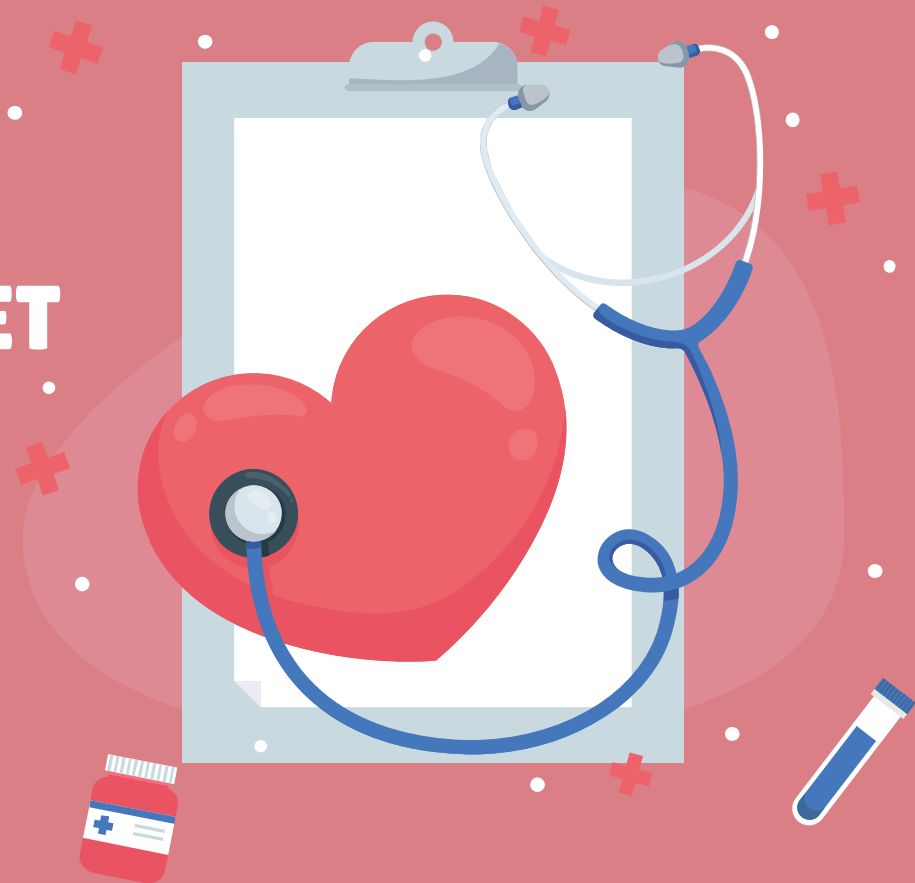


# CARDIOVASCULAR. DISEASES RISK PREDICTION DATASET

**CIS 5450 – Big Data  
Analytics–Project Fall 2023  
Group Members:**

- 1) Samarth Chandrawat**
- 2) Manas P Shankar**
- 3) Hardik Jain**





# OBJECTIVE

- Heart disease is a major global health concern, claiming millions of lives and putting a tremendous strain on healthcare systems.
- Early detection and accurate prediction of heart disease are critical in controlling and avoiding the illnesses beginnings, as well as improving patient outcomes. Advances in medical research, technology, and data analysis have considerably improved our understanding of cardiac disease prediction throughout the years.





# VALUE PROPOSITION

- This brief overview seeks to provide an overview of the progress made in this subject, highlighting significant methodology and approaches used in heart disease prediction.
- We can obtain significant insights into the shifting landscape of heart disease prediction by investigating numerous risk factors, diagnostic techniques, and future trends, opening the way for more proactive and personalised treatment solutions.



# TABLE OF CONTENTS

**1**

The Dataset

**2**

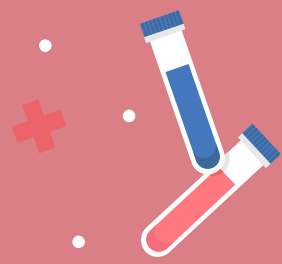
Exploratory Data  
Analysis

**3**

Modelling

**4**

Results and  
Conclusion



1

# The Dataset

Here, we will present the details about the dataset we used for analysing the problem.





# The Dataset–Columns/Features

This dataset is related to health and lifestyle factors, in relation to cardiovascular disease (CVD). Here is a brief description of the columns:

Feature name	About the Feature
General_Health	General health status of the individuals. The values include "Poor", "Very Good", "Good", etc.
Checkup:	Information about the individual's last checkup. The values include "Within the past 2 years", "Within the past year", etc.
Exercise:	Whether the individual exercises or not.
Heart_Disease:	Whether the individual has heart disease
Skin_Cancer:	Whether the individual has skin cancer.
Other_Cancer	Whether the individual has any other type of cancer.

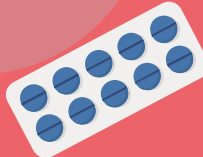




# The Dataset–Columns/Features



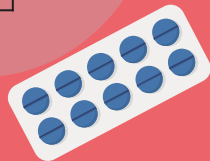
Feature name	About the Feature
Arthritis	Whether the individual has arthritis. Sex: Gender of the individual.
Age_Category:	Age category of the individual. The categories seem to be in ranges such as "70-74", "60-64", "75-79", "80+", etc.
Height_(cm)	Height of the individual in centimeters.
Weight_(kg):	Weight of the individual in kilograms.
BMI	Body Mass Index of the individual.
Smoking_History	Whether the individual has a history of smoking.





