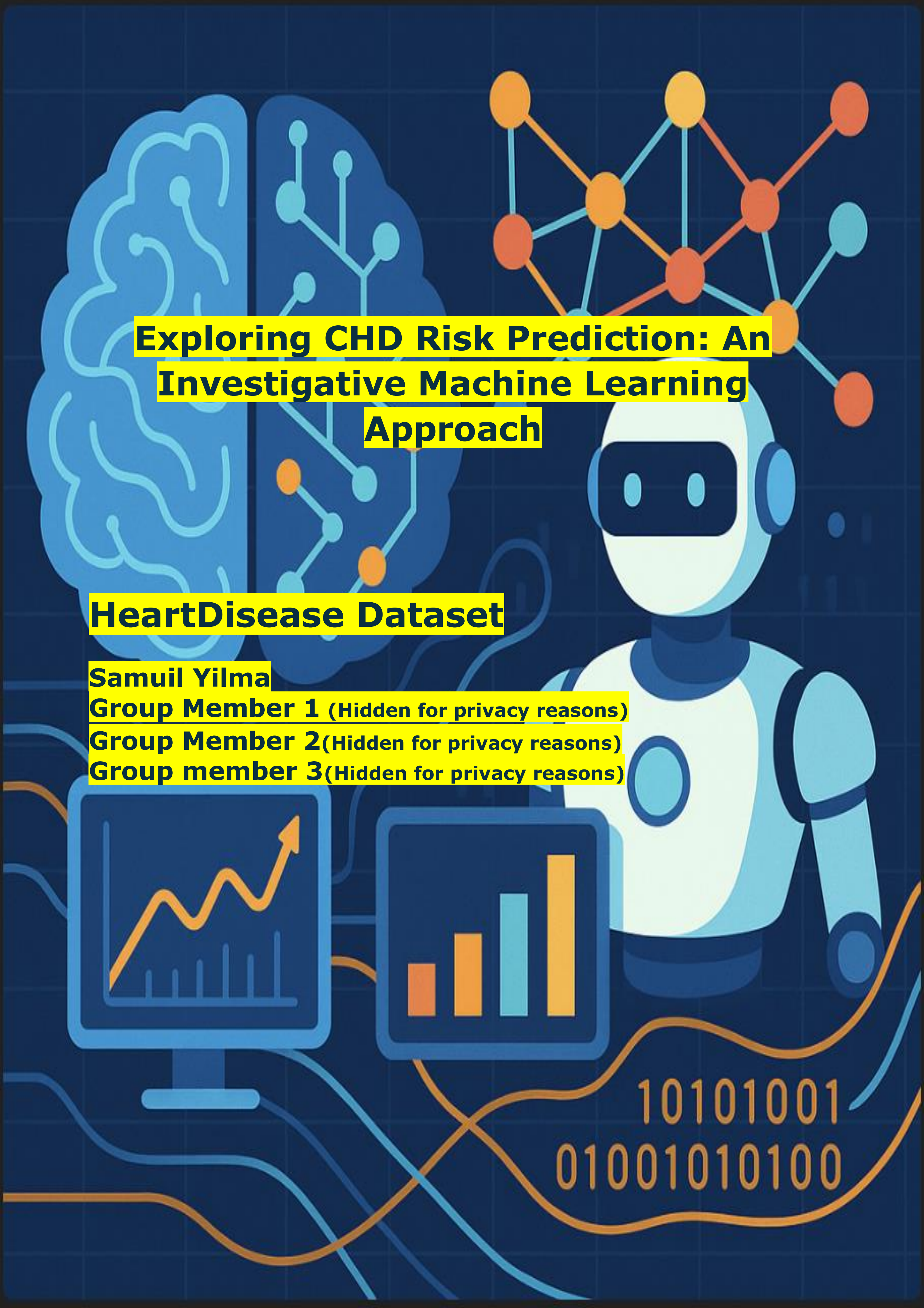


Exploring CHD Risk Prediction: An Investigative Machine Learning Approach

HeartDisease Dataset

Samuil Yilma
Group Member 1 (Hidden for privacy reasons)
Group Member 2(Hidden for privacy reasons)
Group member 3(Hidden for privacy reasons)

Executive Summary

This report explores whether it is possible to reliably predict an individual’s 10-year risk of developing coronary heart disease (CHD) using available health and lifestyle data. Utilising a logistic regression model trained on a heart disease dataset, we assessed a range of predictive factors including age, gender, cigarette consumption, history of stroke or hypertension, diabetes status, use of blood pressure medication, and body mass index. The model demonstrated strong performance, achieving a recall of 73% and a precision of 27%, indicating its ineffectiveness in identifying individuals at elevated risk. The dataset included over 4,000 anonymised patient records from a publicly available source, with a balanced mix of clinical and lifestyle variables. While the model demonstrates good recall, it’s low precision limits clinical utility. It may serve as a starting point for identifying patterns in at-risk populations, but further refinement is needed before deployment in real-world settings.

Table of Contents

Introduction

The chosen topic of this report is to apply supervised machine learning techniques to predict the likelihood of a patient suffering from coronary heart disease (CHD) within a 10-year window, and evaluate the results to determine whether the machine learning algorithm is appropriate.

Coronary heart disease (CHD) remains as one of the leading causes of death worldwide, posing as a major public health challenge due to its long-term progression and often silent onset. The World Health Organization has estimated that 17.9 million deaths occur globally every year due to heart diseases (World Health Organization, n.d.). Nearly half the deaths are within the United States and other developed countries are attributable to cardiovascular conditions (Coronado et al., 2022). This research is intended to pinpoint the most relevant factors of heart disease as well as predict the overall risk using a logistic regression model. As healthcare systems and insurance providers look to improve preventive care and reduce long-term costs, predictive modelling offers a solution. The ability to assess an individual's risk of developing CHD over a 10-year period could enable earlier interventions and more personalised health management strategies.

