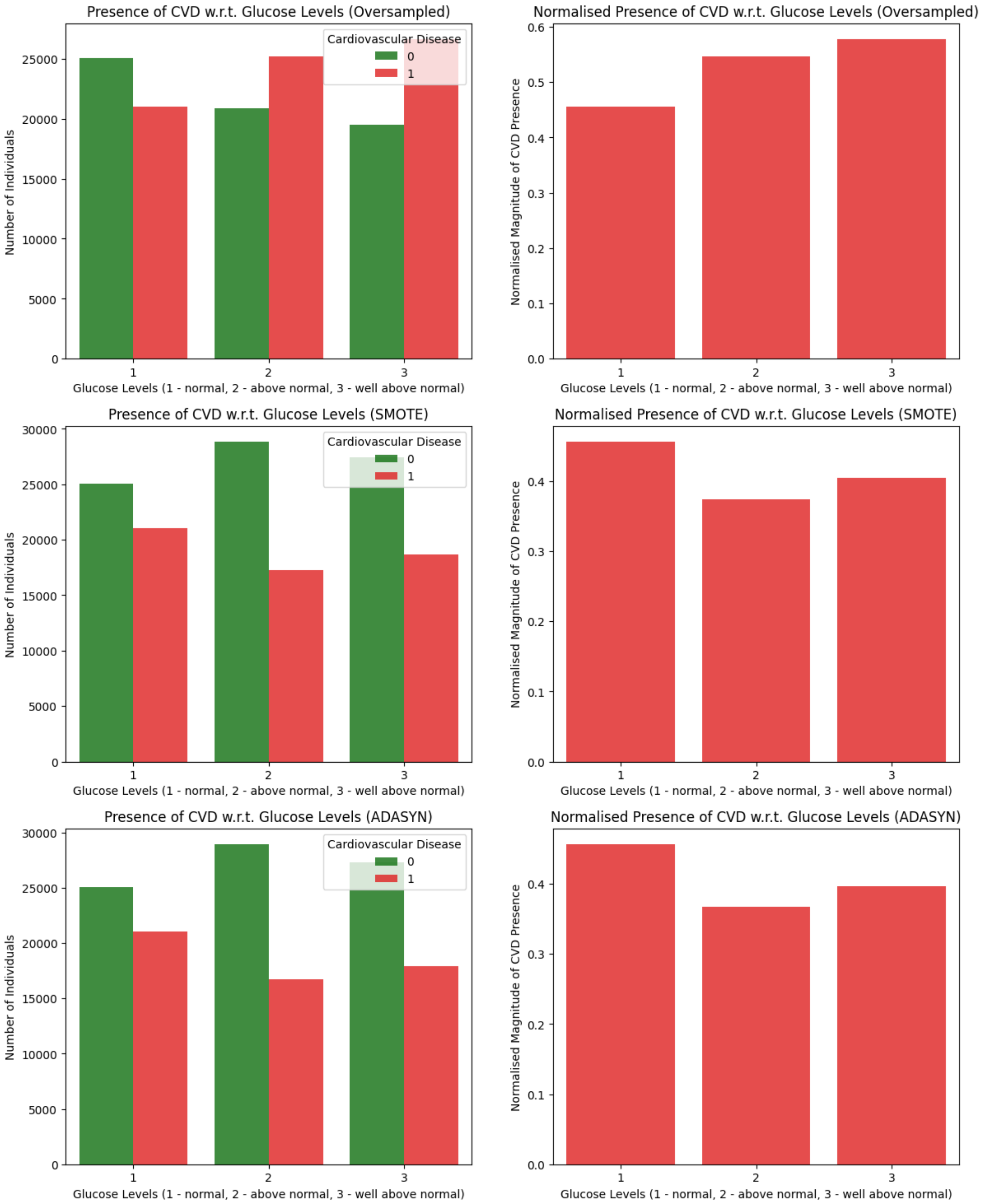


Figure 13 Resampled glucose data.

