Group E

Using Data Science to improve Heart Disease Detection

COMP3608 - Intelligent Systems Project

Josiah Joel - 816030501
Jarrod Moore - 816028020
Raoul Sagram - 816032025

Table of Contents

Introduction

Cardiovascular Disease remains one of the leading causes of deaths worldwide, it represents a group of medical disorders that impact the heart and blood vessels. According to the World Health Organization (WHO), approximately 17.9 million people died from cardiovascular diseases (CVDs) in 2019, representing about 32% of all global deaths. The non-communicative nature of these diseases indicates that identification of risk factors and application of preventative measures is highly important.This project aims to harness the power of machine learning to develop a predictive model for assessing an individual's risk of cardiovascular disease based on their health data and lifestyle choices. Traditional methods for diagnosis and risk assessment of Cardiovascular Disease by human doctors are susceptible to human errors. By leveraging machine learning algorithms and multiple health datasets, this project aims to extract meaningful correlations and create a robust classifier utilising health info and lifestyle choices such as alcohol consumption and smoking. Such a tool will have great impact on how preventative medicine is executed by providing patients and their doctors with actionable insights into their risk which can reduce the prevalence of Cardiovascular Disease. The goal of this research is to maximise the F1 score in 3 traditional prediction algorithms and find the most impactful features to raise awareness and reduce the risk of CVD in Trinidad and Tobago.

Related Works

The exploration of utilising machine learning algorithms in healthcare has become a topic of study within recent years. Particularly within the area of risk prediction of Cardiovascular Disease (CVD), research into how patient data and machine learning models can be coupled to improve health outcomes has been conducted by a variety of teams. This section aims to provide an overview into such works, and their contributions.

In their 2018 paper, Shinde et al provided a review on a number of machine learning models being applied to the goal of predicting health risks or performing diagnosis. Their study offers insights into the application of machine learning models such as Neural Networks and K Nearest Neighbours for a number of purposes including predicting breast cancer, swine flu and comorbid disease prediction.

The research conducted by Adler, E.D. et al, detailed the application of machine learning to improve heart failure risk predictions. Through the use of electronic medical records, the researchers trained a Boosted Decision tree algorithm to perform a binary classification of high versus low mortality risk based on patient medical data.

In their study titled “Risk estimation and risk prediction using machine-learning methods”, Kruppa, J. et al studied the use of machine learning methods in performing disease risk prediction based on genetics, the difficulty in performing risk classification upon genomes and the creation of rules to construct and evaluate medical risk classifiers. The paper goes on to critique the limitations of classical methods of model evaluation such as Area under the Curve (AUROC).

These studies collectively underscore the transformative potential of machine learning in revolutionising cardiovascular risk prediction and management. By harnessing the power of advanced computational techniques and vast datasets, researchers are able to develop robust predictive models capable of identifying individuals at elevated risk of cardiovascular disease. Moreover, these studies highlight the importance of interdisciplinary collaboration between computer scientists, clinicians, and geneticists in advancing the field of intelligent health risk prediction systems.

In summary, the aforementioned studies represent seminal contributions to the field of machine learning-based risk prediction for cardiovascular disease, laying the foundation for future research endeavours aimed at improving patient outcomes and reducing the burden of cardiovascular morbidity and mortality.

Datasets

- Cardiovascular Disease dataset (Dataset 1):
    - https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data
- Heart Failure Prediction Dataset (Dataset 2):
    - https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
- Heart Disease Dataset (Dataset 3):
    - https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Methodology and Experimental Design

In our task of binary classification in the context of heart disease prediction, we used three cutting-edge algorithms: Logistic Regression, Random Forest, and a backward propagating Artificial Neural Network (ANN) configuration. Each algorithm was carefully selected for its distinct strengths and potential contributions to our predictive modelling efforts.

Our decision to employ these three algorithms was grounded in their proven efficacy and versatility across a range of data science applications. Logistic Regression is a fundamental yet powerful method and was chosen for its ability to estimate the probability of binary outcomes based on input features. Meanwhile, Random Forest, using its ensemble learning approach, was deemed ideal for using multiple decision trees to enhance predictive accuracy.

Additionally, we incorporated an ANN architecture featuring 64 nodes in a single hidden layer, trained over 500 epochs with a modest learning rate of 0.001. This configuration was used to capitalise on the ANN's capacity for complex pattern recognition and nonlinear relationships within the data, thus offering a complementary perspective to the other algorithms.

