Business and Local Government Data Research Centre

MATHEMATICAL SCIENCES

Project Title: Cardiovascular Disease Prediction

Organisation: Mid and South Essex Health and Care Partnership

Data Science and AI Masters Project 2022

Summary Report

EXPLORING DATA
ENHANCING KNOWLEDGE
EMPOWERING SOCIETY

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

University of Essex

Author: Adekemi Kadri

Academic Supervisor: Giagos, Vasileios

Date: 23/12/22

With thanks to the Senior Research Officers at the Business and Local Government Data Research Centre.

Table of Contents

1 Introduction

The Data Science and AI Masters Project scheme is a partnership enabled through the Business and Local Government Data Research Centre.

This opportunity allows businesses, public and third sector organisations to work with the University of Essex on a joint, data-related project that would involve the use of advanced data analytics.

The University of Essex is one of the top Universities in the world for data science. And now, thanks to the Business and Local Government Data Research Centre and the Department of Mathematical Sciences (DMS), organisations can benefit from the expertise of the postgraduate students at the forefront of data science and artificial intelligence funded by the Centre to bring the project to life and maximise the value of data as they undertake real-world challenges in an exciting 10 week project.

Moving beyond on the Data Science and AI Masters Projects

The Data Science and AI Masters Project scheme is just one way organisations can benefit from grant funded data analytics and research support.

To find out more about the training, workshops and wider collaborative research opportunities please visit: https://www.essex.ac.uk/centres-and-institutes/business-and-local-government-data

To find out more about longer term projects involving students including placements and recruitment please visit: https://www.essex.ac.uk/business/recruit-our-students-and-graduates/internships and https://www.essex.ac.uk/business/recruit-our-students-and-graduates

To find out more about opportunities to collaborate with wider University of Essex academics through Knowledge Transfer Partnerships (KTPs) please visit https://www.essex.ac.uk/business/expertise/knowledge-transfer-partnerships

2 Background

According to NHS Cardiovascular disease (CVD) can be defined as any disorder that affects the heart or blood vessels. All heart and circulatory conditions, such as congenital heart conditions, strokes, coronary artery disease, vascular dementia, hypertension, heart attacks, and angina, aortic disease are categorized as cardiovascular diseases (CVD) [1]. Cardiovascular diseases include stroke and heart disease. The heart and all blood veins that pump and circulate blood throughout the body make up the cardiovascular system.

As per the details found in [2], the most frequent cause of early death in the UK is Coronary Heart Disease. One in four men and one in six women die from it, and in 2000, it was responsible for around 125 000 fatalities. Each year, there are about 274,000 myocardial infractions. Only 1% of the annual £1.6 billion cost of CHD to the National Health Service is allocated to primary prevention. CHD has an annual economic impact on the UK of almost £10 billion. In the UK, heart disease claims the lives of at least 460 individuals every day and 170,000 people annually, this can be prevented by early detection through the Understanding of the risk factors (high cholesterol, too much alcohol consumption, obesity, smoking, high blood pressure and physical inactivity).

Heart disease is one of the top causes of death in the UK [3], hence it is vital to implement various strategies to decrease the number of deaths caused by this condition. Utilizing data analytic and a data science approach is one of the best methods to accomplish this. This research is crucial in assisting Gp's and CCG's in understanding the population at risk of CVD as early as possible, hence reducing the death rate. A thorough data analysis and creating machine learning model will play a significant role in reducing the deaths rate of heart disease.

3 Research Objectives

- To understand the current population in Mid and South Essex vulnerable to have cardiovascular disease
- To identify the risk factors related to a stroke and a heart attack that could be used as predictors
- To predict the number of people in Mid and South Essex likely to have a heart attack and the number of people likely to have a stroke in the next 5 years and where they are coming from
- To identify any health inequalities for our population with cardiovascular disease and its complications

4 Dataset

I. **2021 Dataset**: This spreadsheet contains 95 columns including detailed information for the year 2021 for five NHS CCGs (NHS Castle Point & Rochford CCG, NHS Southend CCG, NHS Basildon and Brentwood CCG, NHS Thurrock CCG, NHS Mid Essex CCG) and 150 Gps. It includes information on people of all ages measured in years, from 0 to 95+, as well as columns for the population of men and women in each general practitioner. It also includes information on risk factors for cardiovascular disease, such as the number and percentage of people with chronic kidney disease, diabetes, hypertension, smoking prevalence, patients with Coronary Heart Disease (CHD), patients with heart failure, patients over 45 with a history of blood pressure in the past five years, smokers who have received support and treatment in the past, and the smoking status of patients with specific conditions. It also contains information about the percentage of people registered in each GP that are unemployed, percentage of people who have a positive experience of their GP practice, percentage of people reporting good overall experience of making an appointment, percentage of people satisfied with practice appointment times, percentage of people with a long-standing health condition, it contains the number and percentage of patients who checked their last blood pressure that are 80+ years, with CHD in last 12months with a result <=150/90, the number and percentage of patients who checked their last blood pressure that are less than 80+ years and are with CHD in last 12 months with a result <=150/90.

   For ease of analysis and to reduce the number of columns because there were so many in the dataset, the age groups that were broken down for males and females were combined into single columns. Additionally, one of the data wrangling processes involves making untidy data tidy. In this sheet, the untidiness was that the age groups were all in separate columns as well as the gender columns. Before building the model, the dataset was tidied by combining all age groups into a single column and the gender groups into a column as well.

II. **Heart Trend Data**: With 25 columns and 8788 rows, the heart trend data spreadsheet contains information about heart disease from 2009 to 2020. The columns comprise: Indicator Id (Id for the dataset), Indicator name (a categorical column that comprises of cardiovascular disease types (Coronary Heart Disease and Heart failure) with many empty columns that were all deleted before analysis.

