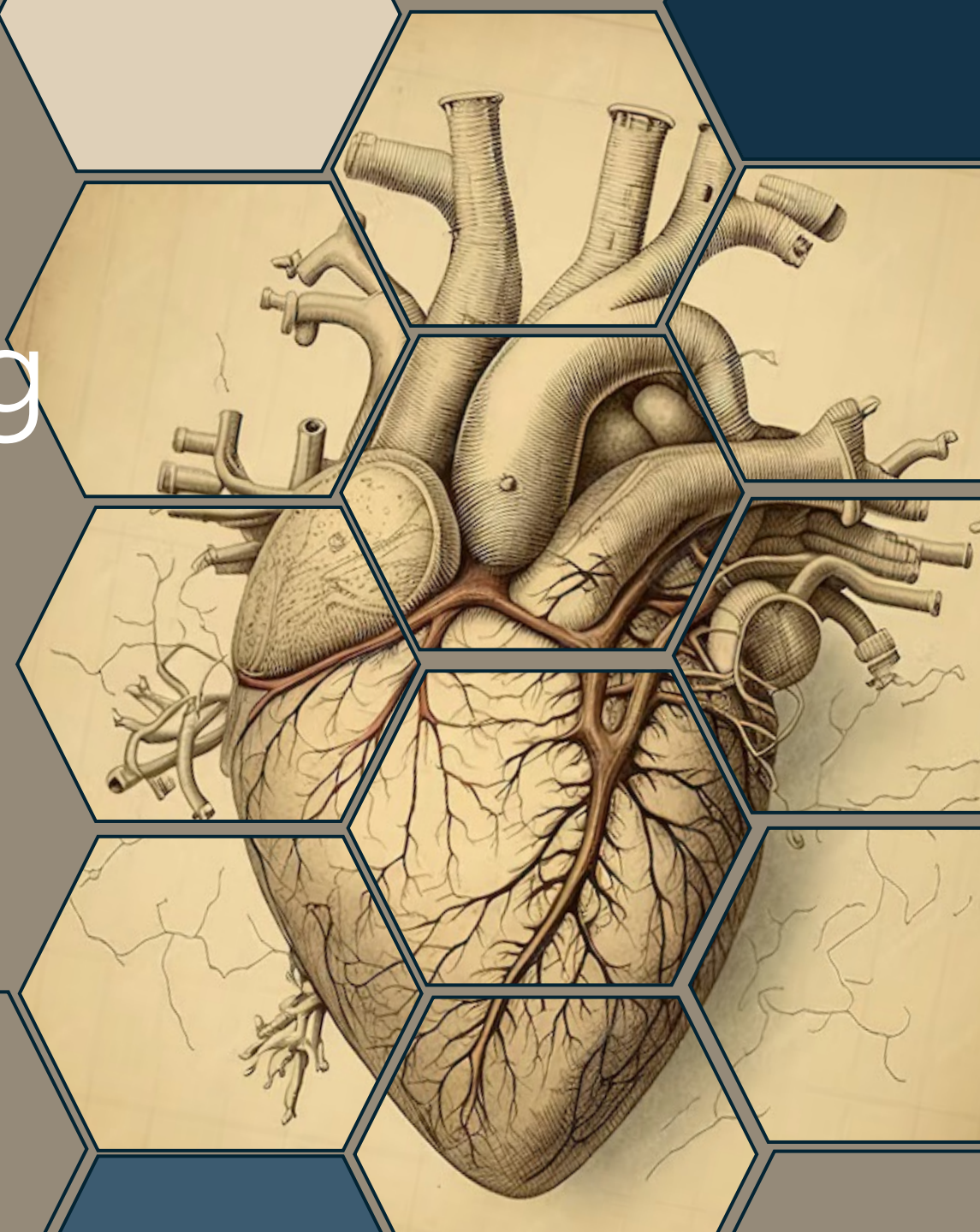
Predictive Modelling For **Cardiovascular Disease** Risk Assessment



1

Introduction

Medical professionals have dedicated enormous effort to the treatment and diagnosis of heart disease. Recognizing the potential impact on humanity.

2

Global Statistics

Total 17 millions premature Deaths Worldwide in 2019. 38% were caused by the CVD. This cause for 45% of deaths in 40-69 years age group.

3

Indian Statistics

In 2016, INDIA reported 63% of total deaths due to NCD. Of which 27% were attributed to CVD.

1

Introduction

Medical professionals have dedicated enormous effort to the treatment and diagnosis of heart disease. Recognizing the potential impact on humanity.

2

Global Statistics

Total 17 millions premature Deaths Worldwide in 2019. 38% were caused by the CVD. This cause for 45% of deaths in 40-69 years age group.

3

Indian Statistics

In 2016, INDIA reported 63% of total deaths due to NCD. Of which 27% were attributed to CVD.

2

Global Statistics

Total 17 millions premature Deaths Worldwide in 2019. 38% were caused by the CVD. This cause for 45% of deaths in 40-69 years age group.

3

Indian Statistics

In 2016, INDIA reported 63% of total deaths due to NCD. Of which 27% were attributed to CVD.

1

Introduction

Medical professionals have dedicated enormous effort to the treatment and diagnosis of heart disease. Recognizing the potential impact on humanity.

Scope and Objective

A

Investigate how lifestyle choices impact heart health and explore ways to reduce the risk of cardiovascular disease (CVD).

B

Evaluate the imp. of early detection in preventing severe outcomes related to CVD.

C

Develop and test statistical models to identify individuals at high risk of developing CVD, aiding in timely intervention and prevention.

D

Provide actionable insights for healthcare prof. and policymakers to make strategies for reducing the global burden of CVD.

Scope and Objective

A

Investigate how lifestyle choices impact heart health and explore ways to reduce the risk of cardiovascular disease (CVD).

B

Evaluate the imp. of early detection in preventing severe outcomes related to CVD.

C

Develop and test statistical models to identify individuals at high risk of developing CVD, aiding in timely intervention and prevention.

D

Provide actionable insights for healthcare prof. and policymakers to make strategies for reducing the global burden of CVD.

Scope and Objective

A

Investigate how lifestyle choices impact heart health and explore ways to reduce the risk of cardiovascular disease (CVD).

B

Evaluate the imp. of early detection in preventing severe outcomes related to CVD.

C

Develop and test statistical models to identify individuals at high risk of developing CVD, aiding in timely intervention and prevention.

D

Provide actionable insights for healthcare prof. and policymakers to make strategies for reducing the global burden of CVD.

Scope and Objective

A

Investigate how lifestyle choices impact heart health and explore ways to reduce the risk of cardiovascular disease (CVD).

B

Evaluate the imp. of early detection in preventing severe outcomes related to CVD.

C

Develop and test statistical models to identify individuals at high risk of developing CVD, aiding in timely intervention and prevention.

D

Provide actionable insights for healthcare prof. and policymakers to make strategies for reducing the global burden of CVD.

Logistic Regression **Assumptions**

01

02

03

04

Logistic Regression

01

02

03

04

Assumption

01

Logistic Regression doesn't Presuppose any linear relationship between the predictor and the response variable.

Logistic Regression

01

02

03

04

Assumption 02

The response variable should exhibit a dichotomous nature, meaning it can be categorised into two distinct groups

Logistic Regression

01

02

03

04

Assumption 03

Categories should be both mutually exclusive and exhaustive, meaning that each case can only belong to one group and every case must be a member of one of the groups .

Logistic Regression

01

02

03

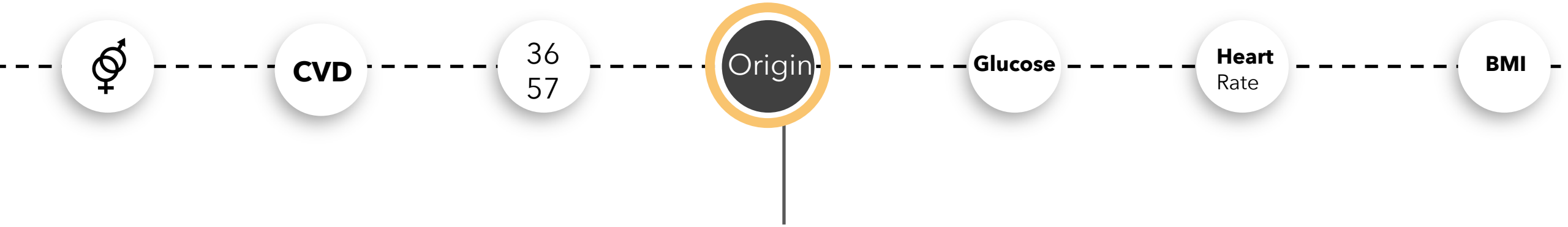
04

Assumption

04

In contrast to linear regression, Logistic Regression often demands larger sample sizes due to the reliance on maximum likelihood estimation. Which is based on approximations derived from large sample theory.