# The Dataset–Columns/Features

Feature name	About the Feature
Depression:	Whether the individual suffers from depression.
Diabetes:	Whether the individual has diabetes.
Alcohol_Consumption:	Amount of alcohol consumption.
Fruit_Consumption	Amount of fruit consumption.
Green_Vegetables_Consumption:	Amount of green vegetable consumption.
FriedPotato_Consumption:	Amount of fried potato consumption.







# The Dataset–Columns/Features

- The dataset contains 308,854 rows (individuals) and 19 columns (features).
- The dataset didn't have any NULL values, but had some rows that had been duplicated.
- The dataset had about 80 duplicated rows and removing these duplicated rows was necessary in order to carry out the analysis and modelling efficiently and obtain the correct results.
- After removing these duplicates we were left with 308,774 rows and 19 columns.

We will now move on to the Exploratory Data Analysis part and see how we can get more insights about our data which will help us to answer some of the important questions we may have related to the problem.



2

# Exploratory Data Analysis

Here, we will try to gain insights about some of the most important characteristics about our data.



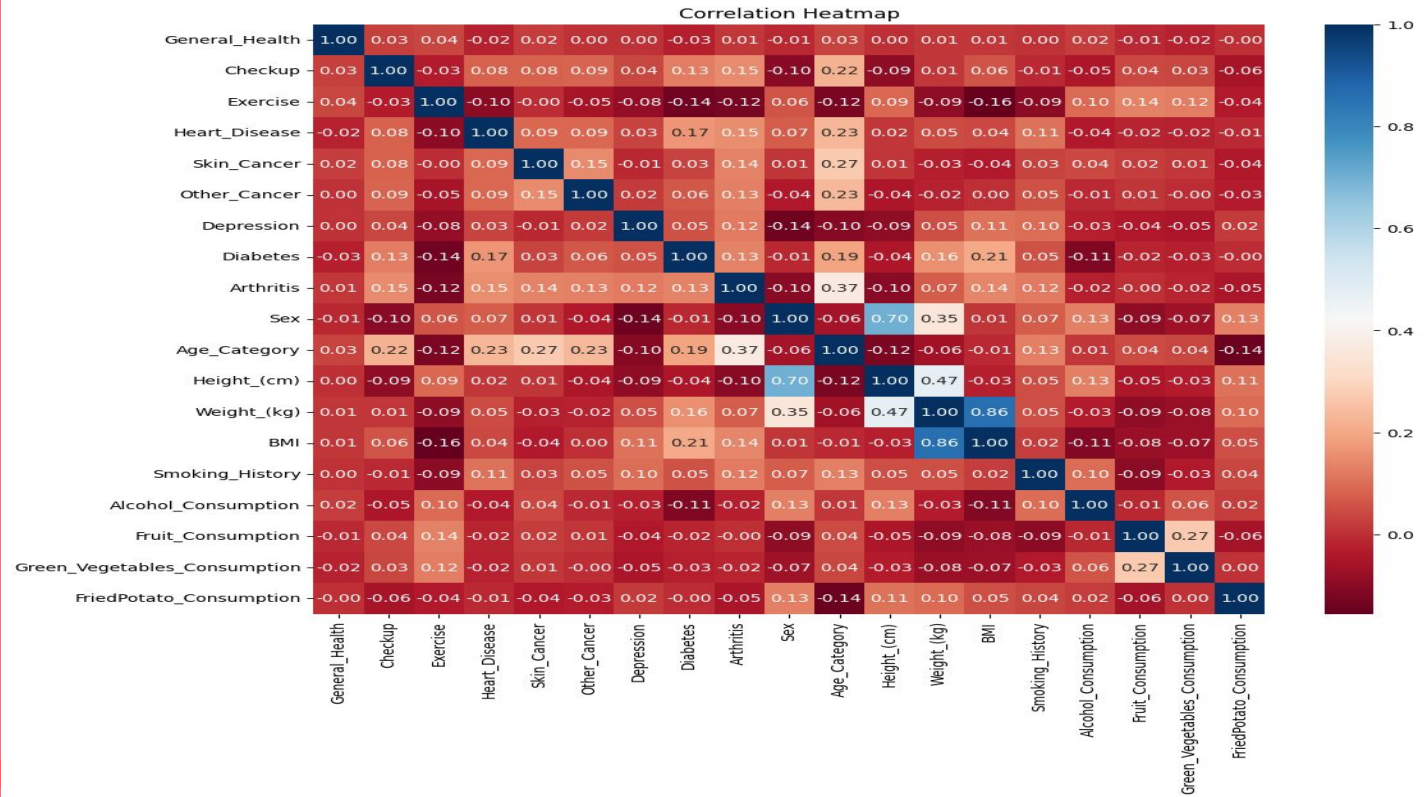


# 1. Initial Analysis

- We first start with our Exploratory Data Analysis(EDA) part by performing a univariate analysis, where we check the count of each unique value in column in the dataset for each of the columns by plotting a bar plot with the count on the y-axis and the value on the x-axis.
- We also performed Bivariate analysis, where we are exploring the connection between various disease conditions and a set of selected variables. The chosen variables include 'General\_Health,' 'Exercise,' 'Sex,' 'Age\_Category,' and 'Smoking\_History,' while the target diseases under investigation are 'Heart\_Disease,' 'Skin\_Cancer,' 'Other\_Cancer,' 'Diabetes,' and 'Arthritis.'
- For each combination of disease and variable, a countplot is generated, providing a visual representation of the distribution of the variables concerning the presence or absence of the disease.
- We also plotted the correlation matrix where the correlation coefficient quantifies the strength and direction of a linear relationship between two variables.
- For our dataset, we would be better able to visualize which columns are more correlated with each other and also check which columns are more correlated with the 'heart disease' column in order to better choose the features for analysis.