III. **Risk Trend Data**: This spreadsheet includes columns that are comparable to those in the data on heart trends and contains information about risk factors for cardiovascular disease which include hypertension, atrial fibrillation,

smoking, heart failure, chronic kidney disease, diabetes, deprivation score and some other details like patients with CHD that have been immunized against flu, last BP reading of patients (<80yrs, with CHD) in last 12months that is <=140/90, last BP reading of patients (over 80 years, with CHD) in last 12months that is <=150/90, CHD patients that has taken aspirin, APT or ACT, heart failure w LVSD: treated with ACE-I or ARB, record of offer of support and treatment for smokers (15+, last 24 months), new hypertension patients, age 30 to 74, these were data from 2009 to 2020. There are 23716 rows and 25 columns in this spreadsheet. Because it contains the same information as these two other spreadsheets, this data was used instead of the smoke trend data and heart data. Most of the columns were deleted due to their emptiness and lack of relevance. For example, the sex column, which did not contain any specific gender information but only contained "all persons" for all rows, was removed because it was irrelevant. The age column, which had only one value ('all ages') for all rows, was similar and was also removed before analysis.

IV. **Cardiovascular Disease 1920**: This spreadsheet, which has 152 rows and 32 columns, contains information on heart failure and stroke for 152 general practitioners. This spreadsheet contains info like patients with Atrial Fibrillation, population of patients, percentage of coronary heart disease patients 80 years of age or older whose most recent blood pressure reading was 150/90 mmHg or below (taken in the 12 months prior), number of patients 45 years of age or older with a blood pressure record from the previous five years, percentage of patients who have had a blood pressure reading in the five years prior and are 45 years of age or older, number of patients with CHD (Coronary Heart Disease), percentage of patients with coronary heart disease who have used aspirin, an alternate anti-platelet medication, or an anticoagulant in the past 12 months, patients with CHD, stroke and heart failure and other indicators.

The data in this spreadsheet is similar to the risk data, with the exception that it only covers 2019 and 2020, so no analysis was done on it.

V. **Smoking Trend Data**: Similar to risk trend data and heart data, the smoke trend spreadsheet includes columns that are comparable to those in the data on heart trend and risk data and this spreadsheet contains information about smokers from 2009 to 2020.

VI. **Atrial Fibrillation**: Atrial fibrillation is a heart condition that causes an irregular and often abnormally fast heart rate. This spreadsheet primarily contains data on patients who have atrial fibrillation, including numbers of patients and the percentage of patients who have the condition in relation to GPs and CCGS. Dates range from 2009 to 2021 for the data. This spreadsheet was not used as

the details of patients with atrial fibrillation is already in the risk trend spreadsheet.

VII. **Definition of terms:** This spreadsheet contained definition of terms in other six spreadsheets for better understanding of the entire datasets.

The 2021 spreadsheet and the risk trend data were the only two of these six spreadsheets that were used in this study. The risk trend data was used because it contained the key information about cardiovascular disease, such as the type of cardiovascular disease and risk factors of cardiovascular disease that were dated from 2009 to 2021. The 2021 data was used primarily because of its age group and gender data in order to understand if there is any relationship between age and cardiovascular disease or relationship between gender and cardiovascular disease.

5 Methods/Methodology

- **Preprocessing of the Dataset:** Both Excel spreadsheets and a Python environment were used to preprocess the data spreadsheets (using google colab). In the 2021 spreadsheet, the age groups that were initially split into male and female brackets were combined into single cells, which means the age groups for male and female were added to create a single cell. This was done in an excel sheet for ease of data exploration and analysis since the spreadsheet initially had more than 80 columns, the split age groups were deleted after the combinations. Additionally, in order to prepare the data, all of the excessively long column names were renamed before being exported into the Python environment for better analysis and visualization. There were just 2 rows with missing values and they were deleted before using in machine learning models

  Other sheets with similar columns (risk trend data, smoking data, and heart data) feature columns that are fully empty were deleted before exporting into python environment. Additionally, the names of some columns' category values were changed to match their shorter, easier-to-understand titles, and the names of some columns' columns were changed to reflect the shorter, more understandable descriptions that were previously included in their names.
  The risk data's indicator categorical column, which includes the various forms of cardiovascular disease and their associated risk factors, was split into separate columns and saved in various data frames. These data frames were later combined again for analysis and input into machine learning models.

- **Data Distribution (Skewness and Kurtosis):** Skewness is a measure of asymmetry of a distribution of a given data that can be interpreted as "skewed to the left" when the skewness value is negative or "skewed to the right" when the

skewness value is positive and can also be zero. When a distribution is highly skewed, its skewness value is less than -1 or greater than +1, whereas a moderately skewed distribution has values between -1 and -0.5 or between +0.5 and +1. A normally distributed distribution has a value range of -0.5 to +0.5. Over 90% of the data in this spreadsheet from the 2021 dataset is skewed, with values greater than 1 and outliers were also observed. As shown in Figure 1,2 and 3, the data on patients with coronary heart disease, stroke and heart failure are positively skewed.