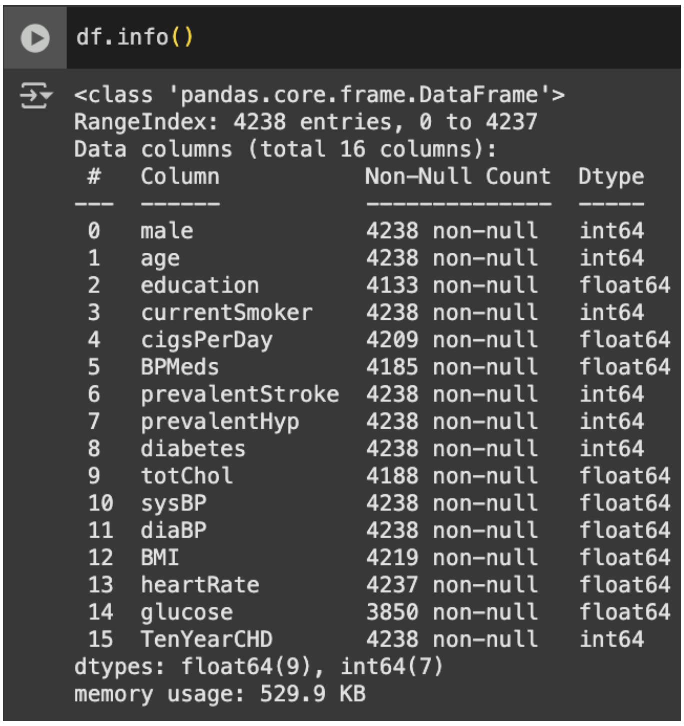
In this report, we address the question: Can we predict the 10-year risk of coronary heart disease using patient health data? To explore this, we use a publicly available heart disease dataset that contains a variety of clinical and demographic features relevant to cardiovascular health. The dataset contains anonymous patient records and also includes variables such as age, gender, cigsPerDay (average number of cigarettes smoked per day), prevalentStroke (whether the individual has had a stroke), prevalentHyp (history of hypertension), diabetes status, BPMeds (use of blood pressure medication), and BMI (Body Mass Index). These features were selected due to their known association with cardiovascular outcomes and their availability in routine health assessments.

Logistic regression was considered particularly suitable for this application not only because it handles binary classification well but also because of its interpretability. In healthcare, where decision transparency is essential, models like logistic regression allow clinicians to understand how each variable contributes to the prediction. Model performance was assessed using recall and precision, with a focus on the model’s ability to identify individuals at an elevated risk of CHD.
This makes the model more trustworthy and actionable compared to more complex black-box algorithms.

Descriptive Statistics

The dataset's sample size is 4238 rows and 16 columns, meaning before any changes or manipulation is done there are 4238 unique patients.

When assessing the dataset's quality, we took 4 dimensions into consideration, Completeness, Validity, Timeliness and Consistency. Completeness refers to the presence of all required information within a dataset. In this case important columns that must be filled were Male(Gender), Age and TenYearCHD. The datasets quality in terms of completeness can be regarded as complete.

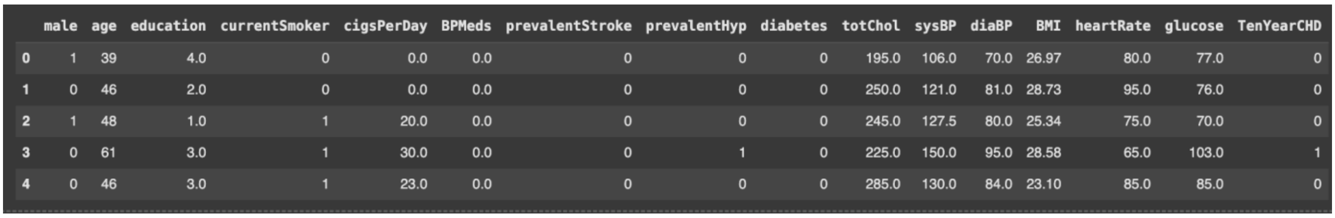


```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4238 entries, 0 to 4237
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   male             4238 non-null   int64  
 1   age              4238 non-null   int64  
 2   education        4133 non-null   float64
 3   currentSmoker    4238 non-null   int64  
 4   cigsPerDay       4209 non-null   float64
 5   BPMeds           4185 non-null   float64
 6   prevalentStroke  4238 non-null   int64  
 7   prevalentHyp     4238 non-null   int64  
 8   diabetes         4238 non-null   int64  
 9   totChol          4188 non-null   float64
 10  sysBP            4238 non-null   float64
 11  diaBP            4238 non-null   float64
 12  BMI              4219 non-null   float64
 13  heartRate        4237 non-null   float64
 14  glucose          3850 non-null   float64
 15  TenYearCHD       4238 non-null   int64  
dtypes: float64(9), int64(7)
memory usage: 529.9 KB
```

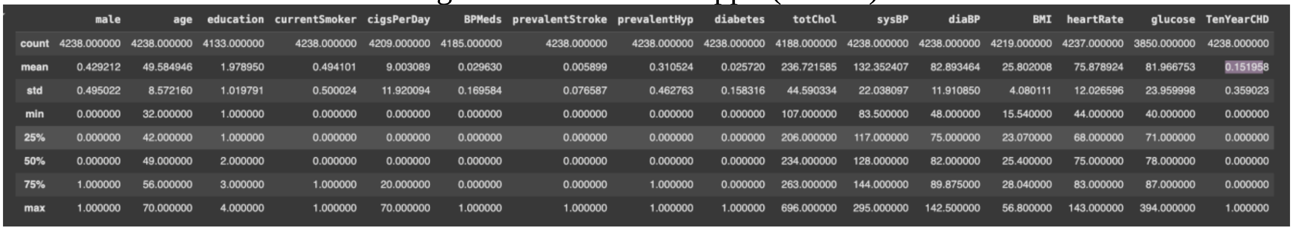
Figure 1: Dataframe Information

Validity describes the way the dataset follows business rules standards such as the format (number of strings/digits), range (minimum & maximum) and datatypes. When assessing the datasets validity we dived into each variable's format, range and appropriate datatypes. When glancing at a snippet of the first 5 patients each variable has the appropriate formats such as Male(gender) being restricted to 1 digit, age being restricted to 2 digits, totChol being restricted to 3 digits, tenYearCHD being restricted to 1 digit. The other variables such as sysBP, BMI, glucose, and heartrate are allowed to have some freedom in terms of digits, where they are all allowed between 2 and 3 digits. All variables have appropriate ranges; there are no variables that out of order ranges.



