ABSTRACT

This report explores the dataset 'Heart Conditions Data' to understand the different risk factors for heart attacks. Heart attacks or myocardial infarctions occur when the blood supply to the heart through the carotid artery is blocked causing heart muscle damage. There are several contributory factors, but our report will focus on the risk factors presented as features on our dataset.

Heart attacks remain one of the major causes of death with an NHS report stating that 25% of deaths across all demographics are caused by heart attacks. This is a serious cause of concern reason why many research studies have been carried out and are still ongoing to fully understand this disease and hopefully offer remedies. While our dataset is synthetic, it mirrors possible real-life data as it attempts to present data which has as features true causes, as seen from other real-world studies, of heart attacks.

The data contains 14 features and 303 instances with the output feature being our target variable. This variable has as outcomes; '1' which refers to more likelihood of having a heart attack and '0' which is the patient is less likely to have a heart attack. The features were reduced by dropping some which had little or no relationship with the target and which may add noise to the data thus making the outcomes of our training model faulty or less accurate.

The model for training of this data was a supervised learning model under classification. The logistic regressor was chosen because the outcome of our training model was going to be binary and categorical and since our data did not have good linear correlation with the target, a classifier in this case is better to train the model.

The results showed the model could predict the outcomes to a high degree. Using a test size of 20% and a random state of 2, the model had an accuracy of 90% which did not change with the tweaking of the parameters up or down. This, while not as high as I would have liked, showed the model was good enough to predict outcomes.

While comparing these results to other benchmark studies on this same topic, I concluded that the risk factors as presented on the data under exploration did indeed match those of some very high-quality studies which also found that risk factors like diabetes, hypertension, high cholesterol etc where indeed risks for heart attacks. There is need for the results of such findings to be shared with the public and measures taken

to ensure that all is being done to reduce the occurrence of heart attacks caused by preventable causes.

Keywords: *Data, Model, logistic regression, Heart attack, Dataset, confusion matrix, research*.

PROBLEM STATEMENT

Heart attacks also known as myocardial infarctions refers to damage to the heart muscle resulting from decreased supply of blood to the heart generally caused by blockages in the coronary artery which is the main vessel supplying the heart.

Myocardial infarctions remain a leading cause of death in the world across all demographics. According to the British Heart Foundation, 25% of all deaths in the Uk are caused by this disease. This is a major public health issue which requires continuous research to ensure the main causes of this problem are fully understood and as many preventable deaths from heart attacks stopped.

This report will aim to answer the question; What conditions, across sex, age and particular health conditions expose or predispose people to having a heart attack.

Objectives and Expected outcomes

The data will be analysed to bring out trends and relationships between the causes and frequency of this condition across the features in the dataset. It will also evaluate how the factors, seen as features in the dataset, increase or decrease the likelihood of having a myocardial infarction.

Methodology

This will be done by exploring the dataset, 'Heart Condition Data' which presents synthetic data that mirrors real world data as can be found on health databases. Statistical tools will be employed to understand the trends in this data. The programming languages used will be python to get a deeper understanding of the data.

Hopefully the report will serve as an informative resource that can be used by both the public and policy makers to understand more about this condition.

DATASET

```
[2]: dataset=pd.read_csv('heart.csv')
```

```
[3]: dataset
```

[3]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

The dataset used for this report is a Kaggle-produced dataset that was made available as one of many datasets' students can use for analysis. This dataset has 14 features and 303 instances. The target is ‘output’ which will tell us the likelihood of a person getting a heart attack or not.

The approach to analysing this data will be both to look at how individual features correlate with the target, how they influence or not the likelihood of getting a heart attack and if the whole dataset can be trained using a classification algorithm to predict the output which will help in both individual cases and in general to highlight danger signs and symptoms which if picked up early can save lives.

A) **The features**

There are 14 main features with 303 data instances. The features are presented as integers and floats except for the sex which is a string. Looking at the data, there will be need to reduce the features for our training model only working with those features that have some correlation or at least some relationship with the output, this is to ensure the noise in our data is reduced which will improve our predictions.

A.1 Age is the first feature of our dataset. There is a wide range of ages which improves the research as it considers all age groups. According to the American Heart Association, the average age of men to have a heart attack is 65 while for women it is 72.

A.2 Sex as the second feature is linked to age, but research has also shown that men are more likely to have a heart than women, this was demonstrated in Havard university research which showed that men were twice as likely to have a heart attack than women.

A.3 The Exang (exercise induced angina) feature measures if a patient has exercise induced angina (pain coming from the heart and radiating to the arm/shoulder). This is a strong indicator of the likelihood of a patient having a heart attack. According to the British Heart Foundation, this pain is caused by hardening of the arteries supplying the heart which reduces blood flow towards the heart.