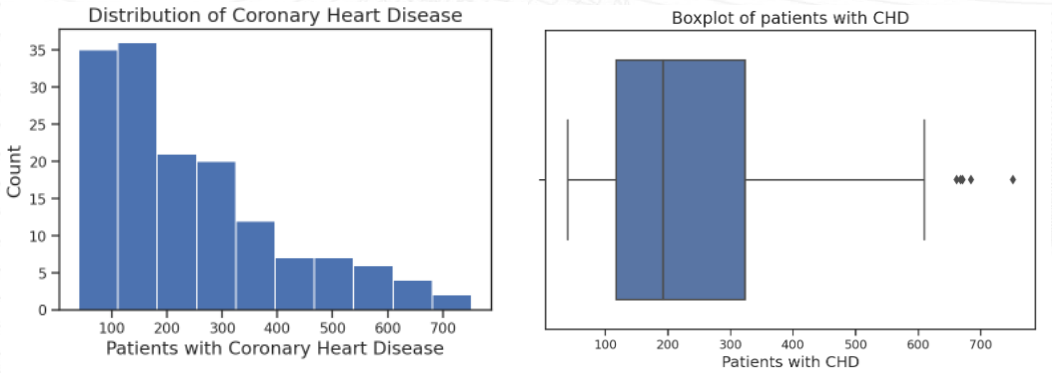

Figure 1: Histogram and boxplot of patients with coronary heart disease

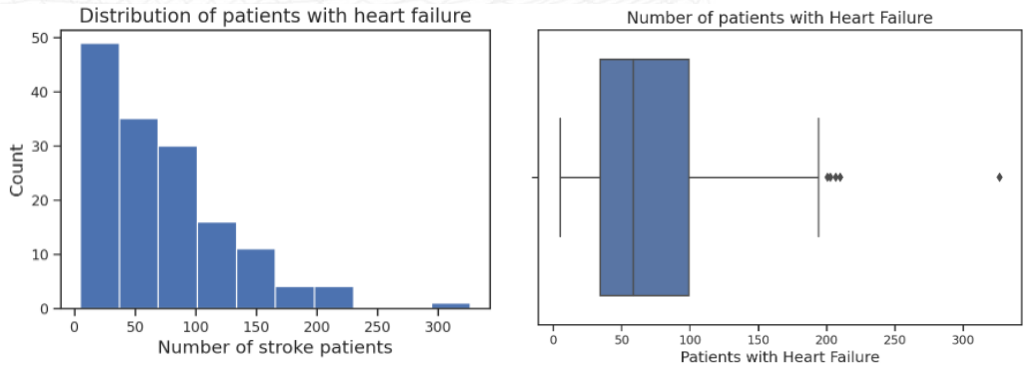

Figure 2: Histogram and boxplot of patients with heart failure

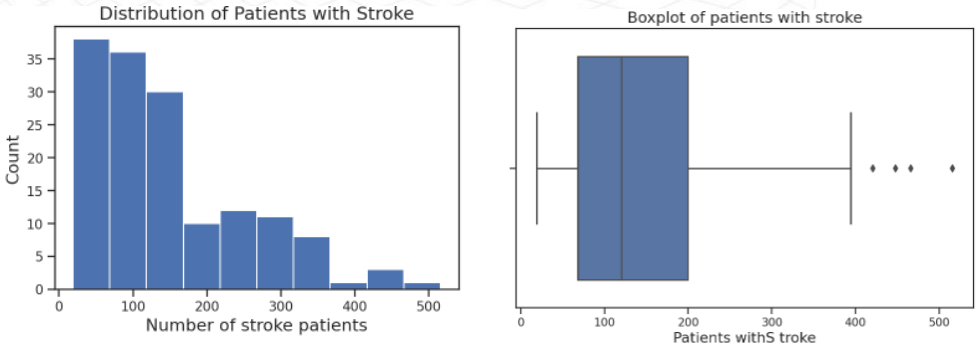

Figure 3: Histogram and boxplot of patients with stroke

- **Summary Statistics**: On the datasets, summary statistics were performed, and the results included the mean, the standard deviation, the interquartile ranges,

the minimum and maximum values in each column as well as other statistics. One finding from the statistical summary was that there were numerous columns with extremely high standard deviations. When these columns were examined closely, skewness and outliers were found in these columns. Additionally, there are a maximum of 327 patients with heart failure across all GPs, with a minimum of 5 patients. While the minimum and maximum numbers of CHD patients are 40 and 752, respectively.

Feature Selection: Feature selection is a machine learning technique that is used to choose relevant features or variables. It seeks to remove redundant or unnecessary features or characteristics that are strongly associated in the data without much information loss. It is frequently used to greatly simplify model interpretation and boost generalisation by lowering variance.

The 2021 spreadsheet has many columns, so it is crucial to use a feature selection model to identify the columns that are most and least relevant before using them in machine learning models to predict and forecast. For this reason, lasso regression model was used for the feature selection.

Lasso Regression: Regression models are a popular model for predicting the risk of a foreseeable result and are frequently employed in statistical analyses [4]. Although there are several regression techniques that may be used to solve these issues, applying traditional regression methods to a set of candidate variables to produce a model frequently results in overfitting in terms of the number of variables that are finally included in the model. As it addresses overfitting and bias, the shrinkage and variable selection method for regression models known as LASSO (Least Absolute Shrinkage and Selection Operator) regression is a desirable choice. As a result of applying the lasso regression model to the data, it was possible to identify the variables/features that are most and least useful for predicting cardiovascular disease. These features are shown in Table 1, 2, 3 for heart failure, coronary heart disease and stroke respectively in the appendix.

Lasso regression was used because it is a perfect model when there is multicollinearity in the columns of the dataset, which can be inferred from the fact that over 90% of the features have very strong correlations when the features of the dataset are checked for correlation.

- **Machine Learning Models**: In this project, three machine learning models were applied: decision tree classifier, linear regression, and FB Prophet time series forecast. The columns with the number of patients with CHD, stroke, and heart failure were employed independently as the target variables in all the machine learning models implemented, and various predictions were made for each of them.

The FB project time series forecast was chosen due to its simplicity and accuracy in making future predictions with only two variables, while the linear regression and decision tree regression models were used because they are best for multiple features with continuous variable predictions (the year column and y, i.e., column to be predicted).

I. **Linear Regression**: Linear regression is a simple and widely used machine learning algorithm for performing predictive analysis; it supports continuous or mathematical variable projections. There are two types of linear regression: Simple linear regression and multiple linear regression. Simple linear regression is used when there is just one predictor (an independent variable) and one regressor (a dependent variable) while multiple linear regression is used when there are several predictors or regressors, the multiple linear regression model is the regression technique that is employed for this project and the result. Table 4 in the appendix contains the results of the linear regression model, which ranks the features/variables in terms of importance from most important to least relevant for predicting cardiovascular disease. Table 4 in the appendix contains the metrics for the model's accuracy.

II. **Decision Tree Regression**: Decision tree regression model is a model in the shape of a tree structure. It gradually creates an associated decision tree while segmenting a dataset into ever-smaller subsets. The result is a tree with decision nodes and leaf nodes [5]. A decision tree classifier variant called decision tree regression can be used to approximate real-valued functions like class proportions. Binary recursive partitioning, an iterative process that divides the data into partitions, is the foundation upon which a regression tree is built. To start, the structure of the tree is established using all the training samples. The algorithm then splits the data into all possible binary splits and chooses the split that divides the data into two parts while minimising the sum of the squared deviations from the mean in each part.