	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Figure 2: Dataframe snipper(5 rows)



	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
count	4238.000000	4238.000000	4133.000000	4238.000000	4209.000000	4185.000000	4238.000000	4238.000000	4238.000000	4188.000000	4238.000000	4238.000000	4219.000000	4237.000000	3850.000000	4238.000000
mean	0.429212	49.584946	1.978950	0.494101	9.003089	0.029630	0.005899	0.310524	0.025720	236.721585	132.352407	82.893464	25.802008	75.878924	81.966753	0.151958
std	0.495022	8.572160	1.019791	0.500024	11.920094	0.169584	0.076587	0.462763	0.158316	44.590334	22.038097	11.910850	4.080111	12.026596	23.959998	0.359023
min	0.000000	32.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	107.000000	83.500000	48.000000	15.540000	44.000000	40.000000	0.000000
25%	0.000000	42.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	206.000000	117.000000	75.000000	23.070000	68.000000	71.000000	0.000000
50%	0.000000	49.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	234.000000	128.000000	82.000000	25.400000	75.000000	78.000000	0.000000
75%	1.000000	56.000000	3.000000	1.000000	20.000000	0.000000	0.000000	1.000000	0.000000	263.000000	144.000000	89.875000	28.040000	83.000000	87.000000	0.000000
max	1.000000	70.000000	4.000000	1.000000	70.000000	1.000000	1.000000	1.000000	1.000000	696.000000	295.000000	142.500000	56.800000	143.000000	394.000000	1.000000

Figure 3:Dataframe summary

Timeliness refers to whether the information is current and up to date for its intended use. All the date is up to date these are all recorded values that don't need updating or changes. Unless there is a significant change in a patient's lifestyle, i.e The patient lost/gained a notable amount of weight affecting their BMI.

Consistency is present when a particular rule or order is applied across all columns and every patient's record abides by the ruleset. The dataset follows a thorough consistency throughout all patients as all values are consistent and realistic.

Overall, we can make a confirmation the dataset's integrity is excellent as the dimensions Completeness, Validity, Timeliness and Consistency are at a high standard.

Our analysis involves many variables regarding patients. The first step required was to determine what our target variable is and identify variables to fit within our model. Our target variable is 10YearCHD and using variables to predict it.

An overview of coronary heart diseases released by the NHS (2024) claims that conditions such as diabetes or hypertension increase the risk of CHD, therefore we have decided to target those variables first and take a look if we can identify whether they're useful target variables.

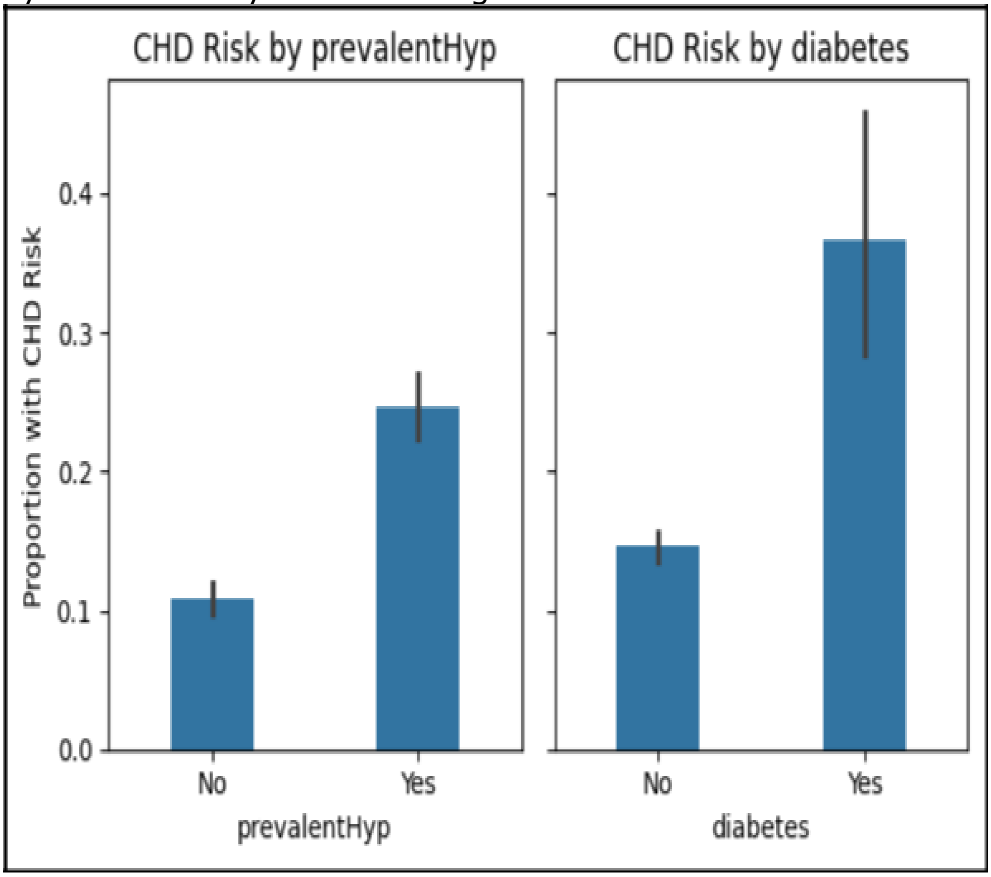


Figure 4: Bar plot of prevalentHyp and diabetes with CHD status

We have decided to include both prevalentHyp and diabetes in our model because each group are displaying a higher proportion of those with a 10 Year risk of CHD. Other variables that seemed interesting were tested out such as prevalentStroke, BPMeds and gender. We made sure to run the same bar plot tests with CHD as we did with hypertension and diabetes. As you can see in the figure below there was the same outcome as before, where one group is displaying a higher proportion to those with a 10-year risk of CHD