A.4 The feature, ca, represents the number of major vessels, which in this data is from 0-3. This is not very clear, and I will not be using this feature as I cannot explain the feature from the available information.

A.5 The cp feature is about the type of chest pain which can be typical or atypical angina, non-anginal pain and asymptomatic. The presence of anginal pain as seen above may point to an imminent heart attack.

A.6 trtbps is the resting blood pressure. Referring to the American Heart Foundation, a high blood pressure can increase the risk of having a heart attack. So, a high resting blood pressure is a sign to be aware of.

A.7 chol measures the amount of bad cholesterol (non-HDL) in the blood which according to the NHS should be at a level below 4mmol.

A.8 The fbs (fasting blood sugar) which indicates the risk for diabetes is also an indicator to be aware of. This is so because diabetes will increase the risk of a patient having a heart attack. According to the European society of Cardiology, diabetes significantly increases the risk of heart attacks.

A.9 rest_ecg measures the heart activity at rest. This is done using an electrocardiogram. This is represented as 0 for normal activity, 1 for having a ST-T wave anomality and 2 which shows a probable or definite left ventricular hypertrophy which if left untreated will lead to heart failure and possibly a heart attack

A.10 Thalach is the maximum heart rate achieved. This feature will be eventually dropped during training to reduce dimensionality and prevent overfitting as other features like resting blood pressure and ecg measurements do present data that can be used in place of the maximum heart rate achieved.

A.11 Finally we have the output which is the target. It is binary as the output is either 1 for more chance of getting a heart attack and 0 for less chance of getting a heart attack.

B. The data will be processed to bring out trends using python. The data will be reduced by dropping some features especially those with no linear relationship with the target. The linear regression classifier will be used to train the model and the confusion matrix used to assess the accuracy and precision of our model.

THE MODEL

After preprocessing, reducing dimensions by dropping some features, the choice of a machine learning model was made, and it was to use the Logistic Regression (LR) classifier. This choice is mainly based on the output which is categorical and binary in nature.

The LR will be used to find the best hyperplane between the different classes making classification easier such that the model can learn better from the training set of data. This model will hopefully improve the predictions. I will be tweaking the parameters including the test size and random state to bring out the best confusion metrics scores.

Using matplotlib, a plot was made including all the values on the dataset before any processing to see the relationship between the various features with the target. Also, the correlation matrix is used to see if there are any strong positive, negative or no correlations between the features and the target in this dataset.

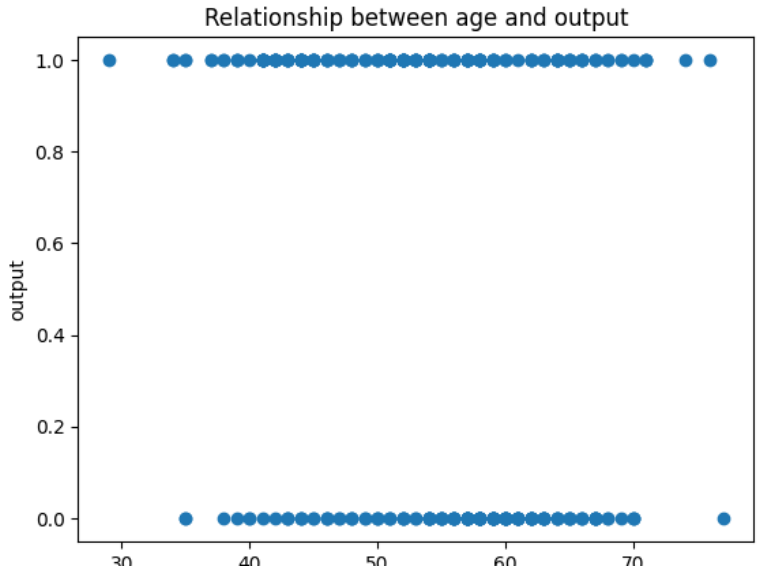
Using pairplot on seaborn we had the graph below.

The correlation matrix shows no linear relationships between the features and the target but generally some relationship which a classifier can separate into classes.

```
correlation_matrix=dataset.corr()
```

```
target_correlation=correlation_matrix['output']
```

[23]:
```
plt.scatter(dataset['age'],dataset['output'])
plt.xlabel('age')
plt.ylabel('output')
plt.title('Relationship between age and output')
plt.show()
```



With the research looking to confirm that the features have a relationship with the target, the various tests using the correlation matrix shows little or no such relations. We will need to use the chosen classification method to see if these features when put in a training model that classifies them will be able to make some predictions as to the risks of having a heart attack depending on risks factors.