III. **FB Prophet Time series forecast**: Time series forecast involves forecasts that are generated for data when the exact outcome may not be known until some future date, but all the earlier observations or features are regarded identically Time-series forecasting involves using previous data to produce forecasts in the future, it has specific methodology for making the predictions and forecasting future events use the date history of the data. FB Prophet was chosen above other time series forecast models because of its ease of use, propensity to handle outliers correctly, and robustness to missing data and trend shifts. This model analyses the trend and pattern from previous years to generate the forecast for future years, considering just the value to be forecasted and the year. In this instance, forecasts for each predictor were created using the numbers of patients with coronary heart disease, heart failure, and stroke, along with time range values. The forecast made for each predictor can be seen in *figure 14* and the metrics for the model can be seen in table 6 in the appendix.

Metrics for machine learning models: The performance of machine learning models is typically measured by different metrics given on a scale of 0 to 1, which can also be calculated in percentage by multiplying the value by 100; the closer the value is to 1, the better the model's performance, and the metrics tell how well the model approximates the relationship; the following metrics were used to measure the performance of the models.

a. Mean squared error
b. R-squared
c. F- statistics
d. Mean absolute error
e. Root Mean Square error

The accuracy measures for each model had high values, indicating that the models were effective in making these predictions. In the appendix's table 6, the outcomes of these measures for each model are displayed.

6 Analysis and Findings

The analysis of the datasets produced the following insights:

- The Gp's from NHS Southend and Mid Essex have the highest number of patients with cardiovascular disease as displayed in *figures 4 and 5*. This is not surprising considering that Southend and Rochford CCG were combined to form one CCG, which has the highest population and the following GPs have been in the top 5 of high record of cardiovascular disease over the years.

  As illustrated in *figure 6*, the same GPs with a high prevalence of coronary heart disease also have a high prevalence of stroke and heart failure.

GP Name	CCG	Number of CHD Patients	Number of Stroke Patients	Population
Chelmer Medical Partnership	NHS Mid Essex	752	516	28135
Beacon Health Group	NHS Mid Essex	752	466	25599
Audley Mills Surgery	NHS Southend	671	448	20685
Dr Puzey,Dr Kothari and Dr Nanda	NHS Southend	668	355	20774
Audley Mills Surgery	NHS Southend	671	448	20685

Fig 4: Top five GPs with high number of CHD patients and their population

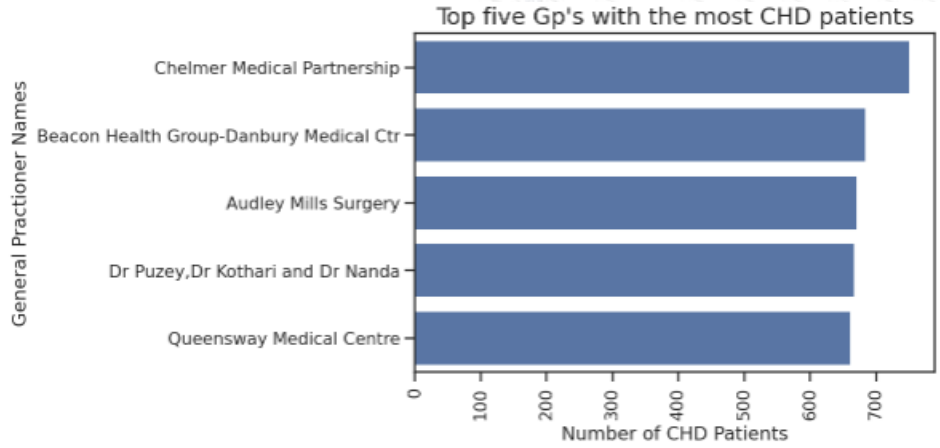

Fig 5: Top five GPs with high number of CHD patients

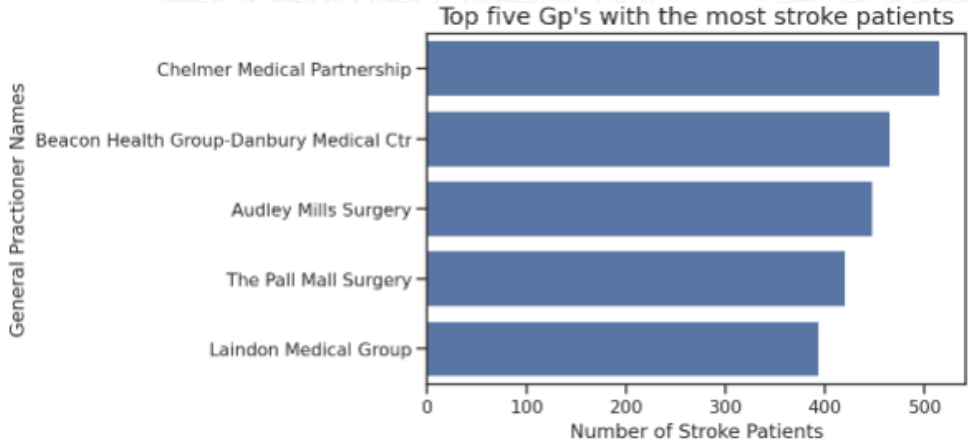

Fig 6: Top five GPs with high number of patients with stroke

- *Figure 7* depicts the proportion of CHD patients in each CCG; as was already indicated, NHS Southend has the highest proportion with a percentage of 33.8% followed by NHS Mid Essex CCG with 32.7%.