GETTING TO KNOW THE **DATASET**

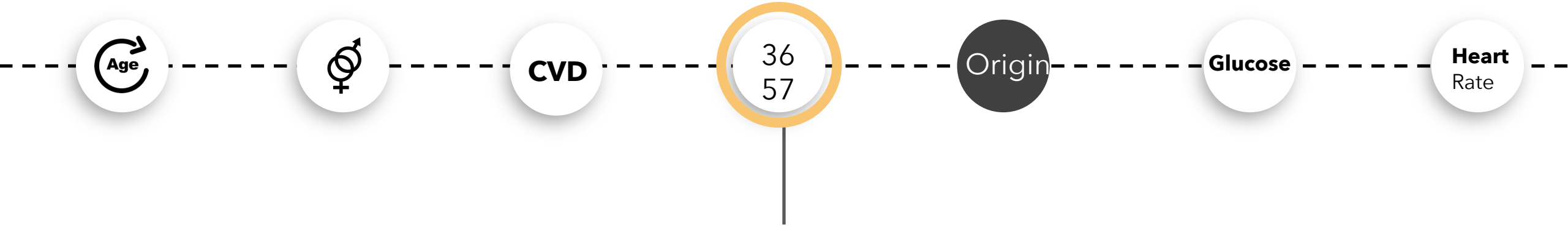


No of Datapoints
In the Dataset, there are in total 3,657 data points available encompassing various demographic, behavioural, and medical risk factors.

Origin Of the Dataset
This Cardiovascular Disease Dataset Dataset Was collected from the UCI Machine Learning Library

No of Datapoints
In the Dataset, there are in total 3,657 data points available encompassing various demographic, behavioural, and medical risk factors.

GETTING TO KNOW THE **DATASET**



Origin Of the Dataset

This Cardiovascular Disease Dataset Dataset Was collected from the UCI Machine Learning Library

No of Datapoints

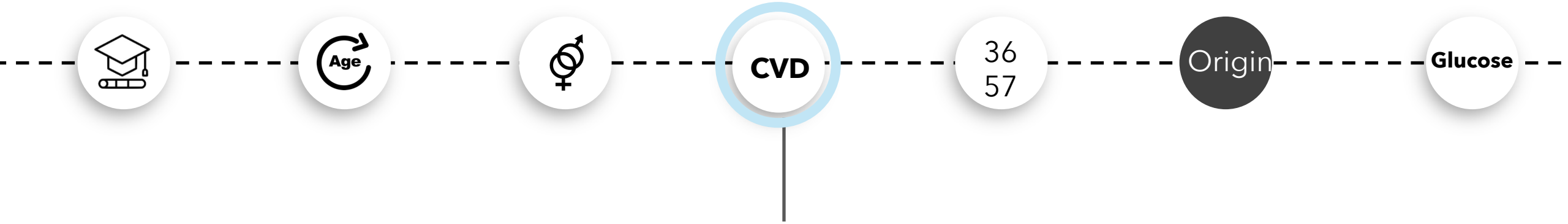
In the Dataset, there are in total 3,657 data points available encompassing various demographic, behavioural, and medical risk factors.

CHD - 10 Year Risk

Binary Variables
Binary Variable indicating the presence (1) or absence(0) of CHD risk.

Response
Variable

GETTING TO KNOW THE **DATASET**



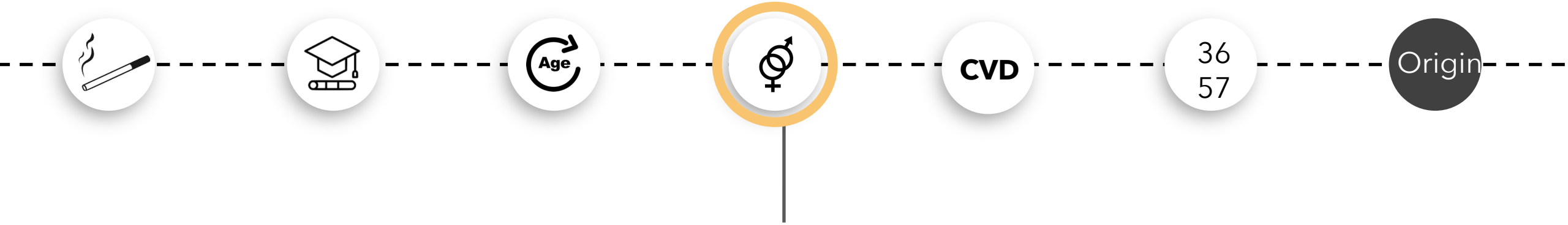
No of Datapoints
In the Dataset, there are in total 3,657 data points available encompassing various demographic, behavioural, and medical risk factors.

CHD - 10 Year Risk
Binary Variables
Binary Variable indicating the presence (1) or absence(0) of CHD risk.

Sex
Categorical (Nominal)
0 = Female , 1 = male

Demographic
Factor

GETTING TO KNOW THE **DATASET**



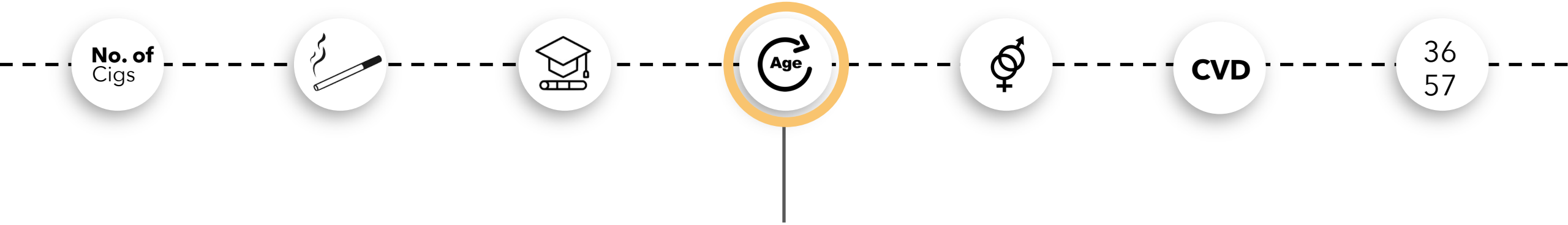
CHD - 10 Year Risk
Binary Variables
Binary Variable indicating the presence (1) or absence(0) of CHD risk.

Sex
Categorical (Nominal)
0 = Female , 1 = male

Age
Continuous
Age in years

Demographic
Factor

GETTING TO KNOW THE **DATASET**



Sex
Categorical (Nominal)
0 = Female , 1 = male

Age
Continuous
Age in years

Education
Categorical (Nominal)
Denotes the Education Level
of the individual.
1 : Pre-College Level
2 : Undergrad Level
3 : Postgrad Level

Demographic
Factor

GETTING TO KNOW THE **DATASET**

BP | No. of Cigs |  |  | Age |  | CVD

Age
Continuous
Age in years

Education
Categorical (Nominal)
Denotes the Education Level of the individual.
1 : Pre-College Level
2 : Undergrad Level
3 : Postgrad Level

Current Smoker
Categorical (Nominal)
0 = Don't Smoke , 1 = Smoke

Behavioral
Factor

GETTING TO KNOW THE **DATASET**

STROKE — BP — No. of Cigs —  — Age

Education
Categorical (Nominal)
Denotes the Education Level of the individual.
1 : Pre-College Level
2 : Undergrad Level
3 : Postgrad Level

Current Smoker
Categorical (Nominal)
0 = Don't Smoke , 1 = Smoke

Cigs. Per Day
Categorical (Nominal)
0 = Don't Smoke , 1 = Smoke

Behavioral
Factor

GETTING TO KNOW THE **DATASET**

HYP — STROKE — BP — No. of Cigs —  —  — Age

Current Smoker
Categorical (Nominal)
0 = Don't Smoke , 1 = Smoke

Cigs. Per Day
Categorical (Nominal)
0 = Don't Smoke , 1 = Smoke

BP Meds
Categorical (Nominal)
Specifies if the individual was using medication for blood pressure management..[0 = Not Diabetic , 1 = Diabetic]