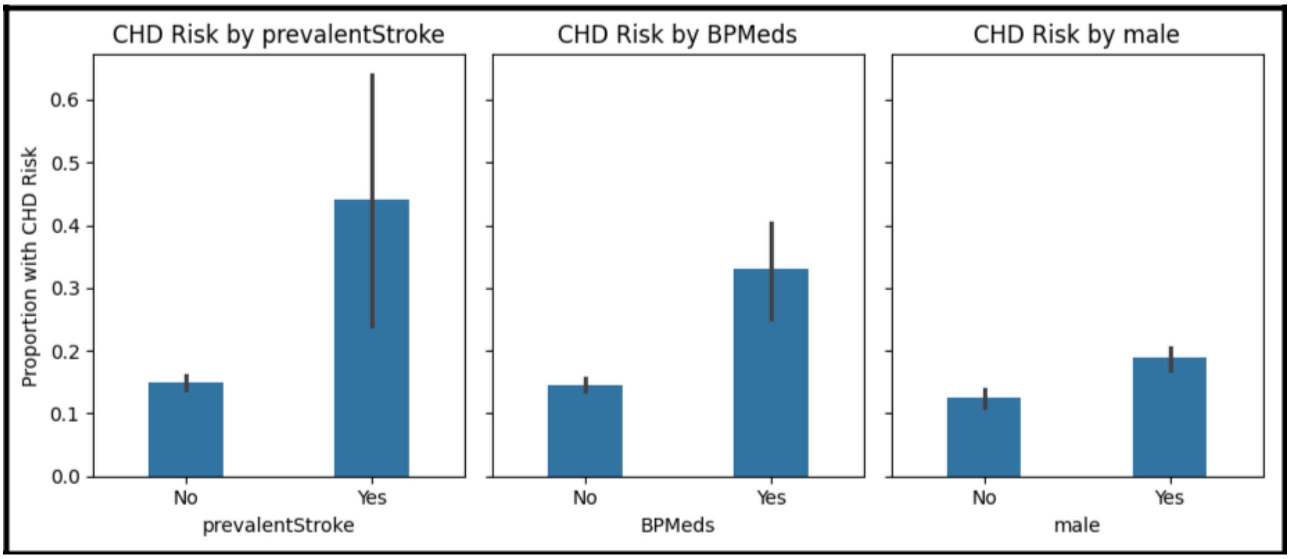


Figure 5: Bar plot of prevalentStroke, BPMeds, gender with CHD status

The bar plot model was run through CHD and variables such as age, totChol, BMI, heartrate and glucose.