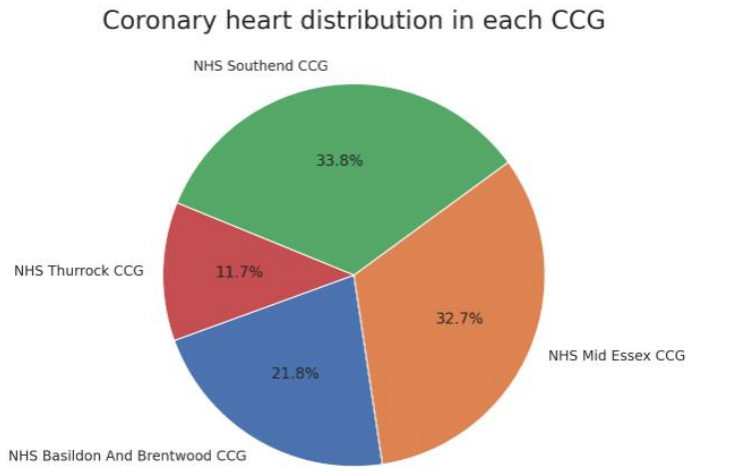

Fig 7: Proportion of CHD patients across all CCGs

- After the analysis, it was discovered that over the years, heart disease patients have been identified as being most prevalent in mid Essex CCGs, followed by stroke and heart failure as ill as shown in *figures 8 and 9*.

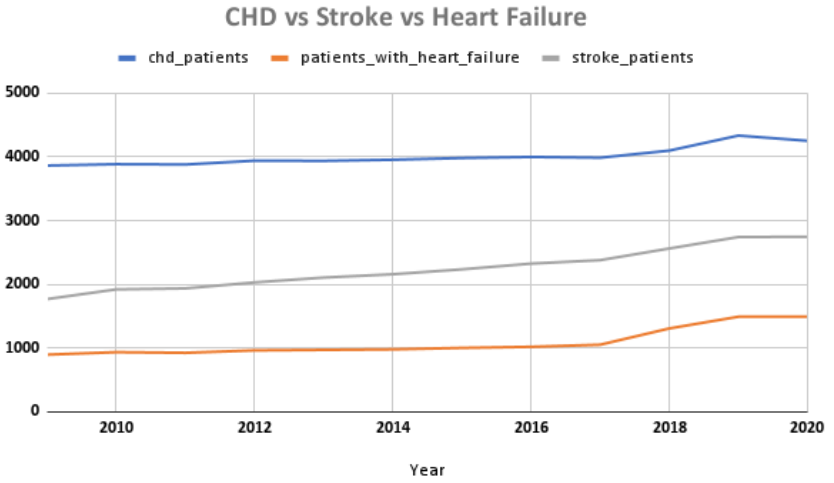

Fig 8: Trend of CHD, Heart Failure and Stroke over the years

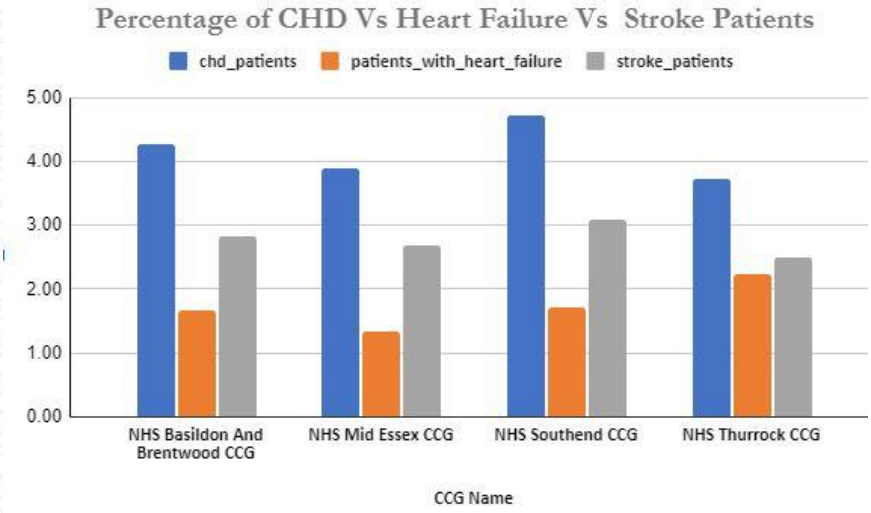

Fig 9: Histogram plot of Trend of CHD, Heart Failure and Stroke yearly

- A notable observation from the data analysis was that when the count of cardiovascular disease is analyzed, there has been an increase in the number of patients over time (see *figure 12*), but when the percentage value is used, there has been a decrease in the percentage of patients with cardiovascular disease over time (see *figures 10 & 11*). It was noticed that the percentage of patients with CHD has been declining over the years across all GPs, this trend was also noticed for of percentage of patients with stroke and heart failure. Figure