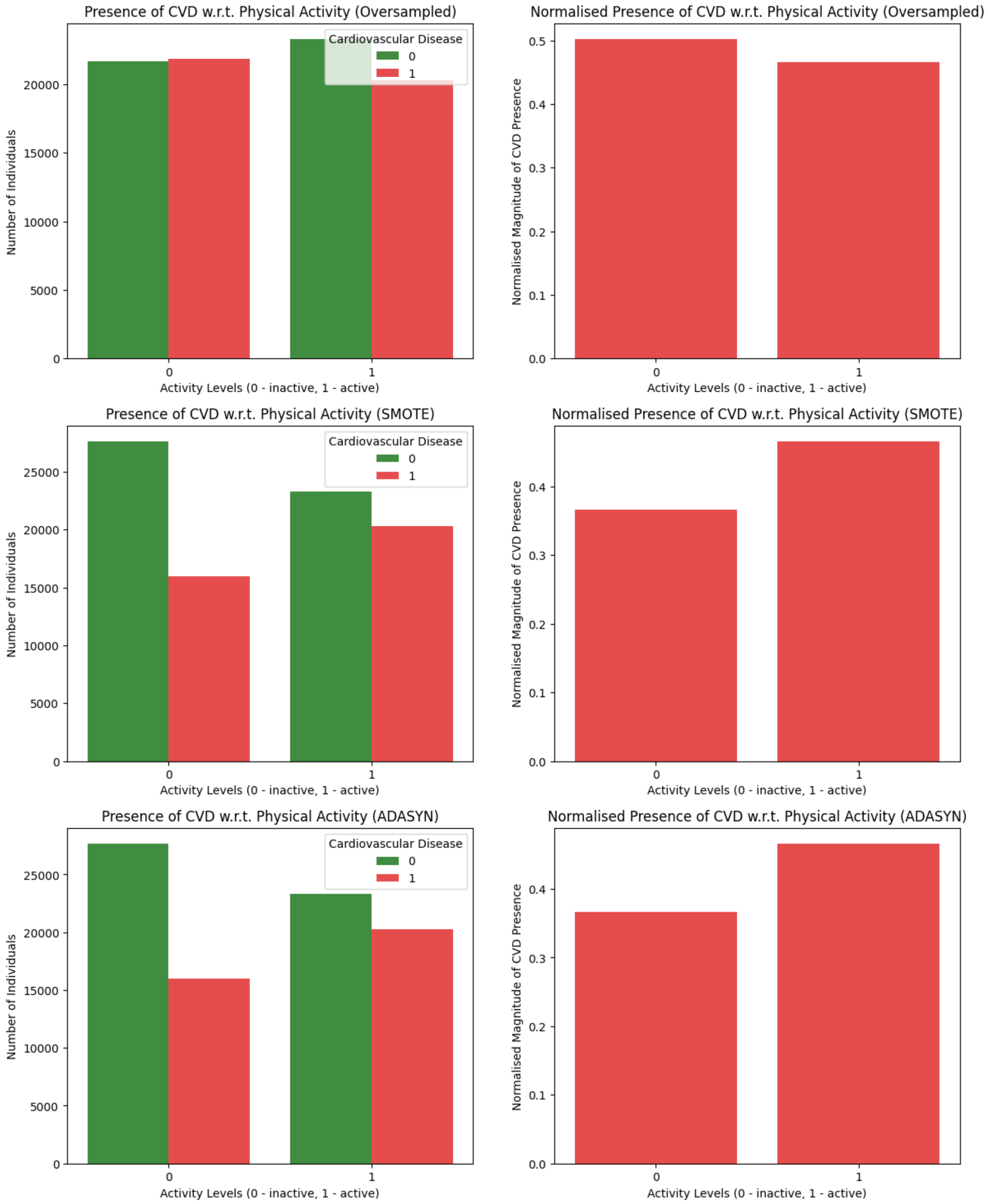


Figure 14 Resampled activity data.