## 2. Correlation Matrix

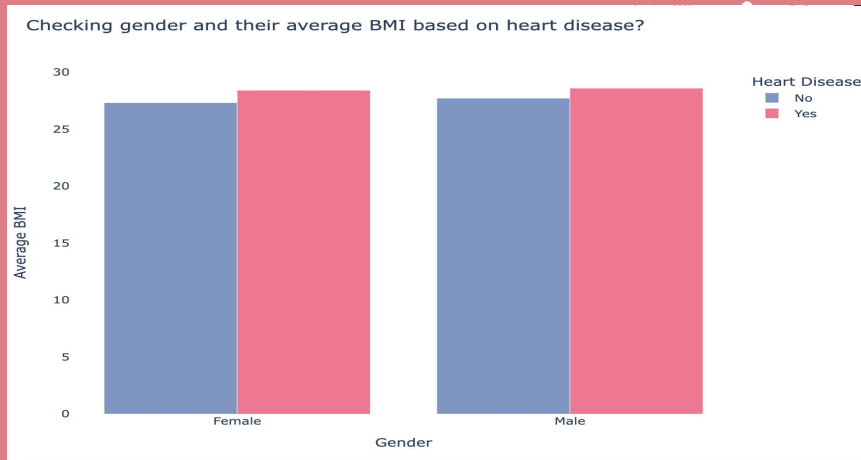




### 3. Analysis of BMI(Body Mass Index)

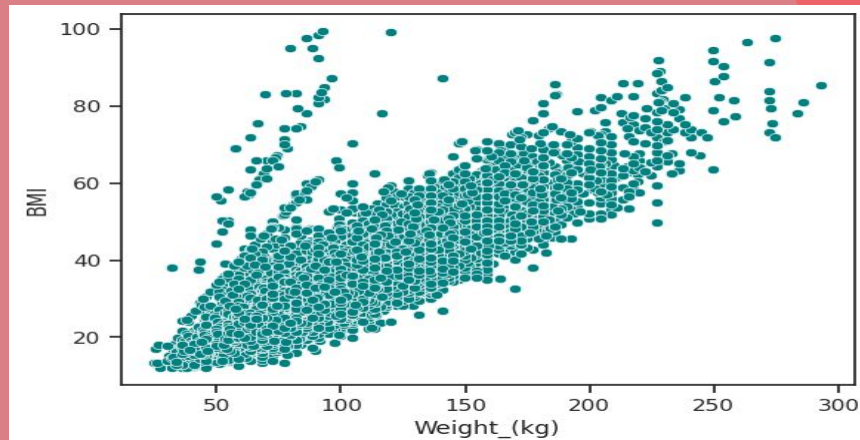
#### 1. Gender and the average BMI based on Heart disease

Individuals with heart disease, regardless of gender, exhibit higher average BMI compared to those without heart disease, suggesting a correlation between elevated BMI and the presence of cardiovascular issues in both males and females.

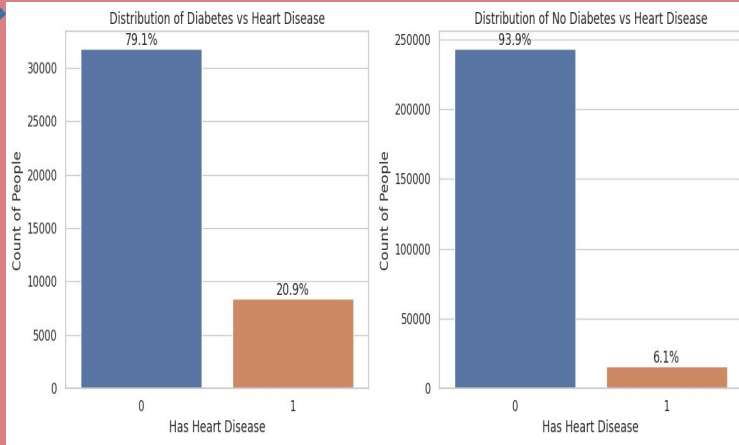


#### 2. BMI and Weight

We find that there is a linear relationship between weight and BMI. The pearson cooeffcient obtained was 0.86.