Medical
History

GETTING TO KNOW THE **DATASET**

Diabetes — **HYP** — **STROKE** — **BP** — **No. of** Cigs —  — 

Cigs. Per Day
Categorical (Nominal)
0 = Don't Smoke , 1 = Smoke

BP Meds
Categorical (Nominal)
Specifies if the individual was using medication for blood pressure management..[0 = Not Diabetic , 1 = Diabetic]

Prevalent Stroke
Categorical (Nominal)
Identifies whether the patient has previously experienced a stroke. [0 = Haven't , 1 = Have experienced]

Medical
History

GETTING TO KNOW THE **DATASET**

Chol. — Diabetes — HYP — STROKE — BP — No. of Cigs — 

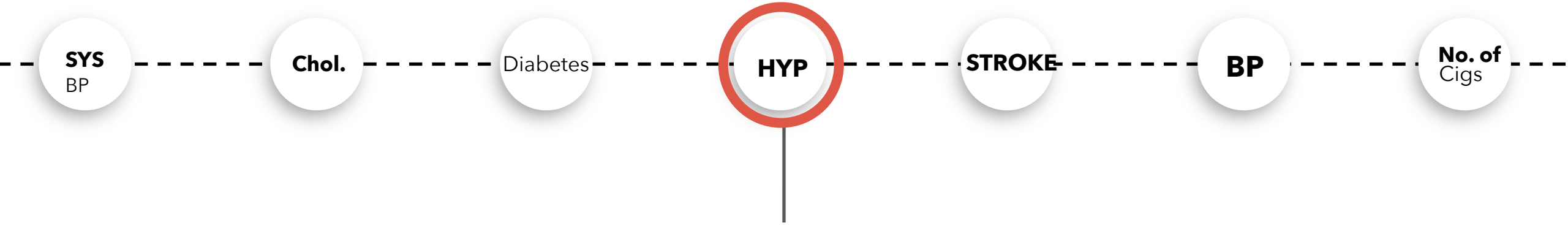
BP Meds
Categorical (Nominal)
Specifies if the individual was using medication for blood pressure management..[0 = Not Diabetic , 1 = Diabetic]

Prevalent Stroke
Categorical (Nominal)
Identifies whether the patient has previously experienced a stroke. [0 = Haven't , 1 = Have experienced]

Prevalent Hypersensitivity
Categorical (Nominal)
Indicates whether the patient is hypertensive [0 =Patient isn't Hypersensitive,1=Patient is Hypersensitive]

Medical
History

GETTING TO KNOW THE **DATASET**



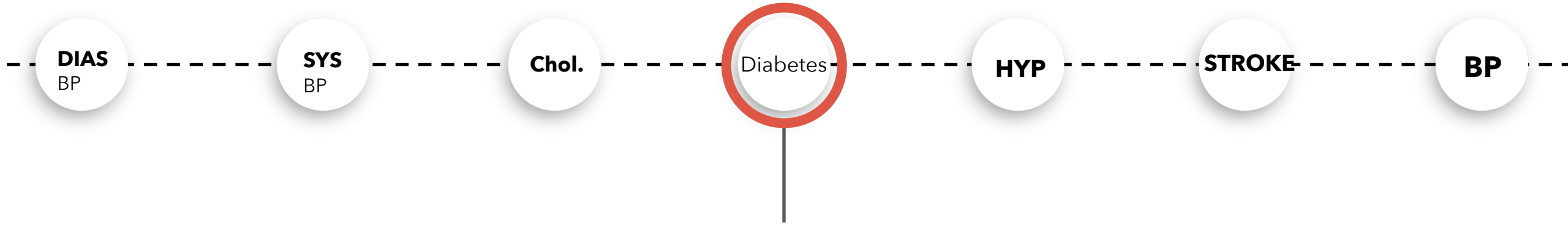
Prevalent Stroke
Categorical (Nominal)
Identifies whether the patient has previously experienced a stroke. [0 = Haven't , 1 = Have experienced]

Prevalent Hypersensitivity
Categorical (Nominal)
Indicates whether the patient is hypertensive [0 =Patient isn't Hypersensitive,1=Patient is Hypersensitive]

Diabetes
Categorical (Nominal)
Identifies whether the patient has diabetes. [0 = Patient isn't Diabetes , 1 = Patient is Diabetic]

Medical
History

GETTING TO KNOW THE **DATASET**



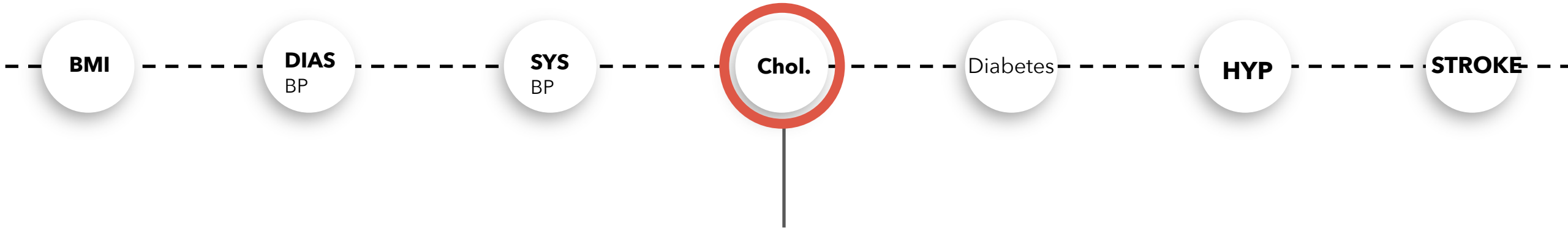
Prevalent Hypersensitivity
Categorical (Nominal)
Indicates whether the patient is hypertensive [0 =Patient isn't Hypersensitive, 1 =Patient is Hypersensitive]

Diabetes
Categorical (Nominal)
Identifies whether the patient has diabetes. [0 = Patient isn't Diabetes , 1 = Patient is Diabetic]

Total Cholesterol
Continuous
Represents the total blood cholesterol level.

Medical
History

GETTING TO KNOW THE **DATASET**



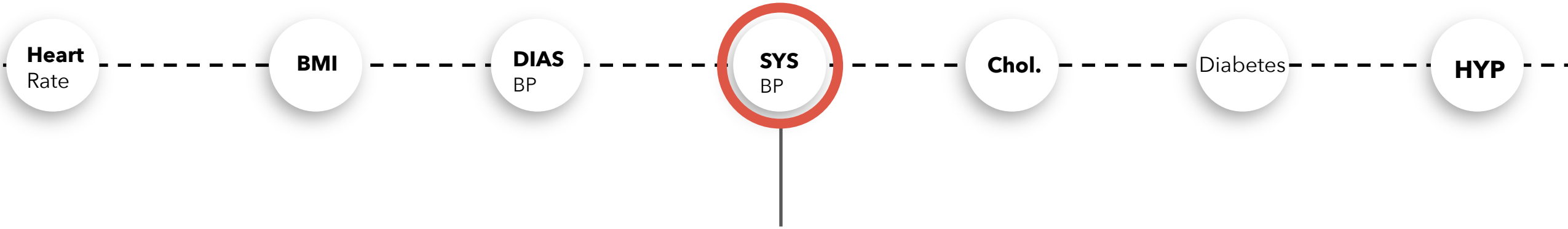
Diabetes
Categorical (Nominal)
Identifies whether the patient has diabetes. [0 = Patient isn't Diabetes , 1 = Patient is Diabetic]

Total Cholesterol
Continuous
Represents the total blood cholesterol level.

Sys BP
Continuous
Represents the systolic blood pressure level.

Medical
History

GETTING TO KNOW THE **DATASET**



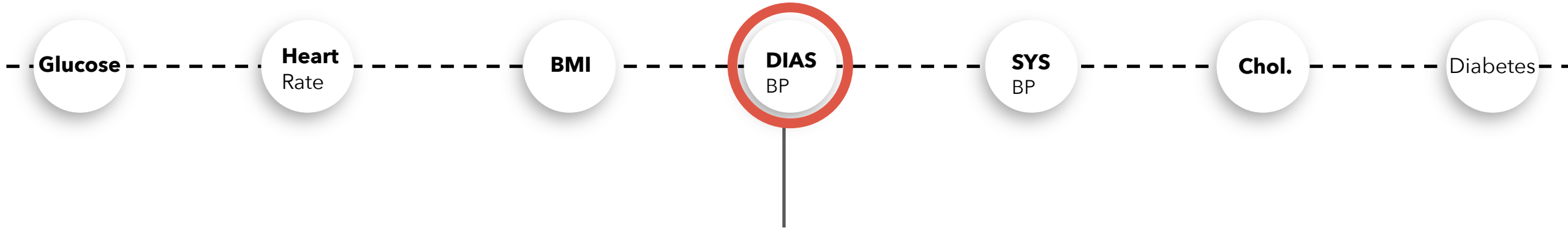
Total Cholesterol
Continuous
Represents the total blood cholesterol level.