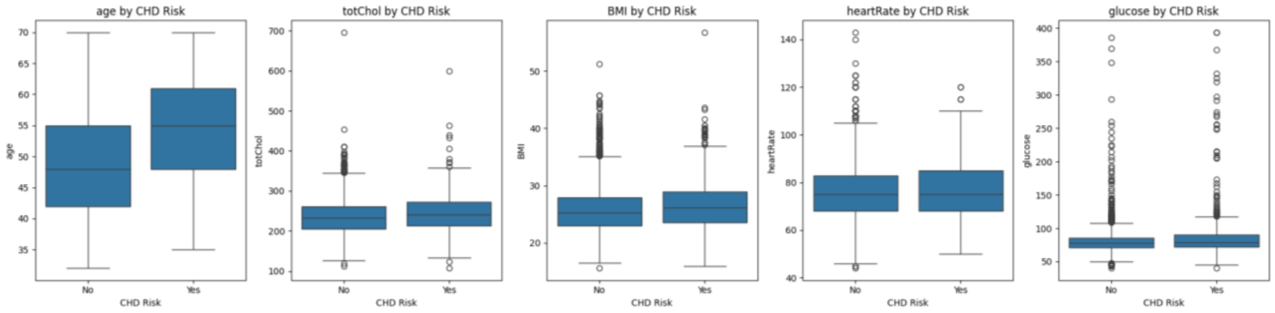


Figure 6: Boxplots of age, totChol, BMI, heartRate, glucose by CHD status

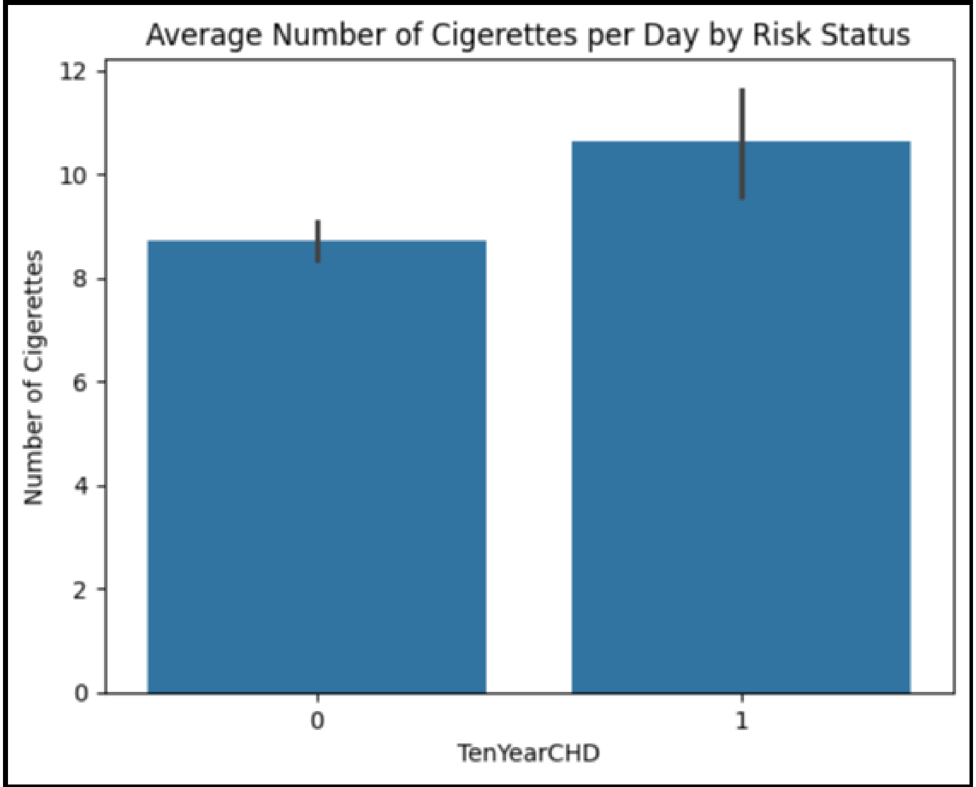


Figure 7: Bar plot of cigsPerDay vs CHD status

We included age and BMI in our final model based on the clear differences between groups. The variable of currentSmoker was removed but cigsPerDay was included. This is due to both variables being multicollinear, they explain the same behavior. When running cigsPerDay through a bar plot model with CHD, we observed that the CHD group tends to smoke more.

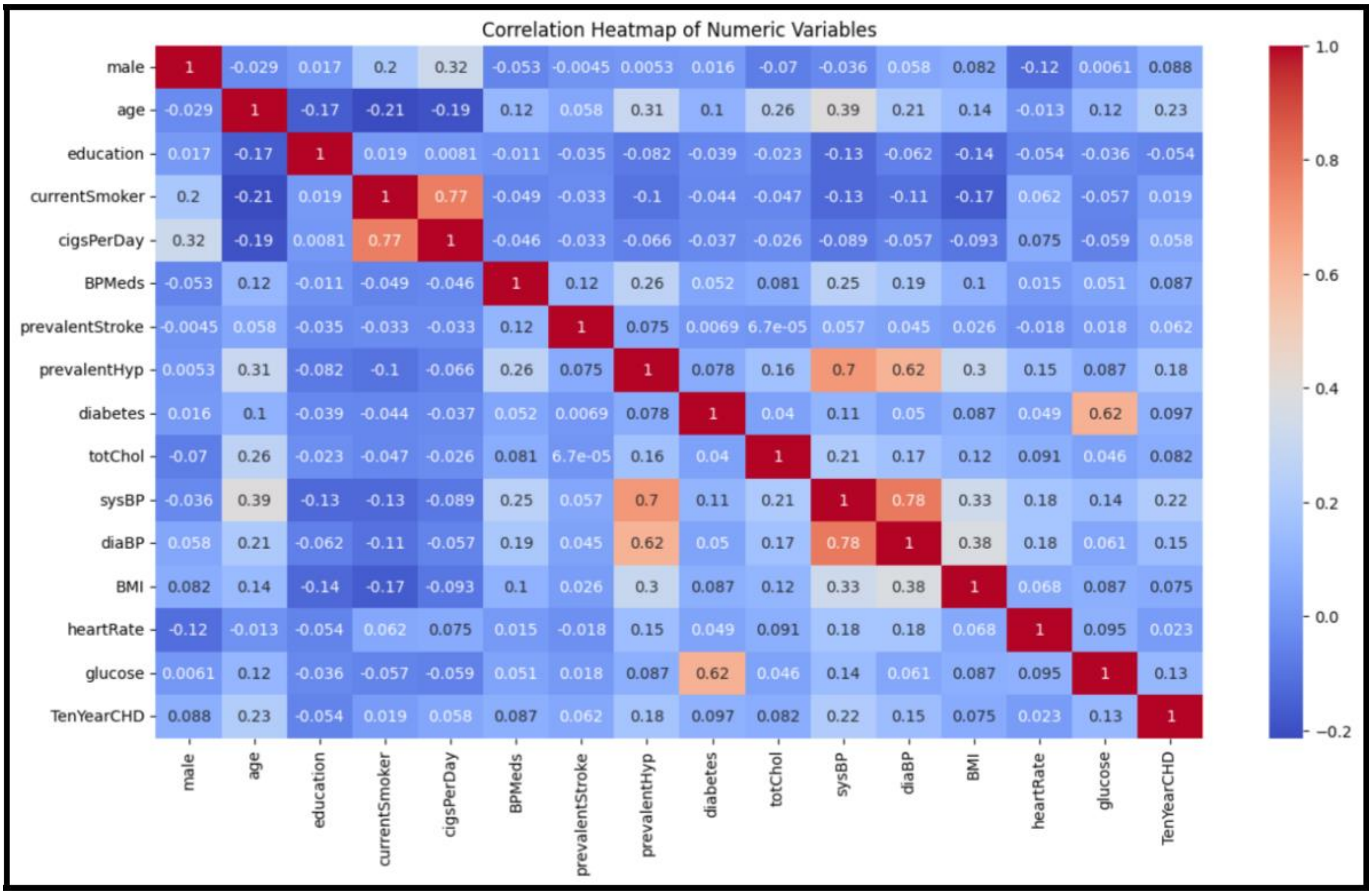


Figure 8: Heatmap of correlation between numeric variables.

