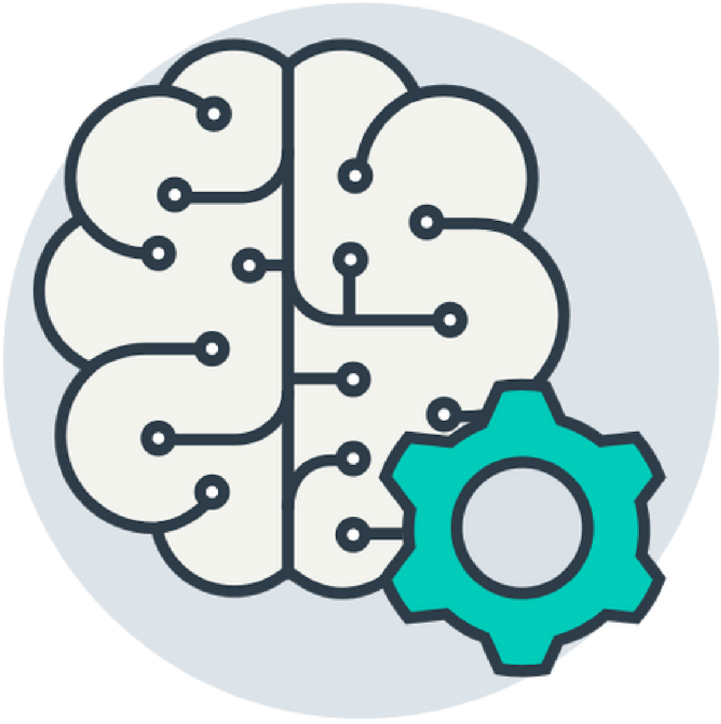
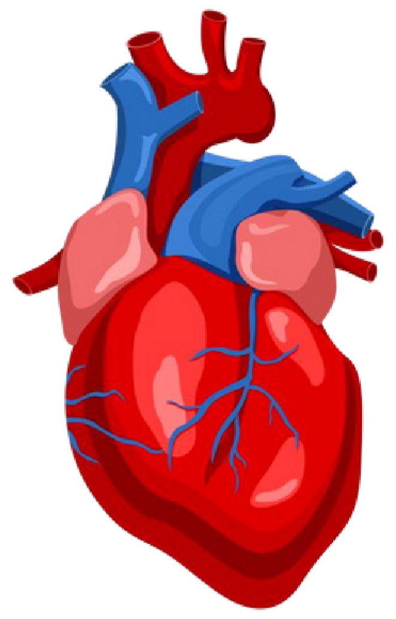
“A Prediction Model to Detect Cardiovascular Disease using Machine Learning”





Introduction

The project I will be undertaking is the creation of a prediction model for cardiovascular disease using machine learning. This poster will detail key aspects of the project.

Background

Around the UK, the current biggest cause of deaths are cardiovascular disease. Also known as heart and circulatory diseases, which affect the heart and circulation. Types of diseases include coronary heart disease (ischaemic heart disease), heart attack (myocardial infarction), abnormal heart rhythm (arrhythmia), stroke (cerebrovascular disease), in addition to many more. Approximately 7.6 million people are living with heart or circulatory diseases in the UK, with coronary heart disease being the most common type of heart and circulatory disease. Moreover, globally it’s estimated 200 million people are living with coronary heart disease, with it being the global leading cause of death in 2019.

The current uses of artificial intelligence in healthcare include; analysing X-ray images to support radiologists in making assessments, acting as remote monitoring technology for patients being cared for at home, helping clinicians read brain scans more efficiently, as well as much more. Furthermore, there are currently artificial intelligence tools being produced to diagnose heart attacks, such as the CoDE-ACS algorithm which was 99.6% accurate at diagnosing patients as tested on a set of 10,286 patients from six countries.

Problem

Worldwide more than 4 in 5 deaths from heart and circulatory diseases are associated with modifiable risk factors. Modifiable risk factors are deemed as risk factors that can be reduced with medical treatment and lifestyle changes. These include:

- High systolic blood pressure (hypertension)
- Dietary risks (poor diet)
- High LDL cholesterol (high cholesterol)
- Tobacco (cigarette smoking/second-hand smoke)
- High fasting plasma glucose (diabetes)
- High body-mass index (obesity and excess weight)
- plus, many more…

Alternate risk factors, such as air pollution, age, gender, as well as family history and ethnicity, also have an effect on the risk of heart and circulatory diseases.

When diagnosing cardiovascular disease, doctors consider various risk factors. Multiple doctors' opinions may be gathered to support the diagnosis. Long wait times for appointments can be dangerous due to the large number of at-risk patients. Individually assessing each patient can be time-consuming, especially when the diagnosis is negative. A machine learning prediction model can help doctors efficiently diagnose patients by identifying high-risk and low-risk cases in medical records. This approach allows doctors to spend less time on individual analyses and decision-making.