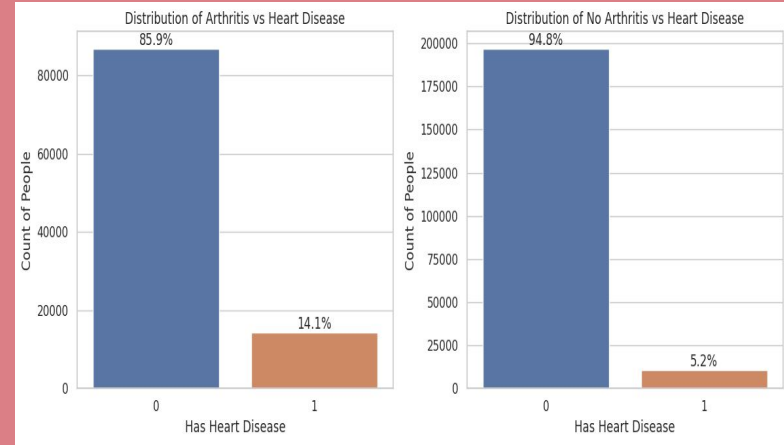


# 4. How having other diseases affect heart disease?



## 1. Heart Disease and Diabetes

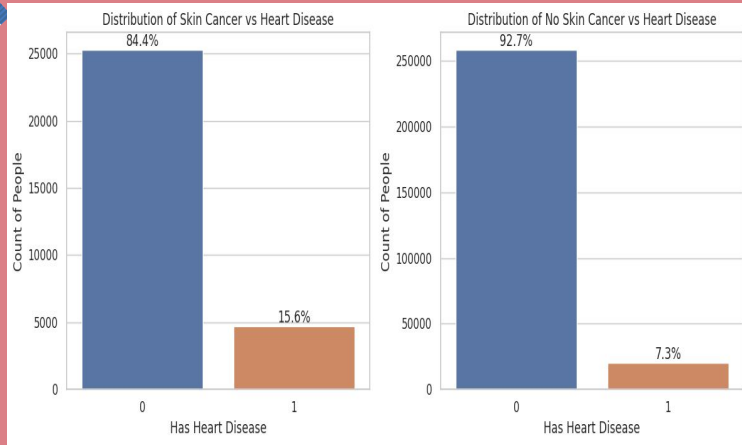
A strong correlation emerges between diabetes and heart disease, revealing a higher prevalence of 20.9% among patients with diabetes compared to 6.1% in those without diabetes, underscoring the notable impact of diabetes on the distribution of heart disease patients.



## 2. Heart Disease and Arthritis

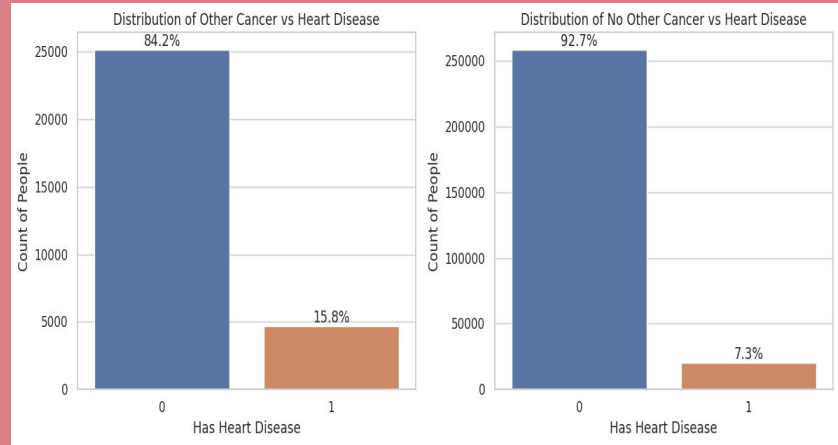
Patients with arthritis exhibit a higher heart disease rate of 14.1%, contrasting with a lower rate of 5.2% in those without arthritis, emphasizing the significant association between arthritis and the distribution of heart disease patients.

## 4. How having other diseases affect heart disease?



### 3. Heart Disease and Skin Cancer

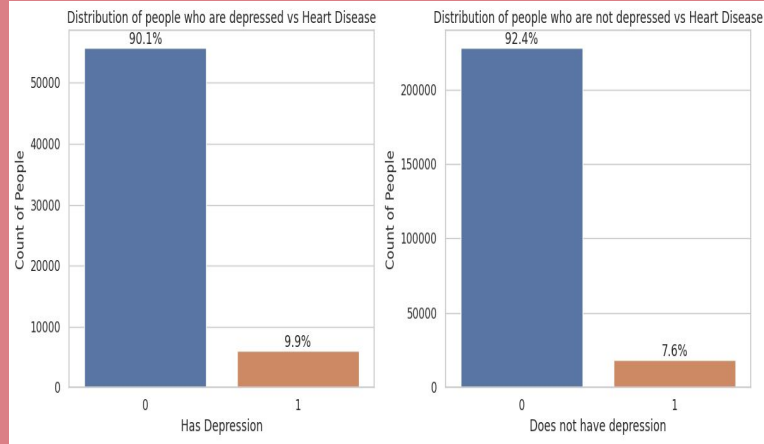
Patients having skin cancer showing a higher heart disease rate of 15.6%, compared to a lower rate of 7.3% in those without skin cancer. This emphasizes a potential association between skin cancer and the distribution of heart disease patients.



### 4. Heart Disease and Other Cancer

Patients with other cancer types display a higher heart disease rate of 15.8%, in contrast to a lower rate of 7.3% in those without other cancer. This underscores a potential link between other cancer conditions and the distribution of heart disease patients.

