Predicting Risk of Cardiovascular Disease (CVD) Using Logistic Regression and Decision Trees with Random Forest

ABSTRACT

This project uses Decision Trees with Random Forest and Logistic Regression techniques on a chosen dataset to predict the risk of heart disease based on a variety of health factors such as Cholesterol, Blood Pressure, age and chest pain. This paper details an introduction on the background of the topic, and the importance of being able to predict such a prevalent disease. This is followed by a literature review and a more detailed look at the dataset and methodologies. Finally follows the project findings and a conclusion.

Keywords: Logistic Regression, Decision Trees, Random Forest, cardiovascular disease, risk factors

INTRODUCTION

Cardiovascular disease (CVD) is one of the main causes of death worldwide, estimated to be responsible for around 32% of deaths globally (WHO, 2021). Most cardiovascular diseases can be prevented by a change in a range of lifestyle factors such as diet, exercise and alcohol etc. Early detection is also very important in preventing the disease from progressing. The sooner a risk of CVD is detected, the sooner a change in lifestyle can take place and any necessary medication can be provided to minimise the likelihood of developing the disease.

This project utilises a synthetic dataset from Kaggle (Tusher, 2024) that contains around 70,000 data samples with information on the leading topics that may influence heart disease risk. The main aim is to use this data to produce an accurate predictor for future cardiovascular disease risk.

This project uses Supervised learning techniques, namely Logistic Regression and Decision Trees, to predict a person's risk of cardiovascular disease based on the analysis of whether common risk factors are present. Common risk factors include factors such as age, gender, genetic predisposition, cholesterol levels, blood pressure, exercise levels and chest pain. The data also includes each person's risk based on these factors as 1.0 (at risk) or 0.0 (not at risk).

OBJECTIVES

The main objectives of this project are to:

- Use multiple supervised machine learning models to accurately predict risk of cardiovascular disease.
- Discover which factors are the most important when it comes to predicting the risk.
- Display the data so that any important trends are visually clear to the reader, this will be performed using python modules such as Pandas or Matplotlib.
- Review the strengths and weaknesses of both supervised machine learning models.

LITERATURE REVIEW

There are a range of resources attempting to predict the risk of cardiovascular disease based on various factors. This section discusses a summary of a few of these research papers.

The first paper, 'Cardiovascular Risk Prediction Using Machine Learning in a Large Japanese Cohort', (Matheson et al., 2022) focuses on assessing the most common risk factors in those who develop cardiovascular disease in the Japanese population using random survival forests (RSF). The two most prevalent risk factors were previous history of heart disease and patients who were taking anticoagulant or antiplatelet medication.

The second, ‘A Comprehensive Review on Heart Disease Risk Prediction using Machine Learning and Deep Learning Algorithms’, analyses and summarises several popular research works that attempt to assess cardiovascular risk (Reddy et al., 2024). The paper concludes that the most common and effective machine learning techniques at predicting risk are Support Vector Machine and Random Forest. Another finding was that effectiveness was also affected by several other factors, such as sample size, chosen features and data preparation.

The third piece of literature, ‘Improving Cardiovascular Disease Prediction with Machine Learning Using Mental Health Data’, uses a different approach to the other papers and focuses on phycological risk factors (Dorraki et al., 2024). These are often overlooked in most attempts to use machine learning to predict CVD risk. Using a UK Biobank data set, the study found a significant increase in model accuracy when mental health factors were taken into consideration; The ML model had an accuracy of 71.31% with commonly used risk factors such as high blood pressure, and this increased to 85.13% when psychological factors were added.

The fourth paper, ‘Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms’, uses a variety of tree-based algorithms, including random forest to predict CVD risk (Asadi et al., 2024). The study determined that generalised mixed effect random forest was the best performing model. The study also used similar risk factors in the dataset, such as age and cholesterol levels. The paper concluded that the most important risk factors for predicting the disease were age, LDL levels (Low density Lipoprotein), family history, activity level and hypertension.