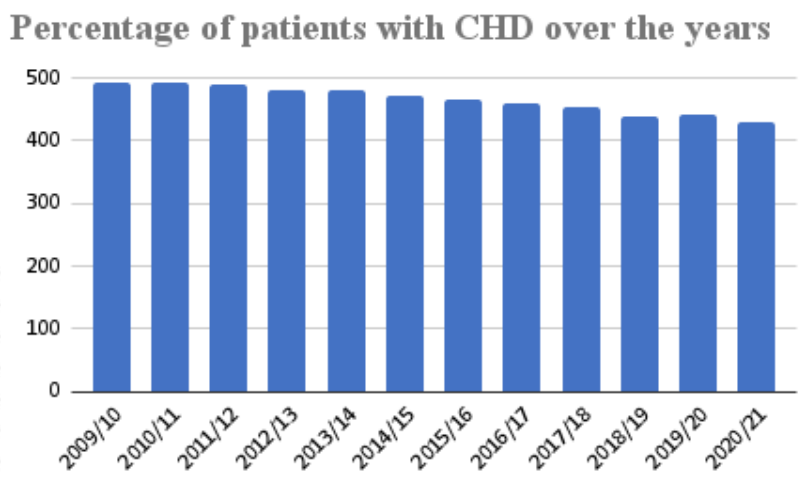

Fig 10: Histogram plot of trend of percentage of CHD patients

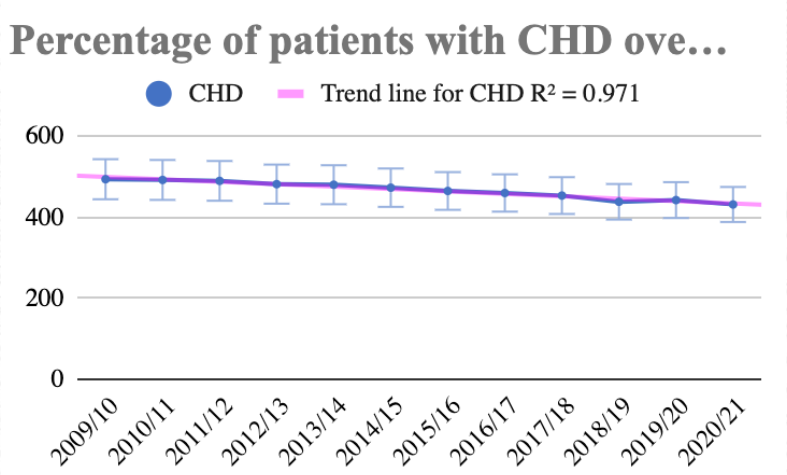

Fig 11: Trend of percentage of CHD patients since 2009

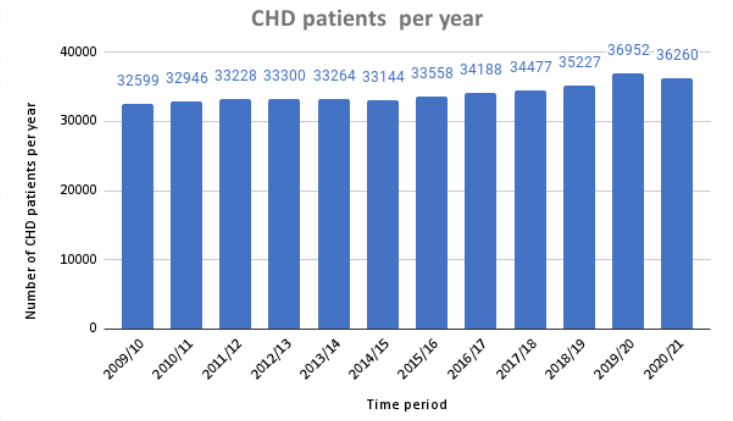

Fig 12: Histogram plot of trend of number of CHD patients yearly since 2009

- All the features in the dataset exhibit strong positive correlation. The correlation heat map in figure 13, which uses count values rather than percentages, reveals a strong relationship between all age groups and gender with CHD, suggesting that anyone inside any age group and either a female or a male can have cardiovascular disease. In order to be sure of whether to use the age group in the machine learning models, a correlation heat map was done for age groups and each of the cardiovascular diseases and the heat map reveals that the age groups with the strongest link have correlation values of 0.95 for each of the following: 60 to 64, 65 to 69, 70 to 74, and 80 to 84, see *figure 14*. Kindly zoom in on the pdf file or view a sharper version of the heat map below in the image that is attached to the report.