Sys BP
Continuous
Represents the systolic blood pressure level.

Dia BP
Continuous
Represents the Diastolic blood pressure level

Medical
History

GETTING TO KNOW THE **DATASET**



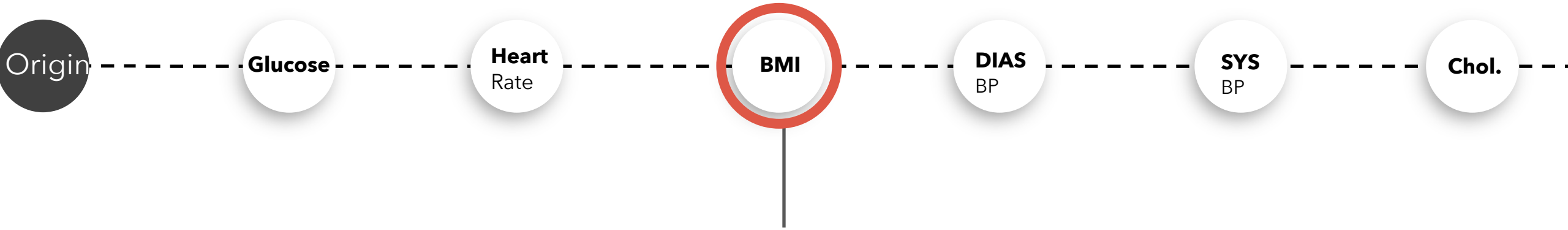
Sys BP
Continuous
Represents the systolic blood pressure level.

Dia BP
Continuous
Represents the Diastolic blood pressure level

BMI
Continuous
Represents the individual's Body Mass Index

Medical
History

GETTING TO KNOW THE **DATASET**



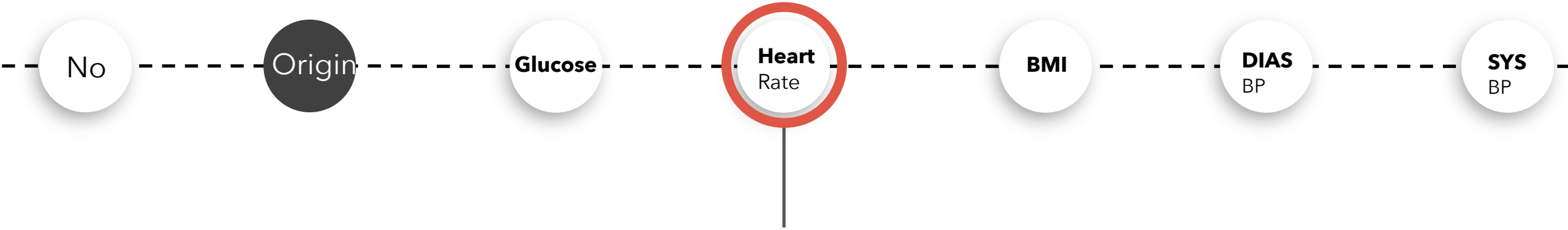
Dia BP
Continuous
Represents the Diastolic blood pressure level

BMI
Continuous
Represents the individual's Body Mass Index

Heart Rate
Continuous
Represents the heart rate treated as continuous due to the large number of possible values.

Medical
History

GETTING TO KNOW THE **DATASET**



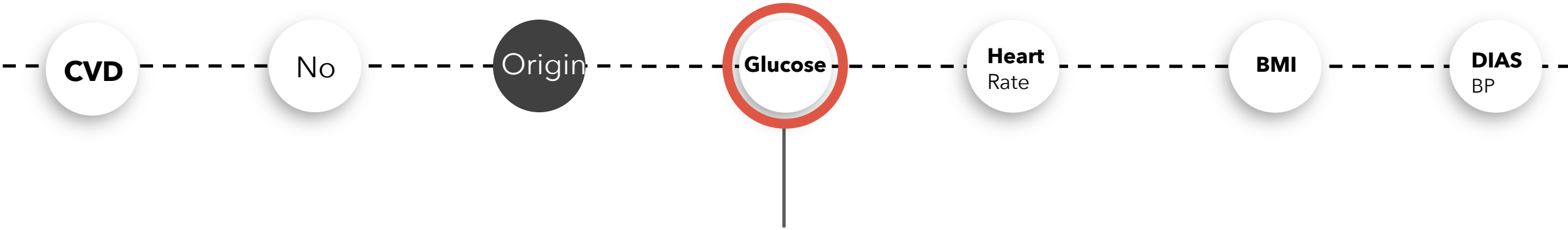
BMI
Continuous
Represents the individual's Body Mass Index

Heart Rate
Continuous
Represents the heart rate treated as continuous due to the large number of possible values.

Glucose
Continuous
Represents the heart rate treated as continuous due to the large number of possible values.

Medical
History

GETTING TO KNOW THE **DATASET**



Heart Rate
Continuous
Represents the heart rate treated as continuous due to the large number of possible values.

Glucose
Continuous
Represents the heart rate treated as continuous due to the large number of possible values.

No of Datapoints
In the Dataset, there are in total 3,657 data points available encompassing various demographic, behavioural, and medical risk factors.

Briefly the Complete **DATASET**

Denoted	Covariate	Nature
x_1	Sex	Categorical
x_2	Age	Continuous
x_3	Education	Categorical
x_4	Smoker	Binomial
x_5	Cigs/Day	Discrete

Denoted	Covariate	Nature
x_6	**Bp** Meds	Categorical
x_7	**Prev.** Stroke	Categorical
x_8	**Prev.** Hyp	Categorical
x_9	Diabetes	Categorical
x_{10}	**Total** Cholesterol	Continuous

Denoted	Covariate	Nature
x_{11}	**Sys** BP	Continuous
x_{12}	**Dias** BP	Continuous
x_{13}	BMI	Continuous
x_{14}	**Heart** Rate	Continuous
x_{15}	Glucose	Continuous