Methodology

Methodologies are used in software engineering provide to a systematic and structured approach to software development, helping software engineers plan and manage the creation of software through the development lifecycle. To help plan and complete this project, the Agile methodology was chosen as it best suits the flexible needs of the project. Which is advantageous, as it offers the ability for the end product to evolve as it aims to meet the requirements of the project through a constantly changing development cycle, without confining the development to specific rules.

To develop the project the following steps will taken:

1.Research and evaluate the benefits of a cardiovascular disease prediction model.

2.Find, evaluate and choose the best dataset to use for the project.

3.Utilise data pre-processing, exploratory data-analysis and feature engineering to clean, format and enhance performance of the data.

4.Analyse and find the optimal and relevant machine learning model.

5.Select relevant features to be used by the model, based upon feature importance.

6.Train and tune the chosen machine learning model using the training data.

7.Test the machine learning model using the test data, and evaluate using classification metrics.

8.Explain the results of the model and what is has predicted.

The performance of the model will be assessed based on classification metrics. The These classification metrics are:

•Accuracy – measures the number of correctly classified instances out of the total number of instances.

•Precision – measures the number of correctly predicted positives out of the total predicted positives.

•Recall – measures the number of correctly predicted positives out of all the predictions in the actual positive class.

•F1 Score – measures a balance of Precision and Recall, a high F1-score means a good Precision value and a good Recall value.

Resources

In order to complete this project, I will be using a variety of machine learning resources. The chosen tools to be used while implementing this project are:

•Python – the Python programming language being easy-to-use and simple to read is one of the reasons it is used for machine learning, as well as its extensive network of libraries containing functions to perform a variety of tasks.

•Google Colab – Google Colaboratory is a free cloud-based platform that hosts the Jupyter Notebook service which allows users to execute code cells individually, commonly used for machine learning and data science.

•Pandas – Pandas is an open-source Python library used to manipulate and analyse data. Pandas functions are commonly used for data cleaning, transformation, analysis and visualisation within machine learning.

•Scikit-learn – Scikit-learn is an open-source Python library used for data analysis and modelling. It contains tools for data preprocessing and machine learning algorithms such as classification, regression, and clustering, used within machine learning.

•TensorFlow – TensorFlow is an open-source machine learning library for developing and training deep neural networks.

These tools were chosen due to their prior experience and ease of use.

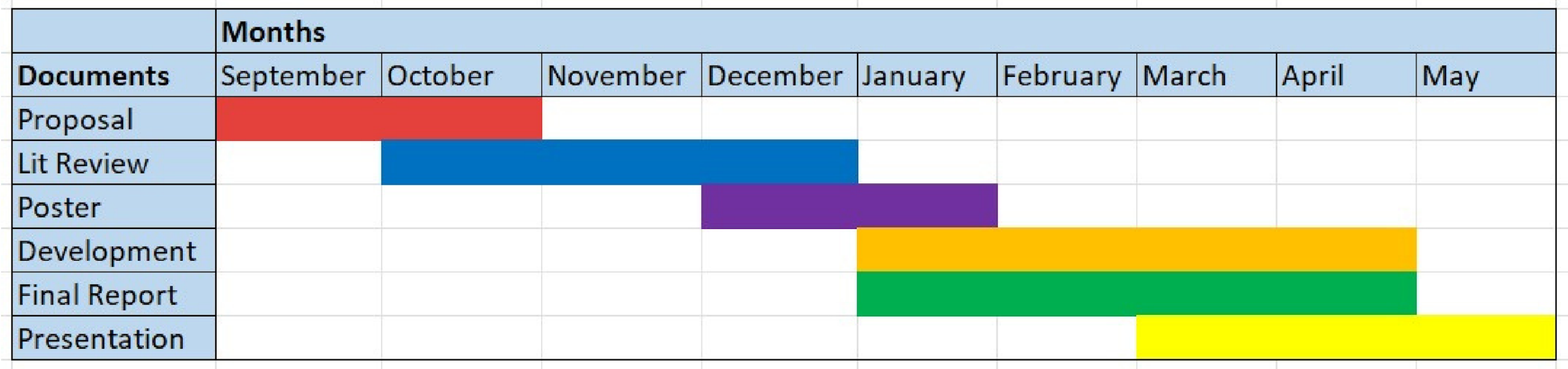
Ethical Issues

One key ethical issue for the project is data privacy when it comes to using patients’ personal health data to train the machine learning model. To resolve this issue, development of the project will cohere to standard data practices, such as anonymizing data where possible to train and test the machine learning model.

Another key ethical issue for the project is explainability, as it can be difficult to determine how a machine learning model comes to a certain prediction, acting as a “black box”. To solve this issue, the project will aim to prioritize explainability in human terms to increase trust.

One other important ethical issue for the project is bias within the training data, which can lead to skewed predictions towards certain groups of people. To fix this issue, the dataset will be assessed to resolve any biases that can negatively affect the prediction outcome.

Time Plan



	Months								
Documents	September	October	November	December	January	February	March	April	May
Proposal	🟥	🟥							
Lit Review		🟦	🟦	🟦					
Poster				🟪	🟪				
Development					🟧	🟧	🟧	🟧	
Final Report					🟩	🟩	🟩	🟩	
Presentation							🟨	🟨	🟨