The final paper, ‘Machine learning approach for predicting cardiovascular disease in Bangladesh’, uses a variety of machine learning techniques, including decision trees and logistic regression, to also predict CVD risk (Hossain et al., 2024). Records from patients with and without cardiovascular disease were analysed to find the factors that were most vital to an accurate prediction. The paper also measured the accuracy of the different machine learning techniques and found decision trees and logistic regression to be two of the most accurate techniques.

DATA DESCRIPTION AND MANAGEMENT

The dataset was obtained from Kaggle (Tusher, 2024) and consists of synthetic data synthesised for the purpose of predicting heart disease risk. All of the risk factors and the data on whether the patient is at risk consist of binary values of either 1, indicating they have the particular risk factor or are at risk, or 0, indicating they do not have the particular risk factor or are not at risk. Age is the only non-binary value in the dataset. There are 13 columns displaying different potential risk factors as shown in figure 1.

heart_disease_risk_dataset_earlymed

Chest_Pain	Shortness_of_Breath	Fatigue	Palpitations	Dizziness	Swelling	Pain_Arms_Jaw_Back	Cold_Sweats_Nausea	High_BP	High_Cholesterol	Diabetes	Smoking	Obesity	Sedentary_Lifestyle	Family_History	Chronic_Stress	Gender	Age	Heart_Risk
0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	48.0	0.0
0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	45.0	0.0
1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	66.0	0.0
1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	60.0	1.0
0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	69.0	0.0
1.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	55.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	51.0	1.0
1.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	67.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	71.0	1.0

Figure 1: First 10 rows from the original dataset

Risk Factor:
Chest pain
Shortness of breath
Fatigue
Palpitations
Dizziness
Swelling
Pain in arms/jaw/back
Cold sweats or nausea
High blood pressure
High cholesterol
Diabetes
Smoking
Obesity
Sedentary lifestyle
Family history
Chronic stress
Gender
Age

Table 1: A comprehensive list of risk factors used in the dataset

The dataset also displayed the heart risk for each individual. Even though age is the only piece of data that is not in binary form, there should be no issues as both models can handle

a mixture of data. Feature scaling was used for logistic regression though as I wanted to ensure that age does not dominate due to having much larger numbers, so the data was standardised for this. This step is less important for decision trees and random forest.

One of the first tasks was to check for null or missing values within the data as I would need to decide what to do with these values. Any null values would need to be either removed or imputed, for example by predicting any missing values. After checking in Python, there were no null values present in the data so nothing further was required.

Another aspect to consider with the data was whether to retain all features/risk factors or to remove some columns. While decision trees are more robust when dealing with a larger number of features, logistic regression can be more reactive. Removing some of the risk factors can help to reduce redundancy and data overfitting. However, based on figure 2, none of the factors are particularly redundant, and none are highly correlated (over 0.75). Therefore, I decided to keep all features.

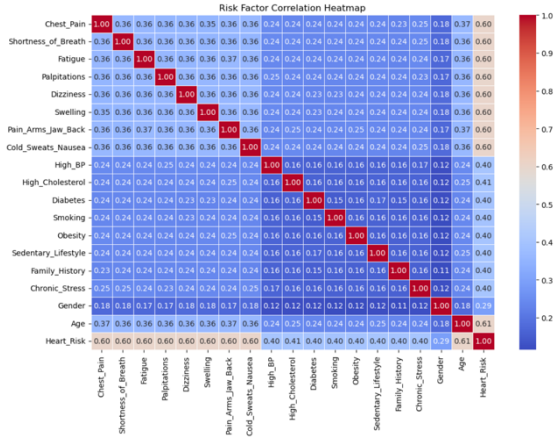

Figure 2: Heatmap showing correlation of all risk factors

DATA LIMITATIONS

There are a few limitations of the data to note; Even though the data has been purposely designed to aid prediction for cardiovascular disease, it has been synthetically produced. The data was made with real life figures in mind and used studies on heart disease as a guideline for data production. However, because of the synthetic nature of the data, it may not be capable of fully taking into account real-life clinical trends or figures. The data also contains no null values, which is rare in a real-life setting. Finally, another limitation to note is that the data consists of binary values for each risk factor (0 or 1), which can limit the real-world complexity of symptoms. Even with predicting risk, with this dataset it is either high or low, but in a clinical setting there may be a large variance in the risk level of an individual in a way that 'yes' or 'no' does not fully account for.