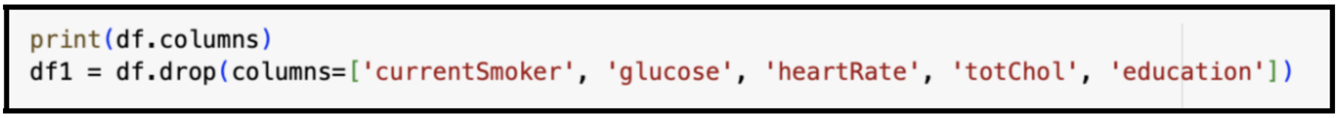
The heatmap above exhibits variables that are correlated together. As expected, we can see that both cigarette (cigsPerDay/currentSmoker) variables are correlated with a rank of 0.77. Our final variables that influence CHD and are going into our machine learning model are gender, age, cigPerDay, prevalentStroke, prevalentHyp, diabetes, BPMeds and BMI.

A review of the target variable distribution revealed that approximately 15% of individuals in the dataset were labelled as having a 10-year CHD risk, while the remaining 85% were classified as not at risk. This highlights a significant class imbalance, which would later impact our model training and evaluation. The imbalance informed our decision to apply techniques like class weighting to improve model performance

Data Preprocessing

Data Preprocessing is an important step and a must in designing a machine learning model. It ensures that the data is structured, clean and reliable for analysis. The steps we proceeded with are dropping variables, dropping null values, removing unrealistic/unnecessary values, and standardising numerical values.

Dropping Variables Not Used

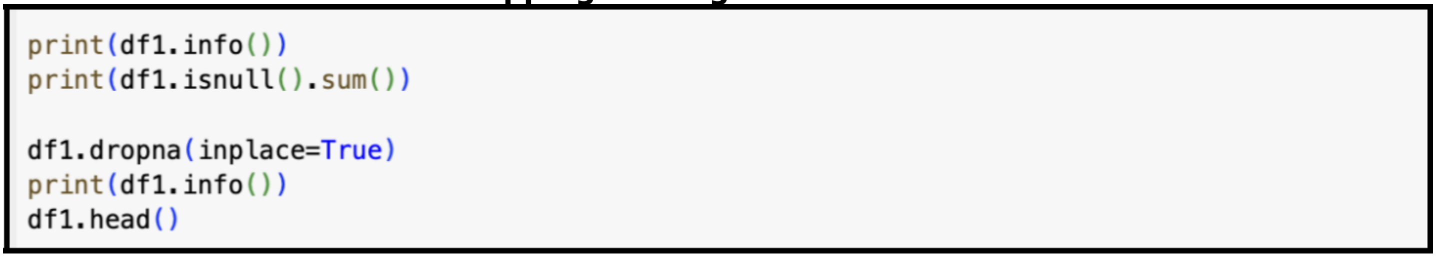


```
print(df.columns)
df1 = df.drop(columns=['currentSmoker', 'glucose', 'heartRate', 'totChol', 'education'])
```

Code Snippet 1: Dropping columns using .drop()

We created a new data frame by removing features that were not useful or lacked metadata: currentSmoker, glucose, heartRate, totChol and education. This reduced noise and minimised multicollinearity.

Dropping Missing Values



```
print(df1.info())
print(df1.isnull().sum())

df1.dropna(inplace=True)
print(df1.info())
df1.head()
```

Code Snippet 2: Checking and removing nulls with .isnull() and .dropna()

We identified missing values in features like cigsPerDay, BPMeds, and BMI. Since these were few and we had a large dataset, we dropped the incomplete rows.

Although data imputation techniques such as mean or median filling were considered, we opted to drop the rows with missing values due to the relatively small number of missing entries and the overall size of the dataset. This avoided introducing potential bias from imputation assumptions, while preserving data integrity.

Checking Unrealistic Values

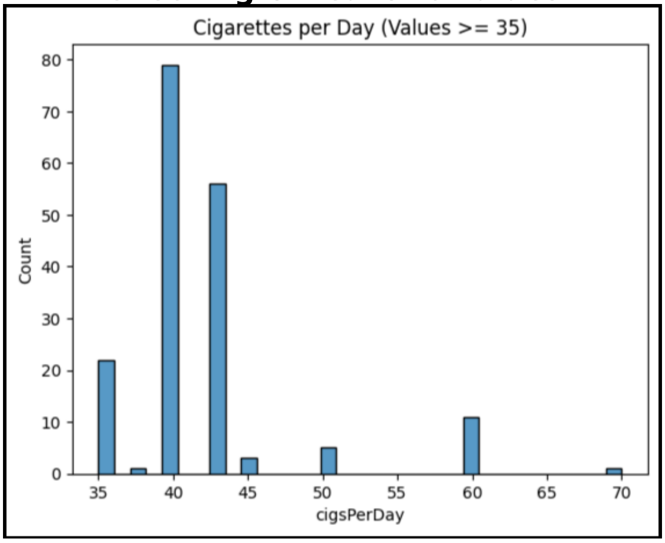
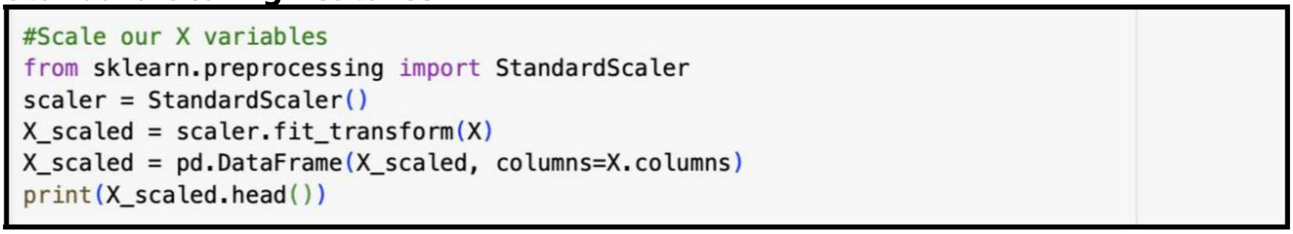


Figure 6: Histogram of cigsPerDay > 35

The maximum value for cigsPerDay was 70, which seemed extreme. However, further inspection showed that many individuals reported smoking 35+ cigarettes a day. This right-skewed behaviour is realistic, so we retained these values.