01

02

03

04

Methodologies Used for Logistic

Regression

02

03

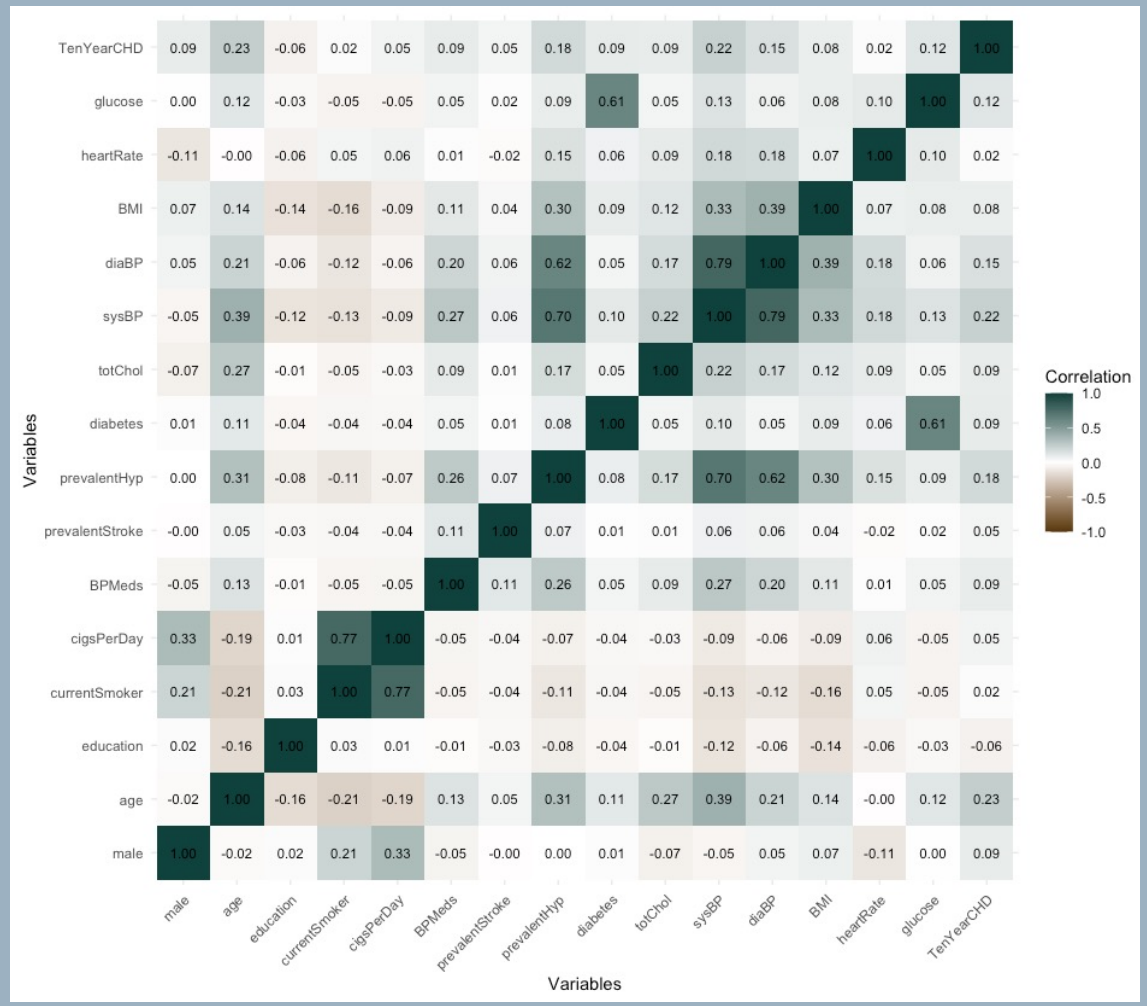
04

01 Checking for **Multicollinearity**

01

Multicollinearity in linear models occurs when predictors are not independent, destabilizing coefficient interpretability. It's detected when the model's R^2 is high (over 0.8), but few predictors are significant, or when pairwise correlations between predictors, assessed using Karl Pearson's coefficient, exceed 0.8. Tools like VIF and correlation heatmaps refine these assessments by visualizing interdependencies. To ensure dataset integrity for linear modelling, a detailed quantitative follow-up analysis is essential.

02 Correlation **Heatmap**



01 Methodologies

02 Logistic **Regression**

03

04

03 Test Significance of **Parameters**

For testing the significance of the parameters, we'll be using wald's t test statistics Now suppose we want to test whether x_i is significant or not. We're to test

H_0 : Covariate x_i is significant

H_1 : Not Significant ; for, $i = 1(1)k$

This is also similar to

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

Test Statistic for testing the significance of the predictor variables then given by

$$t_j = \frac{\beta_j - E(\beta_j)}{S.E(\beta_j)} \sim N(0,1)$$

[Under H0] ; for, $i = 1(1)k$

We'll be rejecting the Null hypothesis and say that the j^{th} covariateis not significant if

$$|\tau_J| > \tau_{\frac{\alpha}{2}}$$

01 Methodologies

02 Logistic **Regression**

03

04

Wald's Test for **Significance**

```
> data.frame(name,walds_t,significance)
             name    walds_t    significance
1            male  1.41443413   Significant
2             age  0.69723314   Significant
3       education -0.17188051   Significant
4    currentSmoker  0.27868337   Significant
5       cigsPerDay  0.17486239   Significant
6           BPMeds  0.31498194   Significant
7  prevalentStroke  0.77475590   Significant
8     prevalentHyp  0.64423554   Significant
9         diabetes  0.32965284   Significant
10         totChol  0.09111104   Significant
11           sysBP  0.21795368   Significant
12           diaBP -0.02802294   Significant
13             BMI  0.09083344   Significant
14       heartRate -0.04796687   Significant
15         glucose  0.12471948   Significant
```

01

02

03

03.1

04

Methodologies

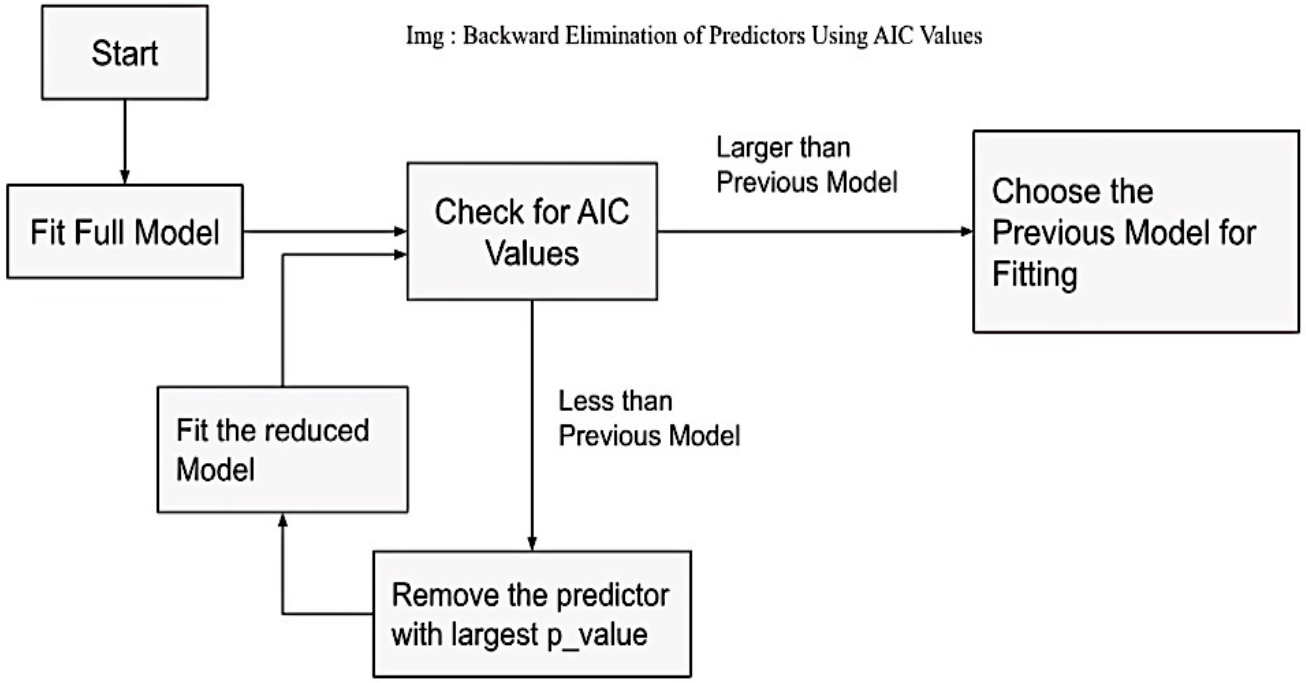
Logistic **Regression**

04 Fitting of Logistic Model

Considering Backward Elimination as the selected stepwise procedure, the initial model encompasses all variables. Utilizing the glm() function in R, I've conducted logistic regression to analyze the association between the binary response variable indicating the presence of heart disease and the fifteen distinct independent predictors [Placed in earlier Slides]. The fitting process is outlined in the code as follows:

```
fullmodel = glm(TenYearCHD ~ .,
data = training_data, family =
binomial(link = "logit"))
```

Backward Elimination **Method**



Img : Backward Elimination of Predictors Using AIC Values

After fitting the full regression model , we go for **Backward Elimination** method. Backward elimination is a method used to identify the independent variables that contribute to the **best regression** model. It is a component of stepwise regression, a feature selection technique employed in ML model development. Backward elimination is a systematic process that begins with a full set of features (i.e full logistic regression Model) and iteratively removes variables until the model's performance is optimized. In ML, there are several other approaches as well.