## 4. How having other diseases affect heart disease?



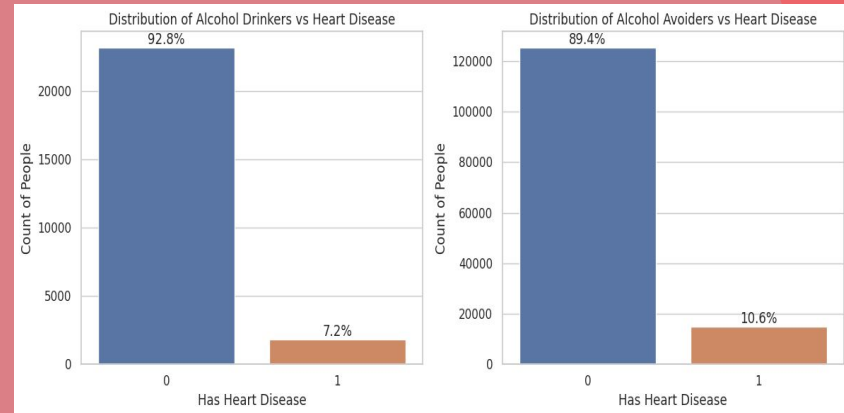
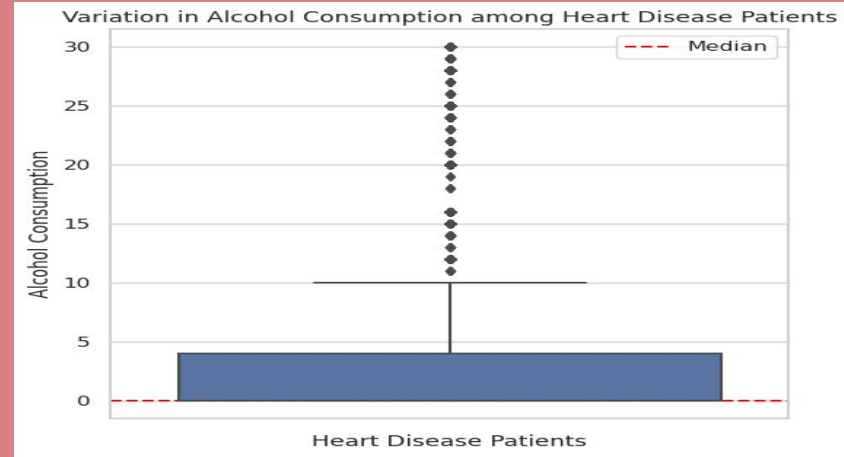
## 5. Heart Disease and Depression

Patients with depression exhibit a slightly higher heart disease rate of 9.9%, compared to a somewhat lower rate of 7.6% in those without depression, suggesting a potential connection between depression and the distribution of heart disease patients.

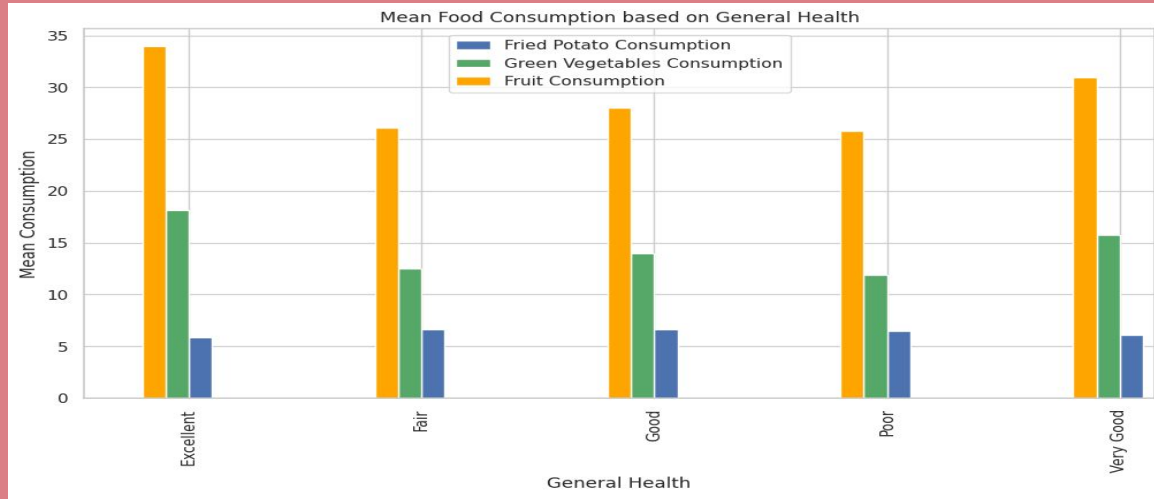


## 5. Impact of Alcohol Consumption on Heart Disease

- A negative correlation is evident in alcohol consumption among heart disease patients, with a higher rate of 10.6% for those avoiding alcohol, contrasting with a lower rate of 7.2% for those who consume alcohol. This highlights a potential inverse relationship between alcohol consumption and the prevalence of heart disease.
- After calculating the p-value, we find that the association between alcohol consumption and heart disease is statistically significant.



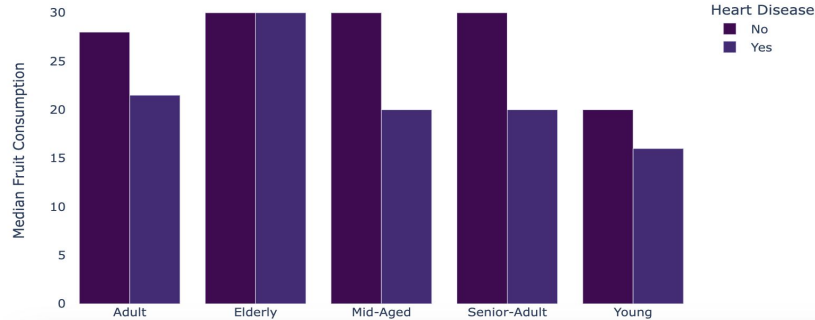
## 6. Mean Food Consumption Based on General Health



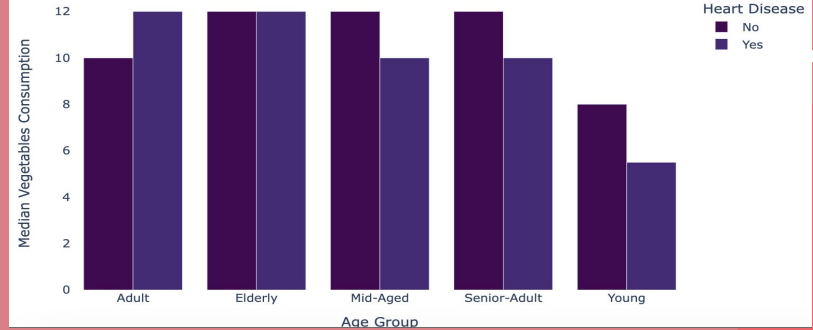
- So, before assessing the impact of Fruit, Fried potato and Green Vegetables consumption on Heart Disease we see what these consumption values tell us about the general health of the people.
- We observed that people who reported 'Excellent' and 'Very Good' general health had very high 'Fruit' and 'Green Vegetable' consumption levels as compared to those who reported 'Fair' and 'Poor' General health levels.
- Similarly, the people who reported 'Fair' and 'Poor' General health levels had high 'Fried Potato' consumption levels as compared to those who reported 'Excellent' and 'Very Good' general health.