METHODOLOGY

The two machine learning methods, decision trees and logistic regression, were chosen because supervised learning is vital for this project. Supervised learning is required because we are dealing with data that has risk factors and a known risk prior to model creation. The model will then use this data to predict the risk of future unseen data. Whereas unsupervised learning is better for projects in which patterns in the data need to be detected, but not necessarily for predicting with new data. I chose decision trees and logistic regression as in the literature review these two methods were consistently some of the most accurate predictors.

Logistic regression is an easy-to-understand technique that can be used to produce a binary output, so is perfect for predicting something like probability or risk. A threshold is normally chosen, like 0.5, and this classifies the binary risk. This is useful because the chosen dataset is presented in binary form and works on the basis of risk being either high or low. The model is also fast and allows for multiple risk factors to be taken into account. Figure 3 shows the logistic regression function.

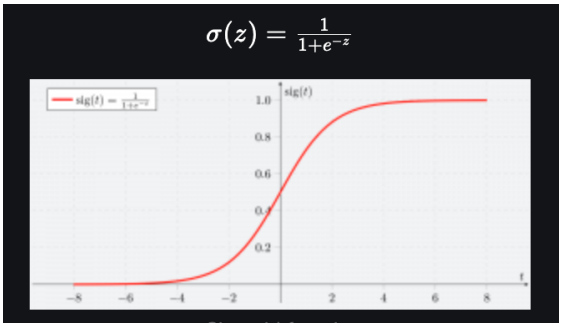

Figure 3: The logistic regression sigmoid function (GeeksforGeeks 2024)

Decision trees with random forests are also easy to understand, and are good at accounting for complex, potentially non-linear relationships between multiple factors. Although they're slower than logistic regression, they are an essential component to supervised learning. I decided to use random forest in addition to decision trees as it can improve the overall accuracy of the model and makes it less prone to overfitting, even when particular risk factors may be highly correlated. This is vital to ensuring the model is applicable to real-world data. Random forest involves the use of multiple decision trees and displays the average of those predictions.

TESTING

Starting with Logistic Regression, the model was implemented in Python using sklearn. The data was split into a training and test set, with 75% of the data being used for training the model, and the remaining 25% being used as the test set to subsequently test the accuracy of the model. The result is shown below in figure 4 and shows a 0.99 score for all factors. This high accuracy level could potentially indicate some challenges with the data and is discussed in further detail in the 'challenges' section of the paper. The figure includes precision, recall and the f1-score, in which logistic regression scores highly in all. Precision refers to the false positives, and a high score shows a low level of false positives. A high recall refers to a lower number of false negatives. Finally, the f-1 score takes into account both precision and recall, as there should be a good balance between the two. The F-1 score is also excellent for the model.

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	8839
1.0	0.99	0.99	0.99	8661
accuracy			0.99	17500
macro avg	0.99	0.99	0.99	17500
weighted avg	0.99	0.99	0.99	17500

Figure 4: Summary of Logistic Regression model performance

The next step was implementing the decision tree. This again used the same split of the training and test sets from the data. Figure 5 shows the accuracy statistics for this model. As you can see, the accuracy was 93%, which is lower than with logistic regression, but still highly accurate and may be more applicable to real-world scenarios. The precision, recall and f1-score were slightly lower, but still excellent for this model.

Decision Tree Accuracy: 0.9283428571428571

	precision	recall	f1-score	support
0.0	0.93	0.93	0.93	8839
1.0	0.93	0.93	0.93	8661
accuracy			0.93	17500
macro avg	0.93	0.93	0.93	17500
weighted avg	0.93	0.93	0.93	17500

Figure 5: Summary of Decision tree model performance

Finally, is a summary of the random forest model performance. As shown in figure 6, there is an accuracy level of 0.99, similar to logistic regression. Again, the precision, recall and f-1 score were all excellent for this model and were higher than just using the decision tree model alone.