Best Fitted **Model**

The distribution of Y is specified by probabilities of success $P(Y = 1) = \pi$ and the probabilities of failure
$P(Y = 0) = 1 - \pi$. $\mathrm{E}(\mathrm{Y}) = \pi$
After fitting the Backward Elimination method we get the multiple logistic regression with binary response is given by :
$Y_i \sim Bernoulli(\pi_i)$; where π_i =
$P(Y_i = 1|X_{1i}, X_{2i} \ . \ . \ . \ X_{7i}) = \frac{e^{n_i}}{1+e^{n_i}}$
Here, η_i is a function of all the predictor variables
$n_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{3i} + \ . \ . + \beta_7 x_{7i}$

```
Call:
glm(formula = TenYearCHD ~ ., family = binomial(link = "logit"),
    data = reduced_data_8)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.732172   0.602685 -14.489  < 2e-16 ***
male         0.502872   0.122016   4.121 3.77e-05 ***
age          0.062626   0.007304   8.574  < 2e-16 ***
cigsPerDay   0.017669   0.004796   3.684 0.000229 ***
prevalentHyp 0.268576   0.153346   1.751 0.079871 .
totChol      0.003190   0.001271   2.511 0.012046 *
sysBP        0.013961   0.003193   4.372 1.23e-05 ***
glucose      0.007262   0.001912   3.798 0.000146 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2379  on 2742  degrees of freedom
Residual deviance: 2112  on 2735  degrees of freedom
AIC: 2128

Number of Fisher Scoring iterations: 5
```

Odds **Ratio**

In short, **OR**(Odds Ratio)

The **Odds Ratio** (OR) is one of the statistical measurements used in clinical research and decision-making. The Odds of the Response variable equating a case (given some linear combination x of the predictors) is equivalent to the exponential function of the linear regression.

Let's consider a binary predictor variable which has values either 0 or 1.

For example, x denote a binary predictor variable defined as follows

$x = 0$ when it is FALSE

$x = 1$when it is TRUE,

and Y denotes the **response variable**.

Then $\frac{P(Y=1|x)}{P(Y=0|x)} = Odds\ of\ \ Y = 1\ against\ \ Y = 0$ for given

Given value of x

So, we get that $O_1 = \frac{P(Y=1|X=0)}{P(Y=0|X=0)}$

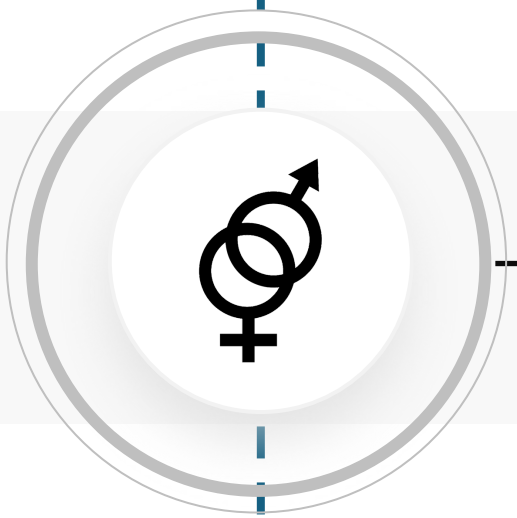
And $O_2 = \frac{P(Y=1|X=1)}{P(Y=0|X=1)}$

Odds Ratio = $\frac{O_1}{O_2}$

```
>                    Odds_Ratio
                  exp(estimates)
male                    1.653463
age                     1.064628
cigsPerDay              1.017826
prevalentHyp            1.308100
totChol                 1.003195
sysBP                   1.014059
glucose                 1.007288
>
```

Demographic
Factor

Interpretation of **Odds Ratio**



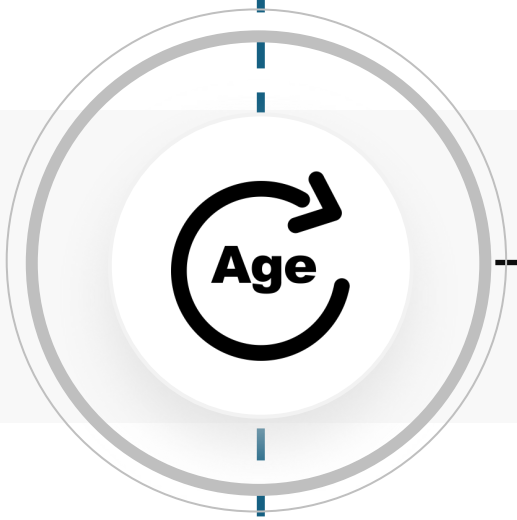
Male

`1.653463`

In this model, with all other factors controlled, males [male=1] have a 1.653 times higher odds of being diagnosed with heart disease compared to females [male=0]. This suggests a 65.34% greater likelihood of heart disease among males relative to females

Demographic
Factor

Interpretation of **Odds Ratio**



Age

`1.064628`

The coefficient for age says that, holding all others constant, we will see a 6.5% increase in the odds of getting diagnosed with CDH for a one year increase in age since exp(0.062625) = 1.064628.

Behavioral
Factor

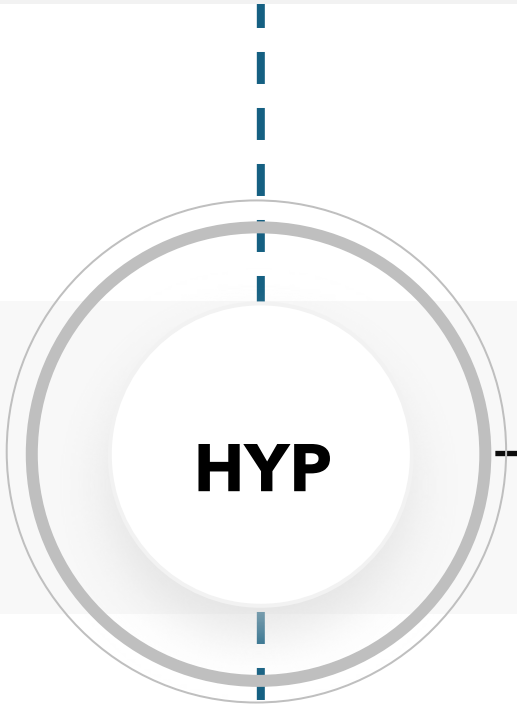
Interpretation of **Odds Ratio**

Cigarettes Per Day

`1.017826`

For each additional cigarette smoked per day, the odds of getting diagnosed with heart disease increase by approximately 1.8%

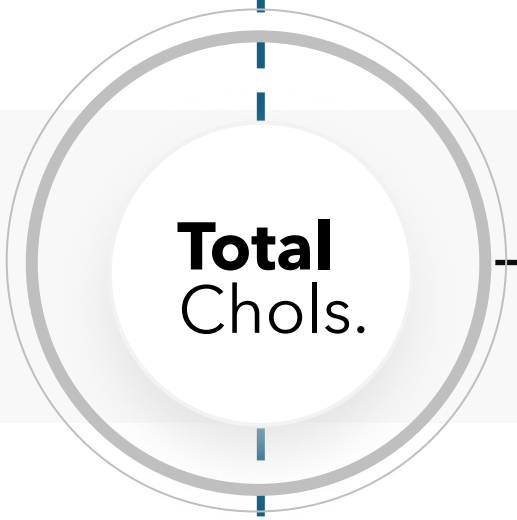
Interpretation of **Odds Ratio**



Prevalent Hypersensitive
`1.308100`
Individuals with prevalent hypertension have approximately 30.81% higher odds of being diagnosed with heart disease compared to those without prevalent hypertension.

Medical
History

Interpretation of **Odds Ratio**

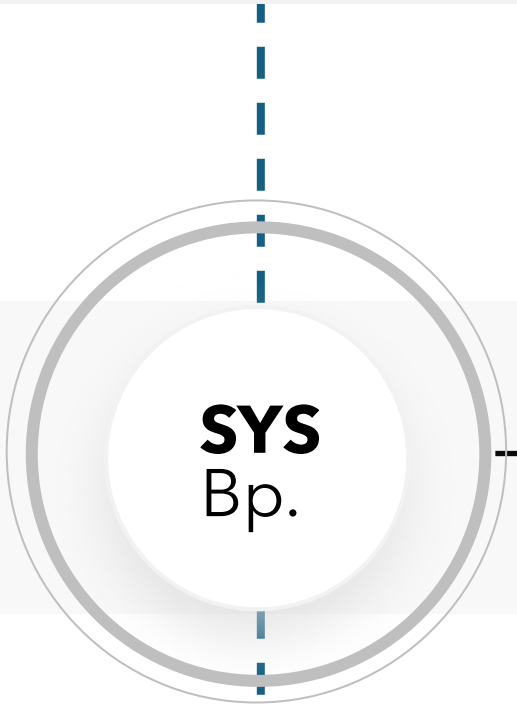


Cholesterol

`1.003195`

With each unit increase in total cholesterol level, the odds of being diagnosed with heart disease increase by approximately 0.32%