To get robust model performance, we prioritised the maximisation of the F1 score.
F1 score balances both precision and recall, ensuring accurate identification of stroke cases while minimising false positives and false negatives, which are critical for timely intervention and treatment.
Precision measures the accuracy of positive predictions, indicating the proportion of correctly identified positive instances among all instances predicted as positive.
Recall measures the completeness of positive predictions, indicating the proportion of correctly identified positive instances among all actual positive instances.

Furthermore, to gain insights into the predictive mechanisms of each algorithm, we identified and scrutinised the four most impactful features utilised in the prediction process. By exploring these critical features, we aimed to enhance interpretability and facilitate domain-specific insights into the underlying factors driving heart disease prediction.

In our approach to data preprocessing and algorithmic selection for heart disease prediction, we prioritised methodological rigour and fairness to ensure robust model performance. The datasets, inherently balanced straight out-of-the-box, so the use of imbalanced learning techniques like oversampling or undersampling was not done, this preserves the integrity of the data distribution and does not add noise. However, to maintain data quality and mitigate the potential impact of outliers, outlier detection was done on the numerical columns, and the removal of corresponding rows.

Central to our experimental design was the utilisation of identical, meticulously cleaned datasets across all algorithmic implementations. This standardised approach aimed to mitigate the influence of data discrepancies and ensure equitable comparisons among the algorithms. To guard against overfitting and uphold the principles of fairness, all algorithms underwent rigorous evaluation through k-fold cross-validation, with 5 folds employed to provide robust estimates of model performance.

An essential consideration in our methodology pertained to the treatment of categorical features, particularly in the context of Logistic Regression. Recognizing the algorithm's requirement for one-hot encoding of categorical features, we transformed these variables into boolean representations. One-hot encoding, a widely employed technique in machine learning, enables the transformation of categorical variables into a binary format.

However, the application of one-hot encoding posed unique challenges in the context of Random Forest, where such encoding is generally discouraged due to its potential to introduce sparsity and inefficiencies in tree-based algorithms. Consequently, to maintain consistency in our comparative analysis, we used one-hot encoding for the Artificial Neural Network (ANN) models. And the experiment was conducted including both one-hot encoded and original categorical datasets in separate instances for the ANN.

By systematically comparing the performance of Logistic Regression, ANN with one-hot encoding, Random Forest, and ANN without one-hot encoding, we aimed to find the relative strengths and weaknesses of each algorithmic approach. This evaluation strategy not only facilitated a comprehensive understanding of algorithmic behaviour but also underscored our commitment to methodological transparency and scientific rigour in the pursuit of accurate heart disease prediction models.

Mathematical Model

Objective function (Z):

$$\text{Maximize } F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Subject to: Constraints for Logistic Regression, Random Forest and ANN

Constraint 1: False Positive (FP)

Constraint 2: False Negative (FN)

Constraint 3: True Positive (TP)

Constraint 4: True Negative (TN)

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The 4 most important features are used from the maximized F1 score

Logistic Regression:

$$\sigma(z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \quad (1)$$

$$|\beta_1|, |\beta_2|, |\beta_3|, |\beta_4|$$

Random Forest:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T(x; \Theta_i) \quad (2)$$

$$\text{importance}_1, \text{importance}_2, \text{importance}_3, \text{importance}_4$$

Neural Networks:

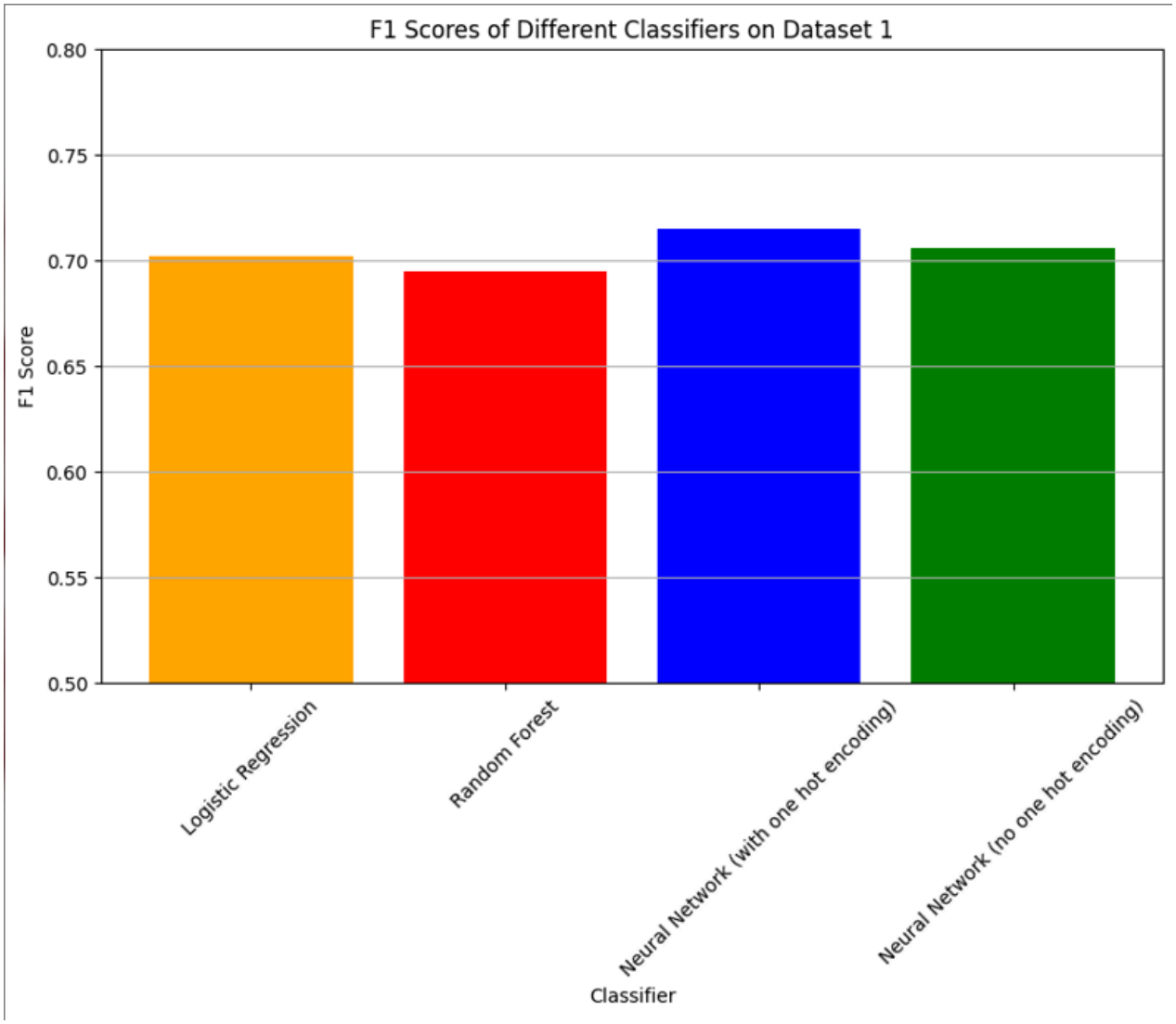
$$\frac{\partial L}{\partial w_{ij}^{(l)}} = \frac{\partial L}{\partial a_i^{(l)}} \cdot \frac{\partial a_i^{(l)}}{\partial z_i^{(l)}} \cdot \frac{\partial z_i^{(l)}}{\partial w_{ij}^{(l)}} \quad (3)$$

$$|w_1|, |w_2|, |w_3|, |w_4|$$

This simply declares the constraints as the values of the confusion matrix and defines the objective function to maximise the F1 score. It also shows that the 4 most impactful features are recorded for the result. Logistic regression predicts the probability of a binary outcome based on input features by fitting a sigmoid curve to the data. Random forest combines multiple decision trees to make predictions by averaging the results of individual trees. Neural networks learn complex patterns by processing data through interconnected layers of neurons, adjusting weights during training to minimise prediction errors