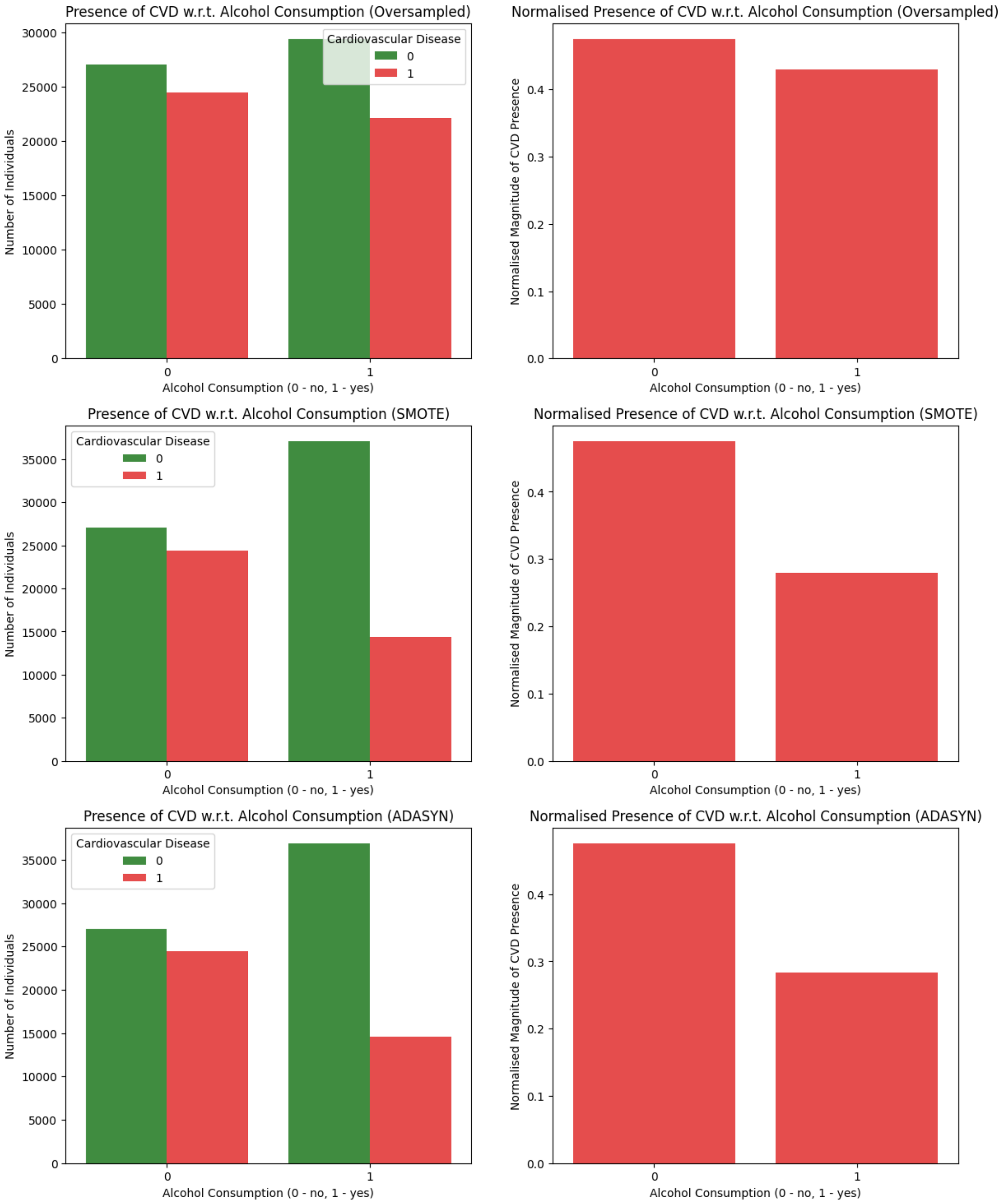


Figure 15 Resampled alcohol data.