Random Forest Accuracy: 0.9927428571428571

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	8839
1.0	0.99	0.99	0.99	8661
accuracy			0.99	17500
macro avg	0.99	0.99	0.99	17500
weighted avg	0.99	0.99	0.99	17500

Figure 6: Summary of Random Forest model performance

CHALLENGES

While testing the data, I came across a few challenges. These were mainly due to the fact that the data had been synthetically produced. I came across the first challenge when assessing the performance of the models; The logistic regression model had a high level of accuracy (0.99), which initially seems positive, but may not be realistic in comparison to a real-life dataset. However, when assessing the decision tree accuracy, the model was 0.93 which is slightly more realistic. The reason for the high accuracy could be because the synthetic dataset may be 'too perfect', making it simple for the model to learn. This ease of learning can lead to overfitting of the data, making it potentially not applicable to real-world scenarios. I did test removing some of the best performing features to see if there was any notable difference to the accuracy, but the difference was minimal.

Another challenge came when synthesising the decision tree, initially the tree was much too complicated and would have been very unclear to extract any useful information out of. This was fixed by pruning the decision tree by adjusting the max_depth to 4, which allowed a clear view of the tree while still maintaining a good level of detail.

RESULTS

Logistic Regression:

After implementing the logistic regression model, the first step taken was to visualise a confusion matrix, as can been seen in figure 7. The matrix indicates 8602 True Positives, 76 False Positives, 8763 True Negatives and just 59 False Negatives. The confusion matrix suggests that the model is predicting with high accuracy and precision.

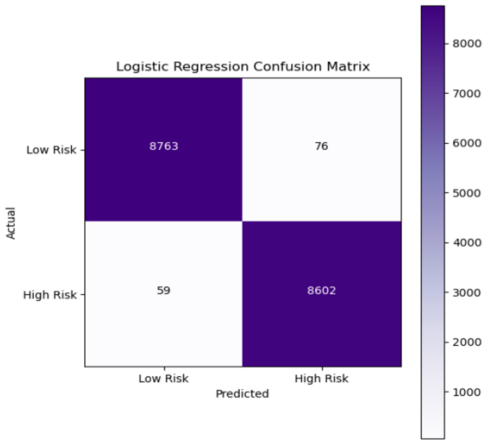

Figure 7: Logistic Regression Confusion Matrix

An important objective for this project was also to discover the most important risk factors when predicting CVD risk. This was visualised using a bar chart showing the top 5 most important risk factors that were used in the logistic regression model. As you can see from figure 8, the top risk factor was fatigue. However, the top 5 features all had a similar level of importance in the model as their coefficient values were all very close in value.

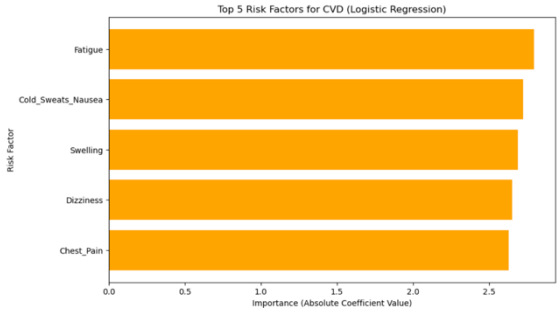

Figure 8: Bar chart displaying top 5 risk factors with the random forest model

Decision Tree with Random Forest:

The confusion matrix was also visualised for the decision tree and random forest. As you can see in figure 9, the matrix indicates 8024 True Positives, 617 False Positives, 8222 True Negatives and 637 False Negatives. This makes the model less accurate than with logistic regression, but it is still performing with a high precision.