Results

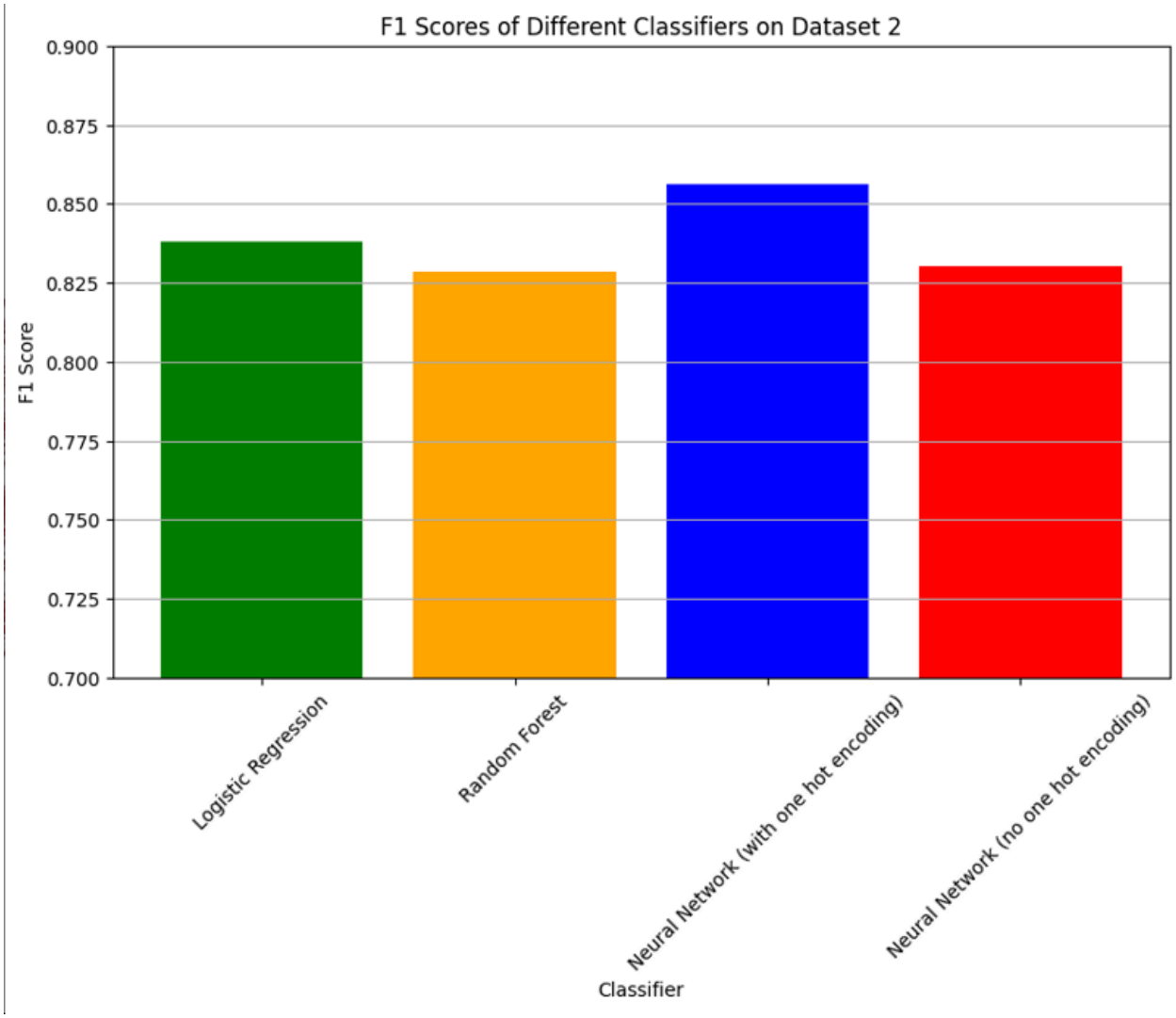
Dataset 1 results



Best Algorithm: ANN with one hot encoding
HIghest F1 score: 0.712
Most Important Features:

- Cholesterol
- Alcohol Consumption
- Gender
- Physical Activity

Dataset 2 results

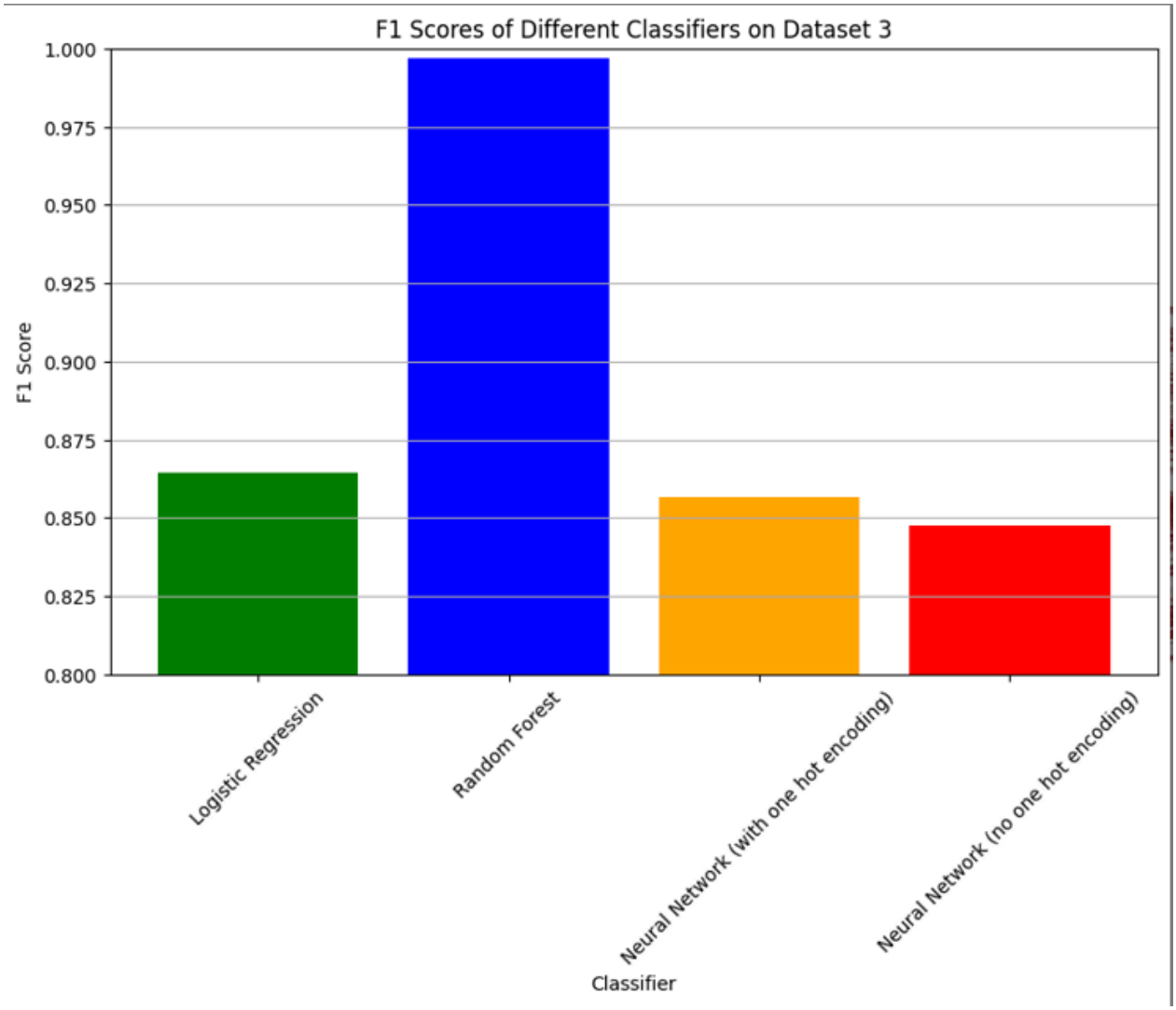


Best Algorithm: ANN with one hot Encoding
Highest F1 score: 0.861
Most Important Features:

- ST_Slope
- Sex
- Chest Pain
- Old peak

Our results were also compared to a Kaggle competition which used a similar dataset to dataset 2:
https://www.kaggle.com/competitions/MLSA-KFS-DS_0/data
The results were comparable but our results were 0.02 off 6th (last) place.

Dataset 3 results



Best Algorithm: Random Forest
Highest F1 Score: 0.997
Most Important Features:

- Chest Pain
- Thalach
- Ca
- Oldpeak

Overfitting Most likely present.