Interpretation of **Odds Ratio**



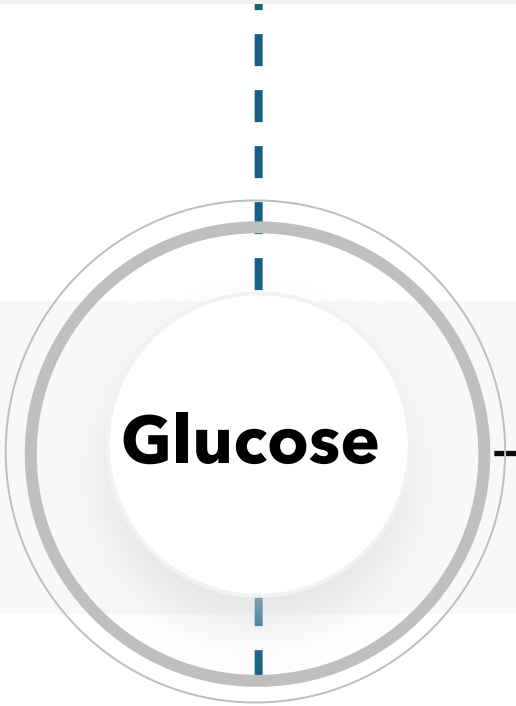
Sys Blood Pressure

`1.014059`

For each unit increase in systolic blood pressure, the odds of being diagnosed with heart disease increase by approximately 1.41%.

Medical
History

Interpretation of **Odds Ratio**



Glucose Level

`1.007288`

With each unit increase in glucose level, the odds of being diagnosed with heart disease increase by approximately 0.73%.

Classification **Statistics**

Before we dive into the Classification Probabilities, let's arm ourselves with some key terms to navigate the terrain more smoothly..

True Positive (TP)

If the predicted value and the actual value are both positive.

True Negative (TP)

If the predicted value and the actual value are both negative.

False Positive (FP)

If the predicted value is positive, but the actual value is negative. *(Type I Error)*

False Negative (FN) :

If the predicted value is negative, but the actual value is positive . *(Type II Error)*

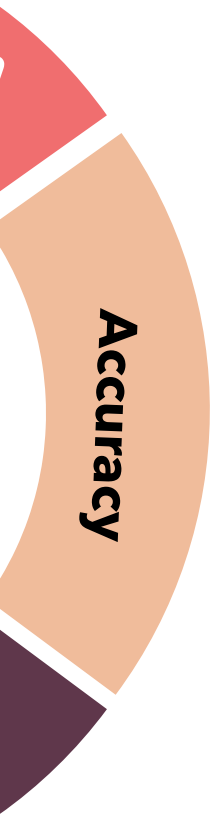
The fitted model is considered as an accurate representation of the data set if the true positives and negatives are maximized, and the false positives and false negatives are maximized. Accuracy can be calculated using the confusion matrix.

For this regression model , taking the Optimum threshold as 0.2977528, the predicted value of the response is calculated as

$$Y = \begin{cases} \hat{Y} = 1 \,; if\ \hat{\pi} > 0.29777528 \\ \hat{Y} = 0 \,; elsewhere \end{cases}$$

	Predicted (No)	**Predicted (Yes)**
Actual (No)	True Negative (TN)	False Positive (FP)
Actual (Yes)	False Negative (FN)	True Positive (TP)

Classification
Statistics



Accuracy

Accuracy provides a detailed summary of the information conveyed by the confusion matrix. It is computed by dividing the sum of true positives and true negatives by the total number of observations, offering a measure of the fitted model's overall correctness. It is measures as follows .

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

For this chosen logistic regression model ,

$$Accuracy = \frac{2308 + 18}{2308 + 411 + 18 + 6} = \frac{2326}{2743} = 0.8479$$

Confusion Matrix of the **Chosen Model**

Y.hat	0 Detected as Non-CVD	1 Detected as CVD
0 Don't have CVD	2308 True Positive	411 False Positive (Type 1 error)
1 Suffer from CVD	6 False Negative (Type 2 error)	18 True Negative

Conclusion

This model can classify and predict the Cardiovascular Disease with 84.79 % Accuracy

Classification
Statistics

Precision

Precision

Precision denotes the accuracy level of a model's positive predictions. It quantifies the proportion of true positive predictions relative to the overall number of positive predictions generated by the model.

$$Precision = \frac{TP}{TP + FP}$$

For this chosen logistic regression model ,

$$Precision = \frac{2308}{2308 + 411} = \frac{2308}{2719} = 0.8488$$

Confusion Matrix of the **Chosen Model**

Y.hat	0 Detected as Non-CVD	1 Detected as CVD
0 Don't have CVD	2308 True Positive	411 False Positive (Type 1 error)
1 Suffer from CVD	6 False Negative (Type 2 error)	18 True Negative

Conclusion

Approximately 89.35% of the positive predictions made by the model are correct. In other words, when the model predicts a positive outcome, it is accurate about 84.88% of the time.

Classification
Statistics

Specificity

Specificity

Specificity, also referred to as the True Negative Rate, gauges the model's capacity to accurately identify negative instances. It is determined by dividing the number of true negative predictions by the total count of actual negative instances.

$$Specificity = \frac{TN}{FP + TN}$$

For this chosen logistic regression model ,

$$Specificity = \frac{18}{411 + 18} = \frac{18}{429} = 0.04195$$

Confusion Matrix of the **Chosen Model**

Y.hat	0 Detected as Non-CVD	1 Detected as CVD
0 Don't have CVD	2308 True Positive	411 False Positive (Type 1 error)
1 Suffer from CVD	6 False Negative (Type 2 error)	18 True Negative

Conclusion

it means that approximately 4.195% of the true negative cases are correctly identified by the model.

Classification **Statistics**

Sensitivity

Similarly termed as the True Positive Rate or Recall, sensitivity assesses the model's ability to correctly identify positive instances. It is computed by dividing the number of true positive predictions by the total count of actual positive instances

$$Sensitivity = \frac{TP}{TP + FN}$$

For this chosen logistic regression model ,

$$Sensitivity = \frac{2308}{2308 + 6} = \frac{2308}{2314} = 0.9974$$

Confusion Matrix of the **Chosen Model**

Y.hat	0 Detected as Non-CVD	1 Detected as CVD
0 Don't have CVD	2308 True Positive	411 False Positive (Type 1 error)
1 Suffer from CVD	6 False Negative (Type 2 error)	18 True Negative

Conclusion

It means that approximately 99.74% of the true positive cases are correctly identified by the model.

Classification **Statistics**

Misclassification Rate

Misclassification Rate

The Misclassification Rate in Logistic Regression is the proportion of wrongly classified observations, indicating how accurately the model predicts outcomes.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\Rightarrow Misc\ Rate = 1 - Accuracy$$