Intriguing results were led to by the experimentation with data balancing via oversampling, the Synthetic Minority Oversampling Technique (SMOTE), and the Adaptive Synthetic Sampling (ADASYN) method. Figure 11 shows the results of balancing the smokers' and non-smokers' data, each method further amplifying the already incorrect trend shown in Figure 2. Figure 12 shows the cholesterol data balancing results, with levels 1 and 2 shifting closer together. The data balancing results for the glucose data in Figure 13 show a correlation loss. Balancing the activity data resulted in unfavourable outputs Figure 14 via SMOTE and ADASYN. The data balancing of alcohol data led to the amplified unfavourable results in Figure 15. In summary, data balancing has resulted in zero tangible improvements.

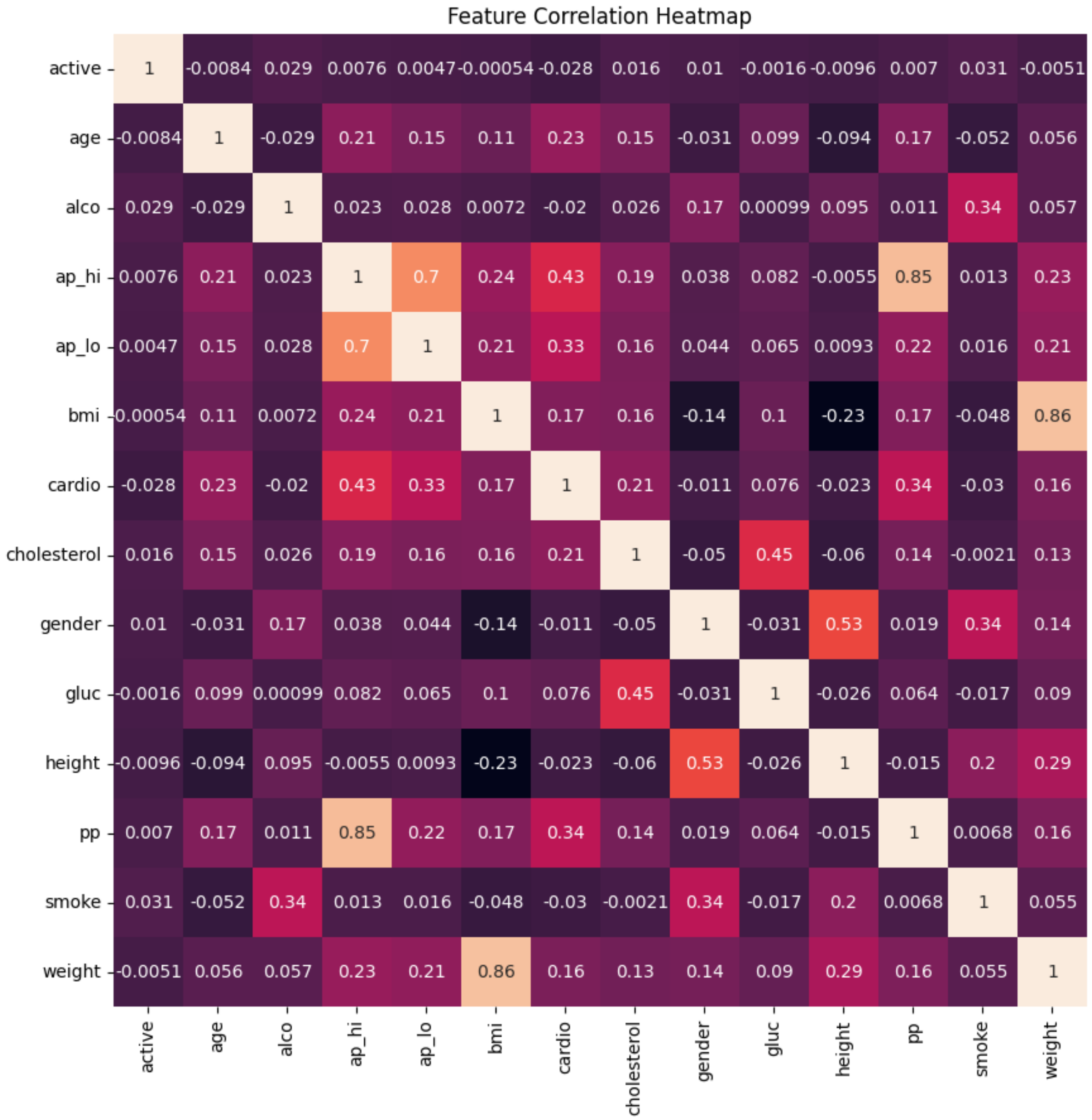


Figure 16 Feature correlation heatmap.

The feature correlation heatmap in Figure 16 allows for selecting optimum individual features. A high correlation is shown with 'cardio' for systolic BP, PP, diastolic BP, age, cholesterol, BMI, and weight, making these features optimum for extraction for classifier use.

Supervised Learning Methods

Supervised learning is a machine learning paradigm where model training utilises a labelled dataset containing input-output pairs. By learning the mapping between input features and corresponding target labels, the primary goal is to make predictions or classify new, unseen instances accurately. Without discovering inherent patterns and relationships within the data, supervised learning leverages known outcomes during training.

		Precision	Recall	F1-Score	Support
Decision Tree	CVD Absent	0.72	0.75	0.73	5709
	CVD Present	0.71	0.68	0.69	5201
	Accuracy			**0.71**	10910
	Macro Avg.	0.71	0.71	0.71	10910
	Wt. Avg	0.71	0.71	0.71	10910
Random Forest	CVD Absent	0.69	0.69	0.69	5709
	CVD Present	0.66	0.65	0.66	5201
	Accuracy			**0.67**	10910
	Macro Avg.	0.67	0.67	0.67	10910
	Wt. Avg	0.67	0.67	0.67	10910
Support Vector	CVD Absent	0.69	0.83	0.75	5709
	CVD Present	0.76	0.58	0.66	5201
	Accuracy			**0.71**	10910
	Macro Avg.	0.72	0.71	0.7	10910
	Wt. Avg	0.72	0.71	0.71	10910
K-Neighbours	CVD Absent	0.68	0.71	0.7	5709
	CVD Present	0.67	0.63	0.65	5201
	Accuracy			**0.68**	10910
	Macro Avg.	0.67	0.67	0.67	10910