Standard Scaling Features



```
#Scale our X variables
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
X_scaled = pd.DataFrame(X_scaled, columns=X.columns)
print(X_scaled.head())
```

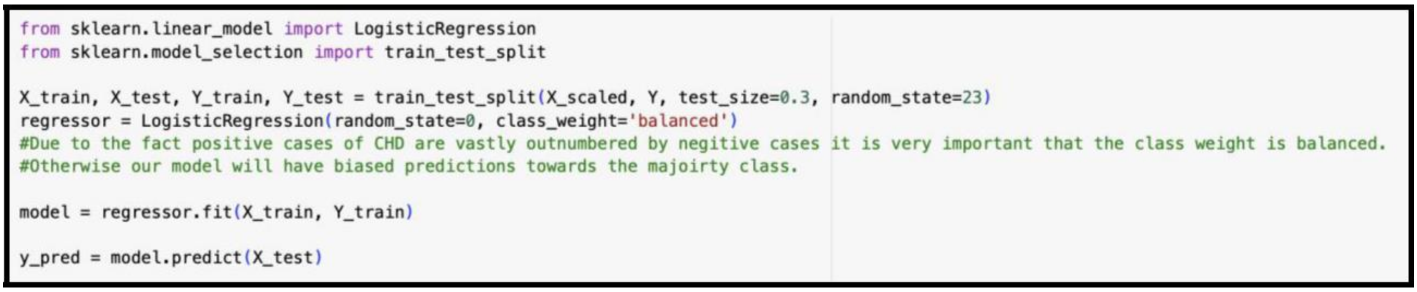
Code Snippet 3: StandardScaler usage on selected variable

We applied standard scaling to normalise the feature ranges, ensuring that no variable would dominate the learning process due to scale differences.

To avoid underfitting which is the process of where the model is not complex enough to learn relationships within the data, we were extremely careful in what we cleaned/removed. In conclusion the finalised dataset is slightly different from the original dataset. The finalised dataset has 4137 patients which is only 101 less patients than the original dataset. The columns Male(Gender), age, CigsPerDay, prevalentStroke, prevalentHyp, Diabetes, BPMeds, BMI were all kept as they were analysed with the dependant variable TenYearCHD and have all shown that they play a role in determining TenYearCHD.

Model Results

After preprocessing the dataset and selecting the most relevant features we implremented and evaluated 3 machine learning algorithms, Logistic Regression, K-Nearest Neighbour (KNN) and Kmeans clustering. The target variable was binary to where 1 represented that the patient had a 10 year risk of CHD and 0 represented none. The dataset was split 70% training and 30% testing to ensure genealisability and reduce the risk of overfitting.



```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(X_scaled, Y, test_size=0.3, random_state=23)
regressor = LogisticRegression(random_state=0, class_weight='balanced')
#Due to the fact positive cases of CHD are vastly outnumbered by negitive cases it is very important that the class weight is balanced.
#Otherwise our model will have biased predictions towards the majoirty class.

model = regressor.fit(X_train, Y_train)

y_pred = model.predict(X_test)
```

Code Snippet 4: Logistic Regression with class_weight='balanced'

Logistic Regression

Logisitc regression is a supervised classification algorithm suited for binary classification problems like this. There was an imbalance to where 15% of patients were CHD positive, the ‘class_weight=’balanced’’ was used to penalise misclassifying the minority class.

Metrics:

Accuracy Score	0.67
Precision Score	0.27
Recall Score	0.73
F1 Score	0.40

Confusion Matrix	[[702, 356] [49, 135]]

K Nearest Neighbour

KNN is a supervised macine learning algorithm, it classifies a data point based on the majority label of it’s k nearest neighbours from the training space. We tested several values of k and selected k=3 as it provided the best validation results. While KNN achieved a high accuracy of 83%, the model’s recall was on 16% meaning it failed to identify most CHD-positive patients

Confusion Matrix	[[1031, 27] [170, 14]]

KMeans

KMeans is an unsupervised learning algorithm used to group data into *k* clusters based on similarity. We attempted to use KMeans with **k=2** (hoping to separate CHD-positive and CHD-negative individuals), but since KMeans does not use label information during training, it is **not inherently suitable for prediction tasks** like CHD diagnosis.

KMeans is an unsupervised machine learning algorithm used to group data into k clusters based on similarity. WE attempted to use KMeans where k=2, but since KMeans does not use label information during training, it is not inherently suitable for prediction tasks like CHD diagnosis. KMeans produced low scores overall, the accuracy was 55% and the recall was 40%, indicating it could not effectively cluster CHD cases. It cannot learn the relationship of between input features and known outcomes.

```
    Confusion Matrix        [[2005,  1511]
                             [ 370,   251]]
```

Model Evaluation, Comparison and Selection

In a healthcare setting, recall is critical. Our logistic regression model had the highest recall, correctly identifying 73% of actual CHD-positive patients. False negatives pose serious medical risks, so this trade-off is acceptable, where a false positive may lead to further tests, but a false negative can delay treatment and increase risk.

Although precision should be not ignored entirely. A model with low precision may result in unnecessary anxiety, costs and medical follow ups. Our goal was to achieve a balance, and among all models tested, Logistic Regression achieved the best trade-off between recall and precision

Conclusion

Our project began with a simple but pressing question: can we accurately predict an individual's 10 -year risk of coronary heart disease (CHD) using patient-level health and lifestyle data? Through careful data analysis, model testing, and evaluation, our team has developed a more nuanced understanding of the factors that influence heart disease risk and the strengths and shortcomings of predictive models within the healthcare domain.