ANN with one hot Encoding F1: 0.869

Full Results

Dataset 1

Algorithm	F1	Features
Neural Network (One-Hot)	0.714	cholesterol, alco, gender, active
Neural Network (No One-Hot)	0.707	smoke, alco, gender, height
Logistic Regression (One-Hot)	0.701	active, alco, smoke, gender
Random Forest (No One-Hot)	0.695	Weight, Height, Systolic Blood Pressure, Age

Dataset 2

Algorithm	F1	Features
Neural Network (One-Hot)	0.848	ST_Slope, Sex, ChestPainType, ExcerciseAngina
Neural Network (No One-Hot)	0.840	ExcerciseAngina, Oldpeak, Sex, ST_Slope
Logistic Regression (One-Hot)	0.838	ST_slope, ChestPainType, Sex, Oldpeak
Random Forest (No One-Hot)	0.828	ST_Slope, Oldpeak, ChestPainType, MaxHR

Dataset 3

Algorithm	F1	Features
Random Forest (No One-Hot)	0.997	cp, thalach, ca, oldpeak
Logistic Regression (One-Hot)	0.864	ca, cp, thal, sex
Neural Network (One-Hot)	0.860	ca, cp, thal, sex
Neural Network (No One-Hot)	0.855	sex, exang, cp, ca

Discussion

Artificial Neural Networks(ANN) utilising one hot encoding were the most consistent best performer for the first 2 datasets, with the remaining dataset having Random Forest Classifier be its best performer.
Nevertheless, ANN had the best consistent performance with F1 scores between 0.71 and 0.87. From this we extracted the top four most impactful features in the classification task for each dataset. The 1st Dataset's top 4 most important features were found to be Cholesterol Level, Patient Gender, whether they consumed alcohol and their physical activity. For the 2nd dataset they were,the direction of the slope of the ST segment of the Electrocardiogram (ECG) taken during peak exercise (ST_slope), the Patient's Gender, the type of chest pain experienced by the patient and the change in the ST segment of the ECG during physical activity compared to the ST segment at rest (old peak). Lastly for the 3rd dataset the most important features were the type of chest pain experienced by the patient again , the maximum heart rate of the patient, the number of major blood vessels coloured by fluoroscopy and the change in the ST segment of the ECG during physical activity compared to the ST segment at rest again.

As we utilised one hot encoding for the ANN, we were able to see what values had the biggest impact on classification. For cholesterol levels, being above normal was correlated with being at risk. For ST_slope, having it be down sloping indicated the greatest risk of having CVD. As for patient gender, males seemed to be more at risk than females. For lifestyle choices, Consuming alcohol or smoking was also found to represent a higher risk of having CVD.

Dataset 1: In Dataset 1, the one-hot encoded neural network emerges as the top performer, showcasing superior F1 score and accuracy. Its key features include cholesterol levels (well above normal), male gender, alcoholism, and smoking. Notably, the inclusion of one-hot encoding enhances the model's performance, emphasising the importance of categorical variables representation.

Dataset 2: Here, both the one-hot encoded neural network and logistic regression exhibit strong performance. Features such as ST_slope, chest pain type, and gender (male) prove influential. These findings underscore the significance of nuanced cardiovascular indicators in predictive modelling.

Dataset 3: Surprisingly, in Dataset 3, the random forest algorithm outshines others despite standard initialization. Key features include chest pain, number of major vessels coloured by fluoroscopy, and thal. The occurrence of such exceptional performance is likely attributed to overfitting.

Conclusion

Cardiovascular disease remains one of the most impactful lifestyle diseases not only in Trinidad and Tobago, but around the world. Predicting a person's risk of the disease from health aspects is important to combating it. Using an artificial neural network with the features one hot encoded is a great way of CVD predictions. From three similar datasets, the features of Cholesterol Level, Gender, Alcohol Consumption, Glucose Level, Type of Reported Chest pain, amongst other features were found to be the best features to collect data on for the creation of an applicable Cardiovascular Disease Prediction model.

Finally, the application of machine learning algorithms in predicting cardiovascular disease (CVD) represents a promising frontier in healthcare. However, success in this endeavour hinges on meticulous attention to dataset nuances and judicious feature selection. By delving deeper into these aspects, healthcare practitioners can unlock the full potential of predictive analytics to significantly impact patient outcomes in cardiovascular health. As technology continues to evolve, the integration of machine learning into clinical practice promises to redefine the paradigm of cardiovascular care, ushering in an era of proactive health management and improved quality of life for patients worldwide.