$$\Rightarrow Misc.Rate = \frac{FP + FN}{TP + FP + TN + FN}$$

For this chosen logistic regression model ,

$$Misc\ rate = 1 - Accuracy = 1 - 0.8479 = 0.1521$$

Confusion Matrix of the **Chosen Model**

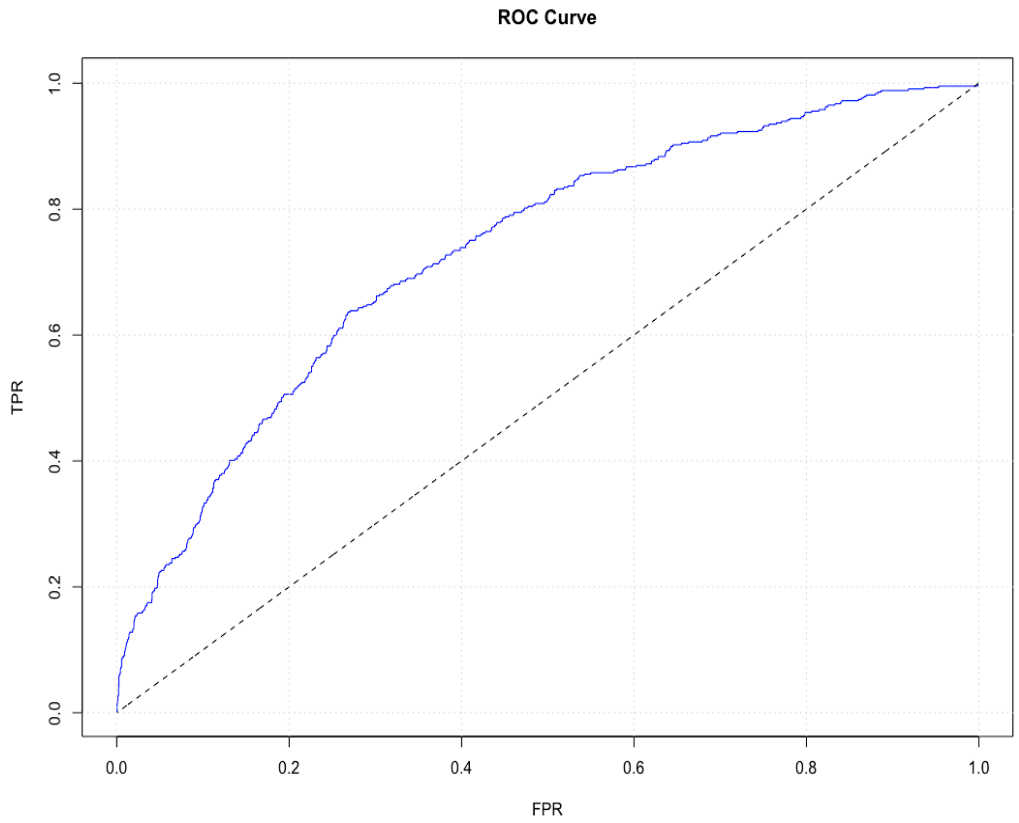
Y.hat	0 Detected as Non-CVD	1 Detected as CVD
0 Don't have CVD	705 True Positive	84 False Positive (Type 1 error)
1 Suffer from CVD	80 False Negative (Type 2 error)	44 True Negative

Conclusion

It means that approximately 15.21% of the observations in the dataset are misclassified by the model.

Receiver Operator **Characteristic Curve**

The ROC curve serves as a vital tool in assessing the predictive capability of a model by visualizing the trade-off between its True Positive Rate (TPR) and False Positive Rate (FPR). A higher area under the ROC curve indicates superior discriminatory power of the model, with values closer to 1 suggesting better performance. In this case, with an area under the ROC curve of 0.7335 for the logistic regression model, it suggests that the model exhibits reasonably good discriminatory ability in distinguishing between positive and negative cases. This value signifies that the model performs significantly better than random guessing and can effectively classify instances into their respective classes, albeit with room for potential improvement.



Testing the Model
In Test Data

```
fitted_prob_test = predict(Selected_Model, newdata =
testing_data, type = "response")

Y.hat_test = ifelse(fitted_prob_test >
optimal_thresold_for_test_data, 1, 0)
confusion_matrix_test_data = table(Y.hat_test,
testing_data$TenYearCHD)
confusion_matrix_test_data
accuracy_test = sum(diag(confusion_matrix_test_data)) /
sum(confusion_matrix_test_data)
round(accuracy_test, 4)
roc_curve = roc(testing_data$TenYearCHD, fitted_prob_test)
auc(roc_curve)
```

Test data ; at the beginning I've split the dataset into 8:2 parts so that I can use the the training data for training purpose and test data for testing the model.

Accuracy : 82.03%

Area Under the Curve : 0.7465

Testing the Model
In Test Data

```
fitted_prob_test = predict(Selected_Model, newdata =
testing_data, type = "response")

Y.hat_test = ifelse(fitted_prob_test >
optimal_thresold_for_test_data, 1, 0)
confusion_matrix_test_data = table(Y.hat_test,
testing_data$TenYearCHD)
confusion_matrix_test_data
accuracy_test = sum(diag(confusion_matrix_test_data)) /
sum(confusion_matrix_test_data)
round(accuracy_test, 4)
roc_curve = roc(testing_data$TenYearCHD, fitted_prob_test)
auc(roc_curve)
```

Test data ; at the beginning I've split the dataset into 8:2 parts so that I can use the the training data for training purpose and test data for testing the model.

Accuracy : 82.03%

Area Under the Curve : 0.7465

Final **Conclusion**

In conclusion, our analysis reveals several significant predictors of heart disease diagnosis.

Gender:
Males exhibit a 65.34% higher likelihood of heart disease diagnosis compared to females, suggesting a notable gender-based disparity.

Age:
Each one-year increase in age corresponds to a 6.5% rise in the odds of heart disease diagnosis, indicating age as a significant risk factor.

Cigarette Consumption:
The odds of heart disease diagnosis increase by approximately 1.8% with each additional cigarette smoked per day, emphasizing the detrimental effects of smoking.

Prevalent Hypertension:
Individuals with prevalent hypertension have a 30.81% higher odds of heart disease diagnosis, highlighting the importance of managing blood pressure.

Total Cholesterol Level:
A one-unit increase in total cholesterol level corresponds to a 0.32% rise in the odds of heart disease diagnosis, underscoring the significance of lipid management.

Systolic Blood Pressure:
Each unit increase in systolic blood pressure is associated with a 1.41% increase in the odds of heart disease diagnosis, emphasizing the importance of blood pressure control.

Glucose Level:
With each unit increase in glucose level, the odds of heart disease diagnosis increase by approximately 0.73%, indicating the relevance of glycemic control.

These findings collectively highlight the multifactorial nature of heart disease risk and underscore the importance of addressing various modifiable risk factors in preventive strategies and clinical management.

Know **How**

I’m very much thankful to you for giving your valuable knowledge and knowhow to fulfil this dissertation topic.

St.Xavier's College(Autonomous), Kolkata

Department of Statistics

Name - **Suchibrata Patra**

Roll No - **454**

Dissertation Supervisor - **Prof. Debjit Sengupta**

Regn. No - **A01 - 1112 - 0867 - 21**

Session - **2021-24**

Semester - **6**

Predictive Modelling for Coronary Heart Disease Risk Assessment: Empowering Healthcare Strategies

Declaration :

"I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials."

Suchibrata Patra

Signature

Backup
Slides

Scope and Objective

A

Investigate how lifestyle choices impact heart health and explore ways to reduce the risk of cardiovascular disease (CVD).

B

Evaluate the imp. of early detection in preventing severe outcomes related to CVD.

C

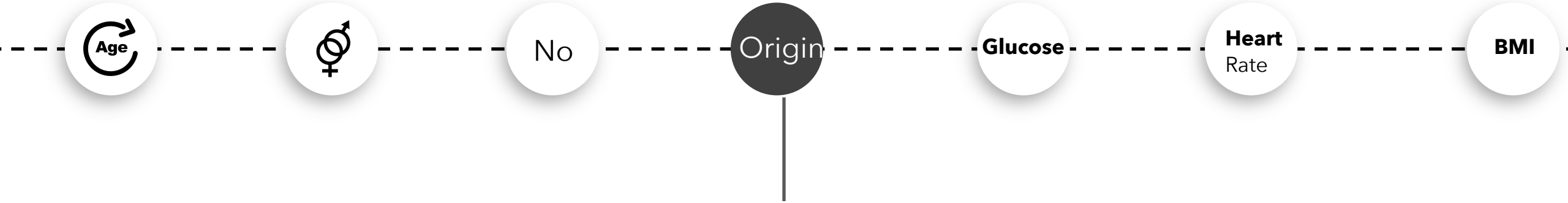
Develop and test statistical models to identify individuals at high risk of developing CVD, aiding in timely intervention and prevention.

D

Provide actionable insights for healthcare prof. and policymakers to make strategies for reducing the global burden of CVD.

Main Slide

GETTING TO KNOW THE **DATASET**



No of Datapoints

In the Dataset, there are in total 3,657 data points available encompassing various demographic, behavioural, and medical risk factors.

Origin Of the Dataset

This Cardiovascular Disease Dataset Dataset Was collected from the UCI Machine Learning Library

No of Datapoin

In the Dataset, there total 3,657 data po available encompas various demograp behavioural, and med factors.

Briefly the Complete **DATASET**

Denoted	Covariate	Nature
x_1	Sex	Categorical
x_2	Age	Continuous
x_3	Education	Categorical
x_4	Smoker	Binomial
x_5	Cigs/Day	Discrete

Denoted	Covariate	Nature
x_6	**Bp** Meds	Categorical
x_7	**Prev.** Stroke	Categorical
x_8	**Prev.** Hyp	Categorical
x_9	Diabetes	Categorical
x_{10}	**Total** Cholesterol	Continuous

Denoted	Covariate	Nature
x_{11}	**Sys** BP	Continuous
x_{12}	**Dias** BP	Continuous
x_{13}	BMI	Continuous
x_{14}	**Heart** Rate	Continuous
x_{15}	Glucose	Continuous