# 7. Impact of Different Food habits on Heart Disease

Impact of Fruit Consumption on Heart Disease

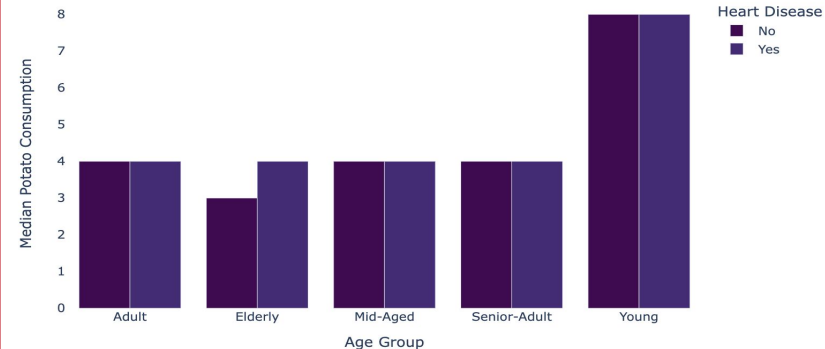


Impact of Vegetables Consumption on Heart Disease



Across all age groups, low fruit consumption is associated with a higher prevalence of heart diseases. In all age groups except adults, low vegetable consumption correlates with an increased risk of heart diseases. While heart rates remain consistent across age groups regardless of fried potato consumption, the elderly group stands out, showing a higher heart disease rate among those who consume more fried potatoes.

Impact of Potato Consumption on Heart Disease



3

# Modelling

Here, we will present details related to feature engineering, model selection, performance, as well as hyperparameter tuning





# Modelling – Feature Engineering

- The first step of the modelling process involved feature engineering, where we structure the data and add/remove features/rows.
- We started by performing one-hot encoding of the categorical variables, to create unique numerical representations, prevent misinterpretation as well as create independent, non-weighted entities. The next crucial step was to handle imbalance in the dataset, via synthetic minority oversampling.
- Next, we removed outliers from the resultant dataset to reduce overall noise, using The Interquartile Range (IQR) method to remove data from the lowest and highest percentile ranges by a threshold factor.





# Modelling – Feature Engineering

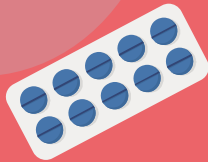
- The next step was dimensionality reduction using PCA. We decided to do this for several important reasons : adjusting the scale of the data (especially between encoded and unencoded features), reducing the number of features/ identifying principal features, reducing noise, as well as improving computational efficiency.
- After performing feature scaling and running PCA, we found 29/37 components that lied at the 80% variance threshold. However, after fitting the models to the scaled and reduced dimension training and test data, we found no notable improvement in accuracy (similar to other metrics). This could have happened because PCA effectiveness depends on the presence of linear relationships and on strong correlation in the data. Due to the lack of both of these characteristics in our dataset, PCA was not effective. Further, a few of our models excel at learning complex relationships themselves, thereby reducing the utility of PCA.





# Modelling – Model selection

- To effectively model the data, we made use of 4 modelling techniques to allow us to accurately represent the data.
- Since the problem intuitively corresponds to a binary classification problem, we made use of **Logistic Regression** as a baseline model. Its simplicity, Interpretability, efficiency, baseline performance, and benchmarking ability make it well suited for this purpose.
- Next, we modeled the data using a **Decision Tree Classifier**. There were several reasons for this: ease of use and interpretability, ability to capture non-linear patterns ( thereby being able to model complex relationships) and handle different data types, insensitivity to scale, computational efficiency and ease of implementation, as well as the fact that they are less prone to overfitting.





# Modelling – Model selection

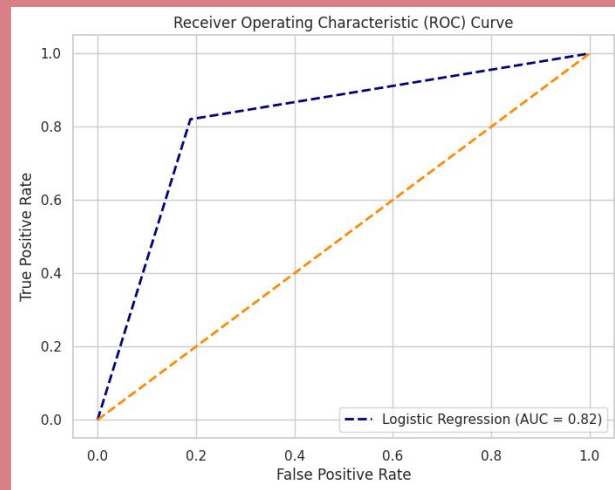
- The next logical step in model choice was to make use of a **Random Forests Classifier**, since it essentially spawns and combines multiple Decision Trees. They retain attributes of Decision Tree classifiers, and also have the benefit of not requiring data imputation, reduced sensitivity to hyperparameter tuning, and reduced overfitting tendency, since the ensemble averaging smoothens the noise in data, and improves generalizability.
- Finally, we also modeled the data using a **Neural Network**, due to its natural ability to capture hierarchical data relationships, represent data efficiently, as well as its ability to handle missing values gracefully. We endeavoured to make use of a FNN primarily because we were curious to understand how well it would capture relationships between the different features, especially when compared against more traditional models like the ones used earlier here.