	Wt. Avg	0.67	0.68	0.67	10910
Multilayer Perceptron	CVD Absent	0.74	0.7	0.72	5709
	CVD Present	0.69	0.73	0.71	5201
	Accuracy			**0.71**	10910
	Macro Avg.	0.71	0.71	0.71	10910
	Wt. Avg	0.71	0.71	0.71	10910

Table 4 Supervised classifiers performances.

The varying accuracies of supervised models, shown in Table 4 Supervised classifiers performances., indicate differences in abilities to capture patterns and predict correct outcomes. The Decision Tree model achieved the highest accuracy of 71.5%, indicating that it performed well in capturing complex decision boundaries inherent in the dataset despite being prone to overfitting. The Random Forest model, an ensemble classifier of multiple decision trees, leverages the combination of many weak learners to create a more robust and accurate model by reducing overfitting compared to the Decision Tree classifier. Unfortunately, the model yielded a lower accuracy of 67.2%, likely suggesting that the ensemble approach might not be the most suitable for this dataset since the Random Forest relies on the diversity of individual trees. The Support Vector Classifier (SVC) achieves, albeit a negligible drop compared to the Decision Tree, 71.2% accuracy, demonstrating its effectiveness in finding a hyperplane that best separates different classes in the feature space. The K-Nearest Neighbour (KNN) classifier brings 67.5% accuracy, implying that it makes effectual predictions based on the similarity to its neighbouring data points due to significant patterns and neighbour relationships within data. Finally, the Multilayer Perceptron (MLP), a type of neural network, achieved an accuracy of 71.2%, which suggests that the neural network architecture is appropriate for this dataset, with scope for further tuning and optimisation to improve performance. A crucial comparison to make is the time it takes to train each model. With each model given 80% of the dataset for training, the Decision Tree classifier took 0.3 seconds, the Random Forest took 5.6 seconds, the KNN classifier took 0.1 seconds, the MLP network took 11.7 seconds, and finally, the SVC took a whopping 1 minute and 10 seconds. The Decision Tree took the shortest time to train and produced the highest accuracy, making it the most suitable supervised classifier for CVD presence prediction.