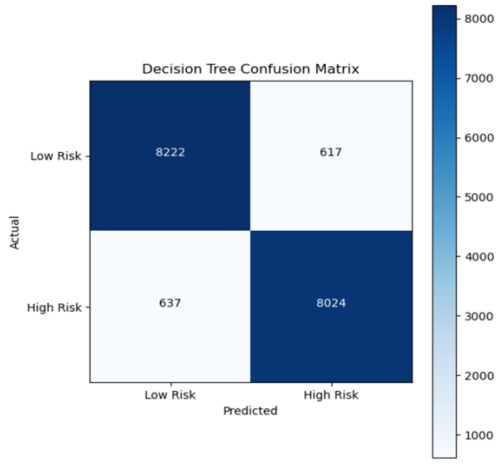

Figure 9: Decision Tree Confusion Matrix

For figure 10, displaying the confusion matrix for random forest, the matrix indicates 8602 True Positives, 68 False Positives, 8771 True Negatives and just 59 False Negatives. This gives a similar accuracy to logistic regression.

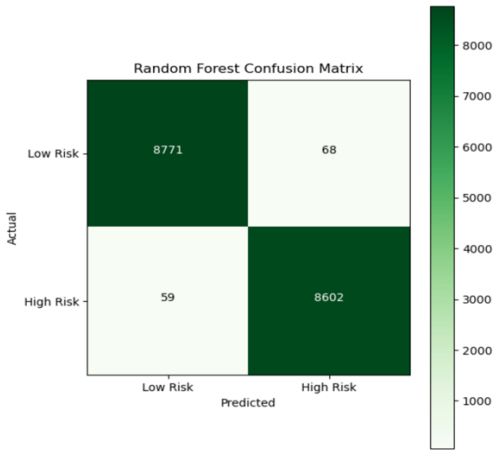

Figure 10: Random Forest Confusion Matrix

The results from the decision tree were visualised with the tree itself and again with a bar chart. As you can see from figure 11 and 12, the top risk factor was cold sweats and nausea. This can also be compared to random forest in figure 13, where the most important risk factor in the model was Age. The top risk factor for each model, including logistic regression, was different. However, logistic regression and random forest both have cold sweats and nausea as the second most important factor which aligns similarly to the decision tree model. Another observation to note was that for the decision tree, cold sweats and nausea was followed by age, but after this the other three factors in the top 5 most important have a significantly less of a difference in importance. However, when using random forest, the bars are more consistent in regard to importance after age.