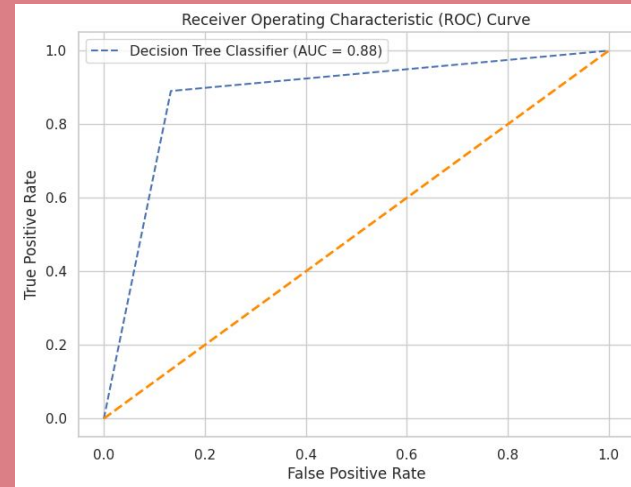
# Modelling – Performance

- We are pleased to report quite high performance with our pre-processing, feature engineering and choice of models.
- The **Logistic Regression model** gave us a baseline performance of 82% accuracy. This indicated to us that it is well suited for the dataset at hand, with sufficient room for improvement. Similar values were obtained for precision, recall, and F-score. Further, AUC-ROC metric (particularly well suited to ignoring class distribution) also gave us a score of 82%, indicating good starting trade-off between True Positive and False Positive Rates.



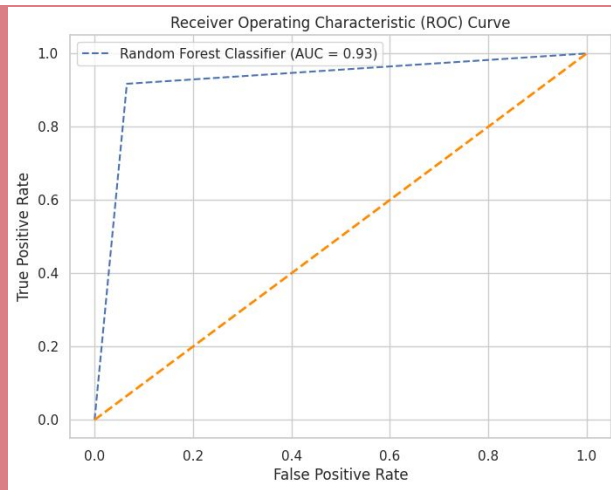
# Modelling – Performance

- The **Decision Tree model** gave us an accuracy of 88%. This indicated a considerable improvement over the baseline model, with further room for improvement. Similar values were obtained for precision, recall, and F-score. AUC-ROC metric gave us a score of 88%, indicating a very good performance in identifying True Positives and False Positives. The higher the skew of the curve to the top left, the better its performance.



# Modelling – Performance

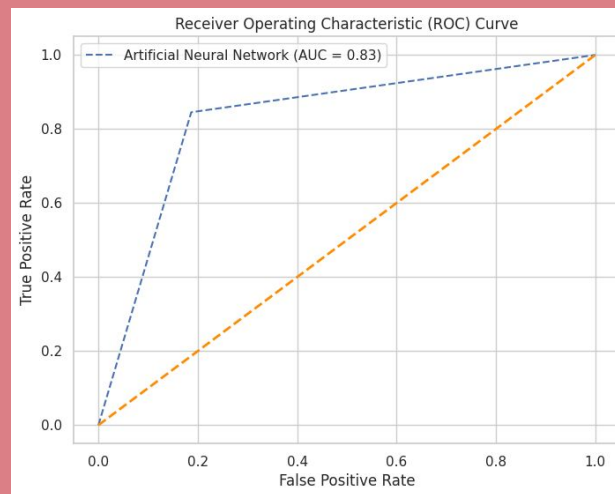
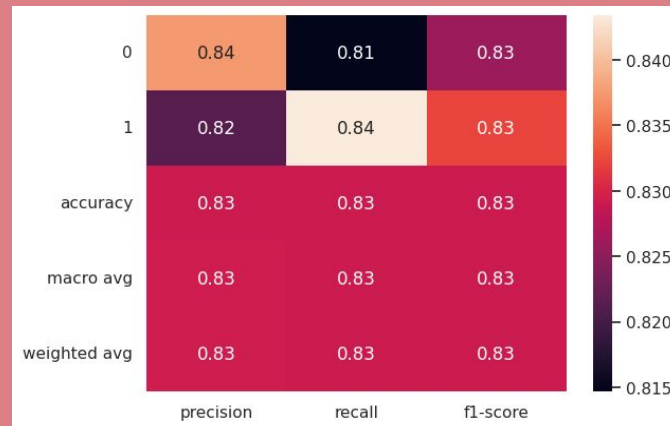
• Our next model, The **Random Forests Classifier**, gave us an accuracy of 93%. This was the highest possible accuracy we were able to achieve on this dataset, and is largely attributable to the characteristics of Random Forest models. Similar values were obtained for precision, recall, and F-score. AUC-ROC metric gave us a score of 93%, indicating an excellent performance in identifying True Positives and False Positives. The curve shown represents the highest the skew to the top left, highlighting the best performance.