Unsupervised Learning Methods

Unlike supervised learning, unsupervised learning operates on datasets without predefined target labels, thus particularly useful in scenarios where the objective is to discover inherent structures or hidden patterns within the data. Allowing models to identify patterns within the data bypasses the time-consuming and expensive stage of labelling data, thus making them well-suited for exploratory data analysis and uncovering hidden knowledge in large, complex datasets.

		K-Means			Agglomerative			DBSCAN		
		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Decision Tree	CVD Absent	0.75	0.92	0.82	0.72	0.75	0.73	0.72	0.75	0.73
	CVD Present	0.8	0.65	0.75	0.71	0.68	0.69	0.71	0.68	0.69
	Accuracy			**0.79**			**0.71**			**0.71**
	Macro Avg.	0.81	0.79	0.79	0.71	0.71	0.71	0.71	0.71	0.71
	Wt. Avg	0.81	0.79	0.79	0.71	0.71	0.71	0.71	0.71	0.71
Random Forest	CVD Absent	0.77	0.83	0.8	0.69	0.71	0.7	0.69	0.71	0.7
	CVD Present	0.8	0.72	0.76	0.68	0.65	0.66	0.67	0.65	0.66
	Accuracy			0.78			0.69			0.68
	Macro Avg.	0.78	0.78	0.78	0.68	0.68	0.68	0.68	0.68	0.68
	Wt. Avg	0.78	0.78	0.78	0.69	0.69	0.69	0.68	0.68	0.68
Support Vector	CVD Absent	0.68	0.82	0.74	0.68	0.82	0.74	0.68	0.82	0.74
	CVD Present	0.74	0.58	0.65	0.74	0.58	0.65	0.74	0.58	0.65
	Accuracy			**0.71**			**0.71**			**0.71**
	Macro Avg.	0.71	0.7	0.7	0.71	0.7	0.7	0.71	0.7	0.7
	Wt. Avg	0.71	0.71	0.7	0.71	0.71	0.7	0.71	0.71	0.7
K-Neighbours	CVD Absent	0.67	0.71	0.69	0.67	0.7	0.69	0.67	0.7	0.69
	CVD Present	0.66	0.62	0.64	0.66	0.62	0.64	0.66	0.62	0.64
	Accuracy			**0.66**			**0.66**			**0.66**
	Macro Avg.	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
	Wt. Avg	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
Multilayer Perceptron	CVD Absent	0.76	0.92	0.83	0.66	0.87	0.75	0.73	0.73	0.73
	CVD Present	0.88	0.67	0.76	0.79	0.52	0.62	0.71	0.71	0.71
	Accuracy			**0.8**			**0.7**			**0.72**
	Macro Avg.	0.82	0.8	0.8	0.72	0.69	0.69	0.72	0.72	0.72
	Wt. Avg	0.81	0.8	0.8	0.72	0.7	0.69	0.72	0.72	0.72

Table 5 Clustering-accelerated classifier performances.