The attachments below show the algorithm used.

[85]:
```
import seaborn as sb
import numpy as np
import matplotlib as plt
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn import svm
```

[10]:
```
Y=dataset.iloc[:,13]
```

[11]:
```
Y
```

[11]:
```
0      1
1      1
2      1
3      1
4      1
      ..
298    0
299    0
300    0
301    0
302    0
Name: output, Length: 303, dtype: int64
```

[12]:
```
X=dataset.iloc[:,0:12]
```

[13]:
```
X
```

[13]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa
0	63	1	3	145	233	1	0	150	0	2.3	0	0
1	37	1	2	130	250	0	1	187	0	3.5	0	0
2	41	0	1	130	204	0	0	172	0	1.4	2	0
3	56	1	1	120	236	0	1	178	0	0.8	2	0

```
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.2,random_state=2)
```

```
classifier=LogisticRegression()
classifier.fit(X_train,Y_train)
```

```
C:\Users\LORA\AppData\Local\Programs\Python\Python313\Lib\site-packages\sklearn\linear_model\_logistic.py:469: ConvergenceWarning: lbfgs failed to conver
ge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

▾ LogisticRegression

```
LogisticRegression()
```

```
predictions=classifier.predict(X_test)
predictions
```

```
array([1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1,
       0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0,
       1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1])
```

```
Y_test
```

```
99     1
296    0
```

```
'from the array above and comparing to the test, the predictions are not completely accurate. We will use a classification matrix to se
ny predictions were correct or not using this model'
```

```
print(classification_report(Y_test,predictions))
print(confusion_matrix(Y_test,predictions))
```

```
              precision    recall  f1-score   support

           0       1.00      0.81      0.90        32
           1       0.83      1.00      0.91        29

    accuracy                           0.90        61
   macro avg       0.91      0.91      0.90        61
weighted avg       0.92      0.90      0.90        61

[[26  6]
 [ 0 29]]
```

```
'''changing the values of the test size and random state gave us the best scores as seen above'''
```

```
'changing the values of the test size and random state gave us the best scores as seen above'
```

```
'''The confusion matrix is a 2 by 2 matrix as our output is either 0 for less chance of heart attack or 1 for  more chance of heart attack'''
```

```
'The confusion matrix is a 2 by 2 matrix as our output is either 0 for less chance of heart attack or 1 for  more chance of heart attack'
```

```
''' The above matrix shows that our model classified 0 correctly 26 times out of a possible 32 and misclassified 6 times with f1-score 0.9 and reca
```

```
' The above matrix shows that our model classified 0 correctly 26 times out of a possible 32 and misclassified 6 times with f1-score 0.9 and recall
81'
```

```
'''and it also predicted correctly the 29 out of 29 times in the test sample the 1's give an f1-score of 0.9 and recall of 0.91'''
```

```
"and it also predicted correctly the 29 out of 29 times in the test sample the 1's give an f1-score of 0.9 and recall of 0.91"
```

```
''' Further analysis of the results in the report'''
```

```
' Further analysis of the results in the report'
```

The above results after changing random state parameter to 2 and test size to 20%, we have a 2 by 2 matrix. The model correctly predicted the less likely to have a heart attack with a f1 -score of 90 and a recall of 81 and predicted the likely to have a heart attack with a f1-score of 91 and a recall of 1.00.

RESULTS

Looking at the results from the confusion matrix which tells us just how well our model predicts the outcomes using the data used, the recall for the predicted target has a recall rate of 0.8 for the '0' and 0.1 for the '1' output. The performance metrics while not at a 100% did bring up good enough scores and further tweaking of the parameters did not change much. Successfully classifying the observations shows our model was trained properly to a high degree and can be used to predict causes of heart attack by looking at the risk factors a patient presents with.

Several research studies have shown the common risk factors for getting a heart attack are about the same as the factors explored here in the form of our features in the dataset.

a. **The INTERHEART** study, which was conducted across 50 countries studying 15000 cases of heart attacks and 14000 controls remains one of the largest studies on heart attacks in the world. This study found that risk factors like smoking, hypertension, diabetes, and poor diet accounted for about 90% of the risk for a heart attack. This is about same with what we see in our dataset as

hypertension, diabetes and poor diet (which can lead to bad cholesterol building up) are main factors that can increase the risk of heart attacks.
The data analysed from our dataset will be nowhere near as comprehensive as the data that the above research would have produced, which must have easily shown correlation between the risk factors and the target.

b. **The Framingham Heart Study**, which started in 1948 is a cardiovascular cohort study which is highly influential in the health world. It provides great insight into general heart diseases including heart attacks and found that hypertension, diabetes, smoking and hyperlipidaemia are major factors that cause heart attacks. This again compares favourably to the data we explored in this study.

c. **The European Heart Journal study on risk factors for heart attacks** was carried out in Europe to find the traditional and non -traditional risk factors for developing a heart attack. The study confirmed hypertension and smoking as one of the more common causes of heart attacks but also found that stress and inflammation are beginning to play a major part in people getting heart attacks. Again, this confirms some of the features in our dataset are indeed major causes of heart attacks.

The above results confirm or at least point to the fact that the features in the dataset above do have an influence on the development of heart attacks. The dataset was not from real world studies but still pointed to some of the major risk factors that cause heart attacks as can be seen from the other benchmark studies.