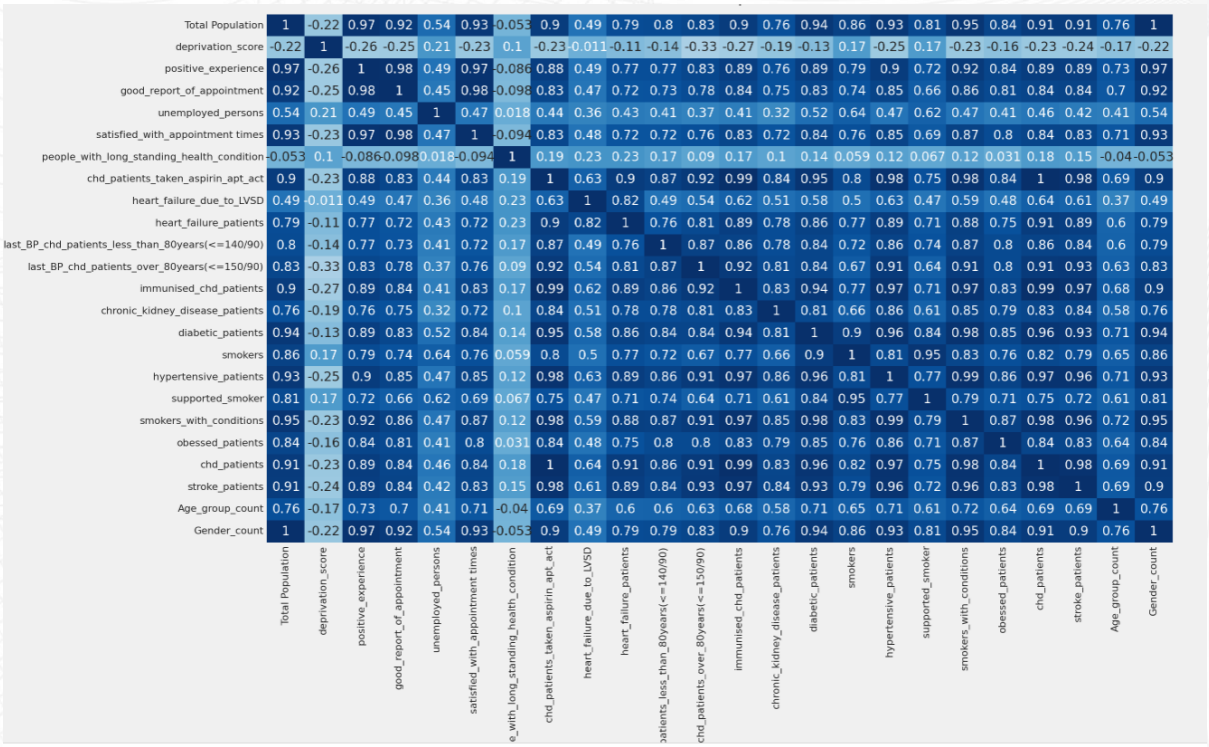

Fig 13: Correlation between all features in the 2021 dataset

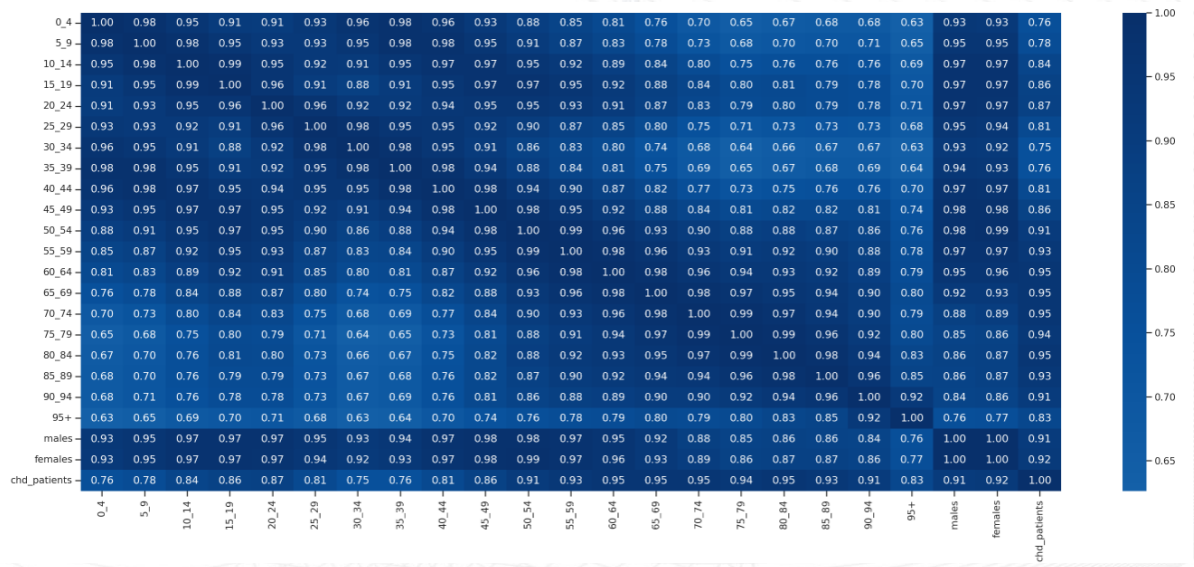

Fig 14: Correlation between all age groups and CHD

7 Machine Learning Model Results

- The time series forecast, which was based on the trend of the count and percentage of cardiovascular disease patients since 2009, predicted that the trend would continue for the following five years, with a declining trend when the percentage value is used but with a continuous increase when the count value is used. These predictions can be seen in *figures 15 and 16*.

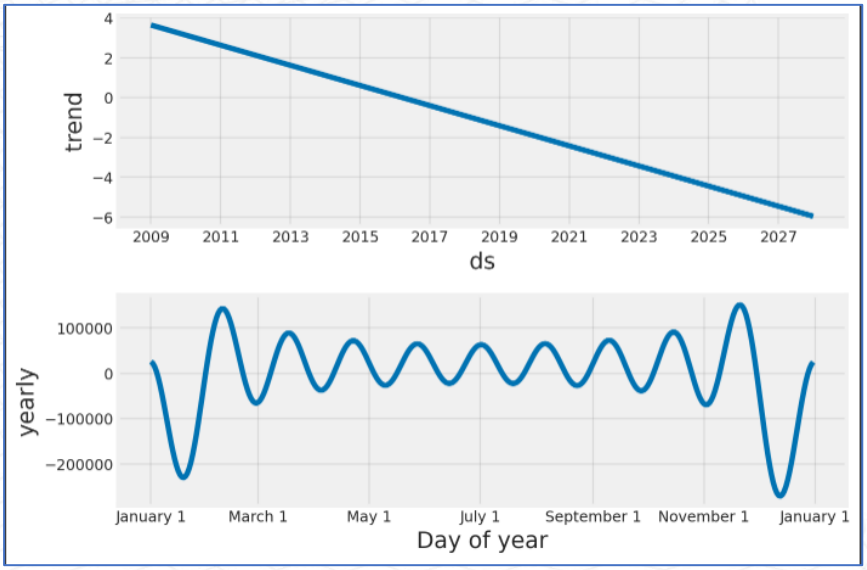

Fig 15: Heart failure and stroke forecast in next five years (Percentage)

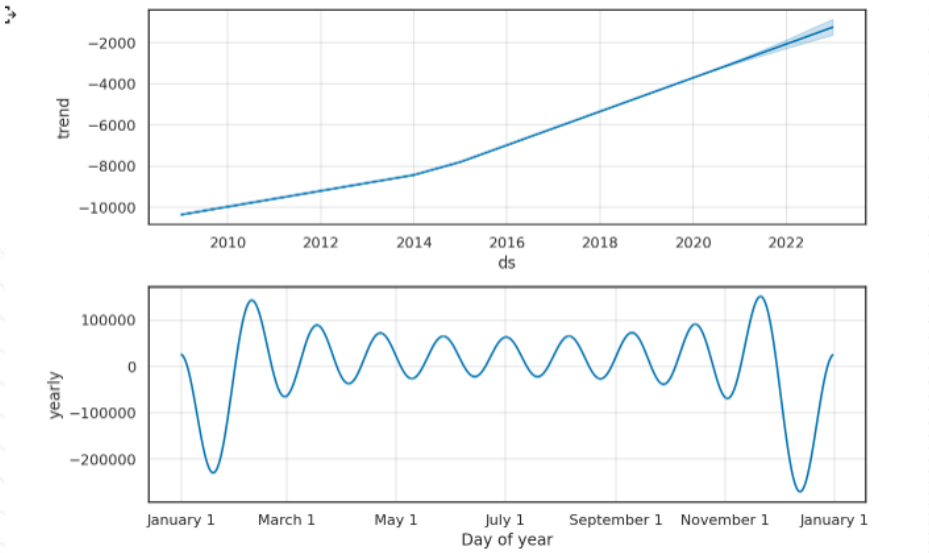

Fig 16: Heart failure and stroke forecast in next five years(count)