Data clustering is an unsupervised learning approach widely popular for accelerating supervised learning models to increase accuracy. The clustering approach involves grouping similar data points into specific clusters based on their inherent characteristics to discover patterns, similarities, and structures within data without predefined labels. Exploring the three main clustering techniques showcases the effect of unsupervised learning fusion with supervised learning.

K-Means is a popular clustering type that aims to minimise the sum of squared distances between data points and their cluster centroids. Applying K-Means clustering has resulted in a dramatic boost in accuracy across the supervised learning models. Shown in Table 5 Clustering-accelerated classifier performances., the Decision Tree model now achieves 79.5% accuracy, the Random Forest model with 78% accuracy, and the MLP model with an accuracy of 80.1%. However, the SVC and KNN models generated a performance drop, featuring 70.6% and 66.5% accuracy, respectively. The 8% increase in Decision Tree accuracy suggests the formation of meaningful clusters by K-Means, resulting in a clear separation of data points into distinct groups. The Random Forest model exhibits a substantial 10.8% accuracy boost. The ensemble nature of Random Forests allows them to leverage the improved clustering to build more diverse and accurate decision trees. The MLP model achieves an 8.9% improvement, indicating that the cluster information helps enhance the representation of complex patterns in data, improving the neural network's ability to learn and generalise. Agglomerative clustering is a hierarchical approach to group data points based on similarity using a distance metric. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based algorithm that groups data points based on their density and identifies outliers as noise. Both agglomerative and DBSCAN clustering approaches brought no tangible improvements to the supervised learning models, signifying the creation of inefficient clusters that do not capture the inherent patterns in the data related to class separations.

The unsupervised acceleration provided with K-Means clustering has accelerated the MLP classifier substantially, resulting in the highest accuracy observed throughout this exploration with 80%, achieved with a quick training time of 3.7 seconds, making it the most suitable unsupervised-accelerated classifier for CVD presence prediction.

Conclusion

Despite accomplishing the primary objective of achieving a robust CVD classifier, significant scope for refinement still exists. For instance, implementing cross-validation and grid search to find the optimum machine learning model hyperparameters can improve accuracy. Additionally, the elbow method optimises K-Means to find the optimum number of clusters, further improving accuracy. Discretising features within the dataset by binning continuous numerical features into discrete categories enables simplified data relationship portrayals and greater robustness to outliers, ultimately improving model accuracy, especially with clustering.

By analysing the dataset, the crucial factors attributing to CVD presence were blood pressure, weight, age, and cholesterol. It was surprising to notice the lack of correlation that other critical features had, such as activeness, alcohol consumption, and smoking, leading to the theory of either incorrect data entries or a dishonest data feed. It's crucial to consider ethical considerations in deploying such models, ensuring fairness, transparency, and the avoidance of biases in predictions. Ongoing monitoring and updating of the model with new and accurate data are essential for maintaining its relevance and accuracy over time, contributing to the overall sustainability of the predictive system. Ethicality is conserved with the address of potential biases in the data, ensuring privacy and consent, and transparent communication of the model's predictions and limitations to the end-user.

References

[1] British Heart Foundation, “Cardiovascular Heart Disease,” British Heart Foundation, October 2019. [Online]. Available: https://www.bhf.org.uk/informationsupport/conditions/cardiovascular-heart-disease. [Accessed 8 January 2024].

[2] NHS, “Cardiovascular disease,” NHS, 22 April 2022. [Online]. Available: https://www.nhs.uk/conditions/cardiovascular-disease/. [Accessed 08 January 2024].

[3] H. K. Walker, W. D. Hall and J. W. Hurst, “Clinical Methods: The History, Physical, and Laboratory Examinations, 3rd ed.,” London, Butterworths, 1990, p. Chapter 16.

[4] Cleveland Clinic, “Pulse Pressure: What It Is and How to Calculate It,” Cleveland Clinic, [Online]. Available: https://my.clevelandclinic.org/health/body/21629-pulse-pressure. [Accessed 8 January 2024].