Reflections

Josiah Joel

In this heart disease prediction project, I explored logistic regression, random forests, and artificial neural networks (ANNs). Logistic regression estimates the probability of an event using a formula involving coefficients and independent variables. Random forests combine multiple decision trees to enhance accuracy, with each tree making decisions based on data subsets. ANNs mimic brain neurons, computing outputs through weighted sums and activation functions.

Logistic regression's formula calculates the probability of an outcome given input variables. Random forests generate predictions by aggregating decisions from multiple trees. ANNs process information through interconnected nodes, each applying weighted sums and activation functions to produce outputs. Understanding these formulas enabled me to comprehend how these methods function and apply them effectively in predicting heart disease.

Using a backward propagating neural network to predict heart disease was like peeking into the future of healthcare. At first, it felt like diving into a deep sea of data, trying to make sense of all the numbers and patterns. But as I navigated through the layers of the network, it became clearer how each piece of information contributed to the overall prediction. It was very cool getting to use a practical technique we learnt to do on paper, using code. Although challenging, the experience was eye-opening, showing me the potential of artificial intelligence in revolutionising medical diagnosis and treatment.

Jarrod Moore

In undertaking this project, I found myself drawing extensively from the knowledge and skills acquired during my coursework in Big Data and Intelligent Systems. Big data provided me with a solid foundation in the principles of data analysis and machine learning and Intelligent Systems, while more theoretical, provided the ‘how’ and ‘why’ for the models we decided to use.

The practical knowledge from Big Data and Intelligent systems was applied in writing the code for the various models and utilising the various machine learning libraries

This project also provided me with a broader understanding of healthcare data, including electrocardiograms (ECGs). Learning about the components of ECG graphs and how to interpret them enhanced my ability to contextualise the cardiovascular datasets used in this project. This knowledge deepened my appreciation for the clinical relevance of the predictive models developed.

Overall, the integration of knowledge from the Big Data course significantly enriched my project experience, enabling me to effectively leverage advanced analytics techniques and domain-specific insights in the pursuit of predictive modelling for cardiovascular disease risk assessment. Moving forward, I am confident that the interdisciplinary skills acquired will continue to serve me well in tackling complex data-driven challenges across various domains.

Raoul Sagram

It was a life-changing experience to work on this heart disease prediction project, which provided insights into several machine learning algorithms such as random forests, logistic regression, and artificial neural networks. I have developed a deep awareness of the subtleties associated with model selection, evaluation metrics, and data preprocessing because of this attempt.

An essential component of the research was the use of random forests, logistic regression, and artificial neural networks. Developing Python code to implement these techniques gave me practical experience in creating and assessing models. By carefully comparing and analysing each strategy, we were able to determine its advantages and disadvantages. While random forests demonstrated resilience in managing nonlinear relationships and feature significance, logistic regression provided ease of use and interpretability. On the other hand, artificial neural networks need a great deal of computer power and fine-tuning to capture intricate patterns with an unprecedented degree of flexibility.

This practical experience helped me become more proficient in Python programming and helped me understand machine learning techniques on a deeper level. The coding abilities and knowledge I've learned from this project will be a great starting point for future projects, enabling me to confidently and competently take on a variety of data science tasks.

Bibliography

World Health Organization. “Cardiovascular Diseases (CVDs).” World Health Organization, World Health Organization, 11 June 2021, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

Shinde, Santosh A., and P. Raja Rajeswari. "Intelligent health risk prediction systems using machine learning: a review." International Journal of Engineering & Technology 7.3 (2018): 1019-1023.

Adler, E.D., Voors, A.A., Klein, L., Macheret, F., Braun, O.O., Urey, M.A., Zhu, W., Sama, I., Tadel, M., Campagnari, C., Greenberg, B. and Yagil, A. (2020), Improving risk prediction in heart failure using machine learning. Eur J Heart Fail, 22: 139-147

Kruppa, Jochen, Andreas Ziegler, and Inke R. König. "Risk estimation and risk prediction using machine-learning methods." *Human genetics* 131 (2012): 1639-1654.

Links

https://colab.research.google.com/drive/1Y-CpWrL091EWAtnmNuQvbyEJmaeqxRjn?usp=sharing
https://github.com/JosiahJoeking/Group_E_COMP3608_Project_Using_Data_Science_to_improve_Heart_Disease_Detection.ipynb