- For the linear regression model, the following features were ranked top features in predicting coronary heart disease
  i. Chronic kidney disease
  ii. Heart failure due to LVSD
  iii. Immunized CHD patients
  iv. CHD patients with record of aspirin taken
  v. Hypertensive patients
  vi. Obesity
  vii. Diabetes
  viii. Stroke
  ix. People with long standing health conditions
  x. Smoking
  xi. Unemployment
  xii. Patients with heart failure

  The following features were ranked top features in predicting stroke

  i. Obesity
  ii. People with long standing health conditions
  iii. Diabetes
  iv. Smoking
  v. Patients with heart failure
  vi. Hypertensive patients
  vii. Chronic heart disease

The following features were ranked top features in predicting heart failure

i. Obesity
ii. Diabetes
iii. Unemployed patients
iv. People with long standing health conditions
v. Diabetes
vi. Smoking
vii. Stroke
viii. Hypertensive patients
ix. Chronic heart disease
x. Heart failure due to LVSD

8 Conclusions and Next Steps

The model suggests that people from Basildon & Brentwood, Mid Essex, and NHS Thurrock CCG are at high risk of stroke compared to Southend.

- The social factor that can affect the prediction of stroke are; unemployment, positive experience at GP, good report of appointment at GP, satisfaction with appointment times
- The age groups that are important in predicting stroke include 0 to 4, 30 t0 44, 50 to 54, 5 to 9, 65 to 69, 70 to 74, 85 to 89, and ages over 95. Despite the fact that it is not common for children aged 0 to 9 to develop stroke, some research claims that any person of any age can develop stroke.
- Females are more related to stroke compared to men according to the feature selection with lasso and linear regression mode
- The feature selection of NHS Southend as the least region to expect stroke patients, despite the fact that this is the most populated CCG in the data, is an intriguing find.
- According to the machine learning models, high number of heart failure can be expected any of the CCG
- Analyses have revealed that approximately 3.5 percent of MID Essex will be at risk of cardiovascular disease in the coming year, with an annual increase of 0.01 – 0.02 % and the CCG to expect maximum percentage of cardiovascular patients from is NHS Southend due to its population.

References

[1] https://www.england.nhs.uk/ourwork/clinical-policy/cvd/#:~:text=Cardiovascular%20disease%20(CVD)%20is%20a,hypertension,%20stroke%20and%20vascular%20dementia. Accessed on December 15, 2022.

[2] Poulter, N (2003). Global Risk of Cardiovascular Disease. *Heart*. 89 (Suppl II). p. ii2-ii5.

[3] Kingston, H (2021). Top 5 Causes of Deaths in the UK. *LetsGetChecked*. Feb 15. https://www.letsgetchecked.com/articles/top-5-causes-of-death-in-the-uk/, accessed on December 15, 2022.

[4]Ranstam, J and Cook, J A (2018). LASSO regression. *British Journal of Surgery*. 105(10). p. 1348, https://doi.org/10.1002/bjs.10895

[5] Xu, M, Watanachaturaporn, P, Varshney, P K and Arora, M K (2005). Decision Tree Regression for Soft Classification of Remote Sensing *Data. Remote Sensing of Environment*. 97(3). p. 322-336. https://doi.org/10.1016/j.rse.2005.05.008.

Appendix – A: Results of Models used for Prediction

Table 1: Lasso regression results on most relevant and least relevant features in predicting coronary heart disease

Most Relevant Features	Least Features
Deprivation score	Positive experience at GP
Good report of appointment	Hypertensive patients
Unemployed people	Stroke patients
People with long standing health conditions	Gender
CHD patients that have taken aspirin	NHS Basildon & Brentwood CCG
Patients with heart failure	Age groups (5-44, 55-59, 65-74, over 80)
Smokers	Females
Patients with chronic kidney disease	Males
Patients with diabetes	
Patients with obesity	
NHS Mid Essex CCG	
NHS Southend CCG	
NHS Thurrock	
Age groups (0-4, 45-49, 50-54, 60-64, 75-79)	

Table 2: Lasso regression results on most relevant and least relevant features in predicting stroke

Most Relevant Features	Least Features
Positive experience at GP	Deprivation score
Good report of appointment	CHD Patients
Unemployed people	NHS Southend CCG
People with long standing health conditions	Age groups (10-29, 35-39, 45-49, 55-59, 60-64, 75-79, 80-84, 90-94)
CHD patients that have taken aspirin	Males
Patients with heart failure	
Smokers	
Patients with chronic kidney disease	
Patients with diabetes	
Patients with obesity	
Hypertensive patients	
NHS Basildon & Brentwood CCG	
NHS Mid Essex CCG	
NHS Thurrock CCG	
Age groups (0-9, 30-34, 40-44, 50-54, 65-74, 85-89, 95+)	
Females	

Table 3: Lasso regression results on most relevant and least relevant features in predicting stroke

Most Relevant Features	Least Features
Positive experience at GP	Deprivation score
Good report of appointment	CHD Patients
Unemployed people	NHS Southend CCG
People with long standing health conditions	Age groups (10-29, 35-39, 45-49, 55-59, 60-64, 75-79, 80-84, 90-94)
CHD patients that have taken aspirin	Males
Patients with heart failure	
Smokers	
Patients with chronic kidney disease	
Patients with diabetes	
Patients with obesity	
Hypertensive patients	
NHS Basildon & Brentwood CCG	
NHS Mid Essex CCG	
NHS Thurrock CCG	
Age groups (0-9, 30-34, 40-44, 50-54, 65-74, 85-89, 95+)	
Females	

Table 4: Linear regression metrics

Metrics measured	Metric value
Mean squared error	0.86
Mean absolute error	0.067
R- squared	1.00

Table 5: Decision tree metrics

Metrics measured	Metric value
Mean squared error	0.99
Mean absolute error	0.07
R- squared	1.00

Table 6: FB prophet metrics

Metrics measured	Metric value
Mean squared error	0.75
R- squared	1.00

Business and Local Government Data Research Centre
Parkside 2C
Knowledge Gateway
University of Essex
Wivenhoe Park
Colchester CO4 3SQ

T: 01206 873859
E: BLGDataResearch@essex.ac.uk
W: www.BLGdataresearch.org
@BLGDataResearch