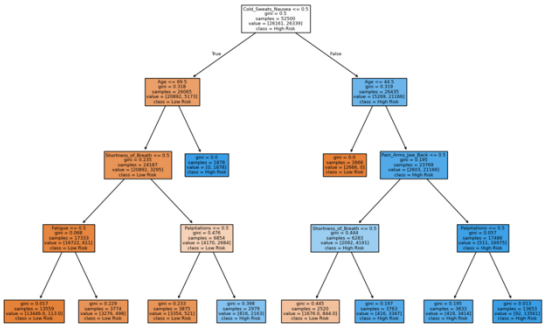

Figure 11: Decision tree of the data

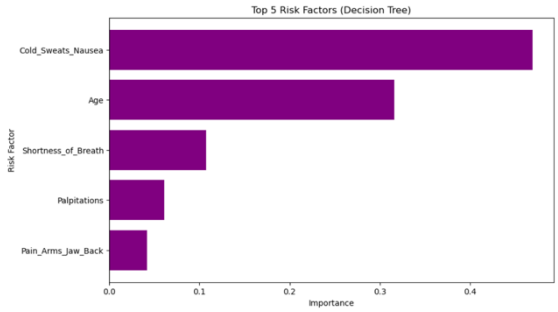

Figure 12: Bar chart displaying top 5 risk factors with the decision tree model

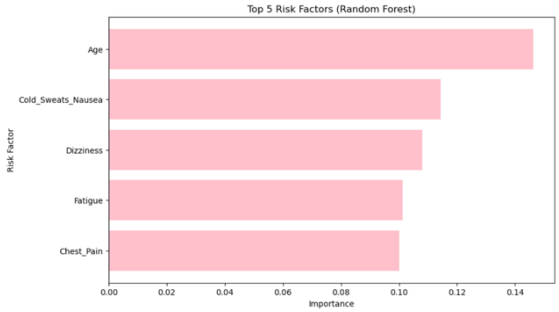

Figure 13: Bar chart displaying top 5 risk factors with the random forest model

CONCLUSION

Overall, when comparing the models used, the most accurate models were logistic regression and the decision tree when used with random forest. The results showed that using random forest with the decision tree improved the prediction accuracy. This may be because using random forest can prevent overfitting of the data. Both models performed exceptionally well and had a high accuracy level with their predictions, with accuracy being over 93%. However, the models still varied with what risk factor they each classed as being the most important in predicting risk of Cardiovascular disease.

As mentioned earlier in the paper, due to some limitations with the synthetic data, in the future I would implement the models again but using different real-world data to compare if the accuracy levels are affected with a different data set. In the future, if using the same dataset, I would experiment more with dimensionality reduction to see if this had any substantial impact on the results.

POSSIBLE APPLICATIONS

As discussed previously in the paper, the main application of this project would be using the model for preventative treatment. If we can predict those most at risk of developing cardiovascular disease in the future, preventative measures can be taken to minimise the risks. Cardiovascular disease is one of the leading causes of deaths worldwide, so it is vital that we have a system that can accurately highlight those most at risk. Early detection can aid in saving lives and lowering the risk of future complications.

REFERENCES

Dataset:

• **Tusher, M.A.** (2024) *Heart Disease Risk Prediction Dataset*. Available at: https://www.kaggle.com/datasets/mahatiratusher/heart-disease-risk-prediction-dataset [Accessed 14 Feb. 2025].

Scientific Papers:

Asadi, F., Homayounfar, R., Mehrali, Y., Masci, C., Talebi, S. and Zayeri, F. (2024) 'Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms', *Scientific Reports*, 14(1). Available at: https://doi.org/10.1038/s41598-024-72819-9.

Dorraki, M., Liao, Z., Abbott, D., Psaltis, P.J., Baker, E., Bidargaddi, N., Wardill, H.R., Anton, Narula, J. and Verjans, J.W. (2024) 'Improving Cardiovascular Disease Prediction With Machine Learning Using Mental Health Data', *JACC Advances*, 3(9), pp. 101180–101180. Available at: https://doi.org/10.1016/j.jacadv.2024.101180.

Hossain, S., Hasan, M.K., Faruk, M.O., Aktar, N., Hossain, R. and Hossain, K. (2024) 'Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023', *BMC Cardiovascular Disorders*, 24(1). Available at: https://doi.org/10.1186/s12872-024-03883-2.

Matheson, M.B., Kato, Y., Baba, S., Cox, C., Lima, J.A.C. and Ambale-Venkatesh, B. (2022) 'Cardiovascular Risk Prediction Using Machine Learning in a Large Japanese Cohort', *Circulation Reports*, 4(12), pp. 595–603. Available at: https://doi.org/10.1253/circrep.cr-22-0101.

Reddy, V., Karna, V.R., Janamala, V.P., Rao, K., Ravi, V. and Tummala, A.B. (2024) 'A Comprehensive Review on Heart Disease Risk Prediction using Machine Learning and Deep Learning Algorithms', *Archives of Computational Methods in Engineering*. Available at: https://doi.org/10.1007/s11831-024-10194-4.

Python Code:

Datacamp (2019) 'Understanding Logistic Regression in Python', *Datacamp.com*. Available at: https://www.datacamp.com/tutorial/understanding-logistic-regression-python [Accessed 3 Mar. 2025].

GeeksforGeeks (2024) 'How to Plot Confusion Matrix with Labels in Sklearn?', *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/how-to-plot-confusion-matrix-with-labels-in-sklearn/ [Accessed 3 Mar. 2025].

GeeksforGeeks (2024). *Understanding Logistic Regression*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/understanding-logistic-regression/ [Accessed 5 Mar 2025]

Organization Website:

World Health Organization (WHO) (2021) *Cardiovascular Diseases (CVDs)*. Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [Accessed 14 Feb. 2025]