Key Learnings and Model Insights

One of the most valuable outcomes of this project was identifying a focused set of variables that meaningfully contribute to CHD prediction. While the original dataset included numerous clinical and lifestyle factors, we learned that not all were equally useful. Through correlation testing, visual analysis (as seen in Figures 1–5), and comparison with medical research, we narrowed down our variables to a strong predictive core: age, gender, cigarettes per day, prevalent stroke, hypertension, diabetes, blood pressure medication use, and BMI.

These variables were not only statistically relevant but also medically interpretable. For instance, higher age and BMI levels were more common in individuals flagged at risk, and average cigarette consumption was notably higher in the CHD-positive group. By dropping weak predictors like glucose, cholesterol, and heart rate, which showed no meaningful separation between classes, we avoided introducing noise and multicollinearity into the model, especially between 'currentSmoker' and 'cigsPerDay'.

Choosing the right model was also a critical part of the learning process. We trialed three different machine learning algorithms, logistic regression, K-Nearest Neighbours (KNN), and KMeans clustering, each with its strengths and limitations. Logistic regression stood out as the most effective approach, delivering a recall of 73%, which significantly outperformed the other models. While its precision was only 27%, the higher recall ensured that the model successfully identified a greater number of true CHD-positive cases, which is of primary importance in a medical setting.

In contrast, KNN produced high accuracy but very low recall (16%), failing to capture most positive cases. KMeans, being unsupervised, struggled with classification accuracy and recall as well, suggesting that clustering was not the right approach given the lack of strong feature separation in the dataset.

Through these results, we gained a crucial understanding: when building health-predictive tools, it is not enough for a model to be statistically robust, it must also align with the ethical and practical needs of healthcare. Specifically, minimising false negatives is vital when lives are at stake.

From Prediction to Prevention

The real-world implications of our findings are substantial. Hospitals, general practitioners, and public health agencies could potentially use a model like ours as a triage tool to flag high-risk patients for early screening and intervention. Although a 73% recall rate means some at-risk individuals will still go undetected, it's a significant improvement over random or uniform screening procedures.

For example, in a population of 5,000 patients, assuming the same proportions as our dataset, the model would correctly identify approximately 945 of the 1,300 individuals at risk. While this leaves about 355 cases undetected, a limitation that we do not take lightly, it also enables more than 70% of those patients to receive preventive care sooner. The cost of a false positive in this context (i.e., unnecessarily flagging someone as high-risk) may simply result in an extra check-up or lifestyle assessment. On the other hand, a false negative could lead to long-term hospitalisation, intensive care, or even premature death. This trade-off justifies the model's recall-first approach.

Ethical Considerations and Limitations

Despite the promising findings, our project has several limitations that require acknowledgment. The first is the simplicity of the dataset. While useful for introductory machine learning applications, it lacks deeper features that influence cardiovascular risk, such as physical activity levels, stress markers, genetic predispositions, socioeconomic status, and access to healthcare. Adding such features would undoubtedly improve predictive power and reduce both false positives and false negatives.

Another limitation relates to bias and fairness. The dataset does not capture demographic diversity beyond gender, and even gender was reduced to a binary male/female classification. This raises ethical concerns, especially in real-world deployments where predictive tools must serve broad and inclusive populations. If unbalanced data or underrepresented groups are not properly addressed, the model could inadvertently reinforce existing disparities in healthcare.

Data integrity also presented occasional concerns. For instance, the variable 'cigsPerDay' had a maximum value of 70, which initially seemed implausible. However, further inspection revealed a right-skewed distribution, with multiple patients reporting 35 to 60 cigarettes per day. We chose to retain these observations, but this highlighted the broader challenge of distinguishing between true outliers and data-entry errors, especially when working with self-reported behavioral data.

Additionally, while we attempted to address class imbalance by using class_weight='balanced' in our logistic regression model, future work could benefit from more advanced resampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique) or ensemble methods that penalise false negatives more explicitly.

Opportunities for Future Work

There are several avenues for expanding and refining this project. From a technical standpoint, future models could explore the use of ensemble techniques such as Random Forests or XGBoost, which may provide higher precision without sacrificing recall. These models often perform well on tabular healthcare data and can handle non-linear relationships more effectively than logistic regression.

In terms of feature engineering, incorporating longitudinal data (e.g., changes in BMI, blood pressure, or smoking habits over time) would enable dynamic risk prediction rather than static snapshots. This could open the door to time-series modelling using recurrent neural networks or other temporal algorithms.

From a systems perspective, embedding this model into an actual hospital workflow would require creating a user interface, validating the model on local patient data, and training clinical staff on how to interpret and act on predictions. This real-world integration poses both technical and human challenges, but it is essential for translating academic projects into actionable health tools.

Finally, partnerships with public health researchers and ethicists could help guide responsible deployment. Questions such as "Who owns the data?", "How is consent obtained?", and "What happens when a patient disagrees with their predicted risk?" are all crucial for ensuring that predictive models support, not undermine, individual autonomy and public trust.

Final Reflections

At its core, this project was about using data for good. We set out not only to build a predictive model but also to understand how machine learning can support preventive medicine. Along the way, we confronted the complexities of real -world data, wrestled with trade-offs between different metrics, and came to appreciate the ethical landscape of healthcare analytics.

While our logistic regression model is not perfect, and we do not recommend deploying it in isolation, it offers a promising foundation. With further refinement, additional data, and careful integration into medical practice, such tools could help clinicians identify at-risk patients earlier, save lives through proactive care, and reduce the economic burden of chronic disease.

In conclusion, data is only powerful when transformed into action. By moving from prediction to planning, and from planning to prevention, our project lays the groundwork for future systems that use AI not only to diagnose disease, but to avoid it altogether.

References

Coronado, F., Melvin, S. C., Bell, R. A., & Zhao, G. (2022). Global responses to prevent, manage, and control cardiovascular diseases. *Preventing Chronic Disease, 19*, 220347. https://doi.org/10.5888/pcd19.220347

World Health Organization. (n.d.). *Cardiovascular diseases*. Retrieved June 2nd, 2025, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1