# Modelling – Performance

Our final model, The **Full Neural Network**, gave us an accuracy of 83%, with similar values for other metrics.



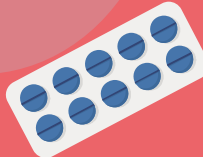
- While this does seem surprising, we suspect a few reasons for this. The size of the dataset is a factor, since Neural Nets tend to require large amounts of data to perform well. Random Forests, being an ensemble method, are more robust and interpretable (along with Decision Trees). Since the task at hand is solved adequately by simpler models, perhaps simpler models are the better choice.
- The AUC-ROC metric gave us a score of 83%, indicating moderate performance in identifying True Positives and False Positives, in line with Logistic Regression.






# Modelling – Hyper-Parameter tuning



- There are several important reasons to perform hyperparameter tuning : optimizing a model, generalization ability, avoiding underfitting/overfitting, improving convergence as well as improving model robustness. All of these reasons made us sought to perform hyperparameter tuning on a few of our models.
  - To start off, we wanted to understand if the baseline performance of our logistic regression could be improved. We did this via an automated search in scikit-learn, seeking to find the best hyperparameters for the Regularization hyperparameter (which controls overfitting), as well as the penalty hyperparameter (L1 or L2). After performing this search, we found that the optimized model remained at an accuracy of 82%, with similar values for other metrics, indicating that either default values were good to begin with, or that there was no room for improvement with Logistic Regression/ the dataset.
- 
- 
- 

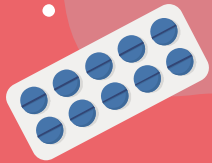


# Modelling – Hyper-Parameter tuning

- Next, we carried out hyperparameter tuning on the Decision Tree Model, to potentially improve upon its interpretability, generalizability, as well to control its depth and its minimum splitting samples.
  - We again made use of an automated search function to find the best hyperparameters, in this case by employing a grid search over all possible combinations of hyperparameters. After deriving the optimized model, we attempted to fit the refit the model on the data. We observed that there was no notable increase in the accuracy of 88% (similar to other metrics). This again indicates that perhaps the default parameters were already good enough for this model, or that there was no more room for improvement with this model/ the dataset.
- 

# Results and Conclusion

4



# Interpretability

Interpretability in ML models is the capacity to elucidate and make the model's predictions comprehensible and transparent to humans.





# Interpretability

It tends to decrease as model complexity increases.

1. Logistic Regression - Highly Interpretable
2. Decision Tree - Highly Interpretable. We can visually interpret in a graphical manner.
3. Random Forest - Random forests are very less interpretable compared to decision trees. Can still get feature importance scores.
4. Neural Networks - Generally considered as not interpretable / blackbox.

# Predictive Power or Performance

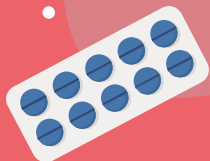
**Logistic Regression:** It is a simple and linear model. It performs well when the relationship between features and the target is approximately linear. In our case, it achieved an accuracy of 82%, which is a decent performance.

**Decision Tree:** It can capture non-linear relationships and interactions between features. It provided an accuracy of 88%, indicating an improvement over logistic regression. Decision trees are prone to overfitting, but pruning techniques can help mitigate this issue.

**Random Forest:** Being an ensemble of decision trees, often provide improved performance by reducing overfitting. The accuracy of 93% suggests that it's effective in capturing complex patterns and generalizing well to new data.

**Neural Networks:** Can learn intricate patterns and representations from the data. Achieving an accuracy of 83%, which would increase with more data, indicates that the neural network is highly effective for our dataset. However, it comes with the cost of increased complexity/low interpretability.

**We need Interpretability over  
Performance**



# Why ?

**Clinical Decision** -: We would value models that provide clear insights into the variables influencing heart disease. Interpretability would be key for making these informed clinical decisions.

**Feature Importance for Prevention** - In the realm of heart disease prevention, knowing which attributes contribute significantly is paramount. (Logistic Regression, Decision Trees, Random Forest.)

**Patient Communication** - Explaining the risk factors to patients is an integral part, and depends a lot on interpretability.

**Regulatory Compliance and Ethical Considerations** - In the healthcare industry, compliance with regulations and ethical considerations is paramount. Transparent models not only facilitate adherence to regulatory standards but also contribute to ethical machine learning practices.

**Collaboration with Multidisciplinary Teams** - Heart disease analysis involves collaboration among diverse healthcare professionals, including doctors, nurses, and data scientists. Interpretability, facilitates effective communication across disciplines.

# Challenges – Limitations – Future work

- 1. Limited Demographic Insights** - Analysis can be further enhanced with more detailed demographic data to explore regional variations in health and lifestyle patterns affecting heart disease rates.
- 2. Neural Network and Data Limitations** - The accuracy for the neural nets can be further improved by acquiring a larger dataset, unlocking its full potential for more precise predictions.
- 3. Addressing Potential Bias** - Potential biases can be recognized and mitigated in the dataset, ensuring model predictions are more representative and applicable across diverse populations.
- 4. Temporal Analysis** - A Temporal Analysis can be performed to understand how heart disease rates have evolved over time. This could involve examining trends, seasonal variations, or identifying any significant changes in risk factors.
- 5. Health Intervention Strategies** - There is a potential for recommending health intervention strategies based on the analysis. We may identify actionable recommendations for individuals or communities to reduce heart disease risk, considering cultural, economic, and lifestyle factors.



**THANK YOU !**