```
'from the array above and comparing to the test, the predictions are not completely accurate. We will use a classification matrix to se
ny predictions were correct or not using this model'
```

```
print(classification_report(Y_test,predictions))
print(confusion_matrix(Y_test,predictions))
```

```
              precision    recall  f1-score   support

           0       1.00      0.81      0.90        32
           1       0.83      1.00      0.91        29

    accuracy                           0.90        61
   macro avg       0.91      0.91      0.90        61
weighted avg       0.92      0.90      0.90        61

[[26  6]
 [ 0 29]]
```

The confusion matrix above shows the accuracy of our model which is 90%. Not the best score in this case but does show that the model is right in its predictions most of the time. We can have a degree of confidence in saying that our features which are the independent variables will indeed cause changes in our target, the dependent variable 'output'. This model can be improved by dropping even further the variables that are not particularly needed in improving our model like 'cp' and 'Thalachh' while adding features like smoking, obesity and hyperlipidaemia.

CONCLUSION

Heart attacks remain a major public health concern affecting millions of people a year. There is always ongoing research to understand the causes and hopefully prevent the occurrence of this killer disease. There have been numerous studies in the world which have pointed to hypertension, smoking, cholesterol levels,stress,diabetes to name a few as major risk factors for heart attacks. Our data, while synthetic, did mirror real world happenings as was confirmed from comparing our features to trusted benchmark research studies.

The data from the ‘heart condition 'dataset which we have explored in this report does indeed point to some of this same risk factors. The features in our dataset included most of the risk factors as seen in other studies. These were explored in relation to the target which was the output (in this study it was binary; less or more chances of getting a heart attack).

The results from our model (Logistic Regression) training and confusion matrix points to a model that was trained to predict outcomes from this data. While the accuracy was not as high as we would have wanted, it still showed quite a high level of accurate predictions. This suggests our data is good enough to be trained to predict the likelihood or not of occurrence of a heart attack. The binary output was categorical which informed our choice of a machine learning model.

The conclusions drawn here corroborate the findings in other real world research data as presented in the results. This means changes in certain habits like smoking, overeating without exercises leading to obesity, uncontrolled diabetes, high consumption of bad cholesterol producing meals etc can go a long way in reducing the prevalence of heart attacks.

The question must be asked though; why are heart attacks still happening at the rates we know today? Is it because there is nothing that can be done or more education is needed concerning these risk factors? Can tobacco companies be forced to reduce sales to certain demographics especially teenagers who are introduced to smoking way too early? Can there be mandatory levels of weekly exercises as is done in some offices and schools already to reduce obesity? These questions and more need to be answered to ensure that the numerous research studies carried out on this topic do not end in academia, but lessons are drawn and implemented in the public to ensure the number of people dying from preventable heart attacks reduce.

REFERENCES

1.Benedikt et al. (2020). '*C*omparison of Cardiovascular Risk Factors in European Population Cohorts for Predicting Atrial Fibrillation and Heart Failure, Their Subsequent Onset, and Death'. J*ournal of the American Heart Foundation,* Vol9, number 9.

2.G, Aurelien. (2019) Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.2nd Edition. O'Reilly Media.

3.National Heart,Lung,and Blood Institute. Framingham Heart Study. Htpp//:www.nhlbi.nih.gov/science/Framingham-heart-study-fhs. Accessed:6th January 2025.

4. NHS (2024) *Causes of heart attacks,* Available at: Https//: www.nhs.uk/conditions/heart-attacks/causes/ Accessed: 6TH January 2025).

5.O, Stephanie., N, Abdissa.,Y,Salim.(2021) 'INTER-HEART: A global study of risk factors for acute myocardial infarction'. *American Heart Journal,* Vol.141, Issue 5, Pages 711-721.