

---

# PM PROJECT REVIEW I

GI4: HEART FAILURE PREDICTION

19162121038 Dishwa Shah

19162121044 Tanishk Sethiya

19162121047 Yash Talati

20162122006 Jeel Patel



# AIM

CREATE A MODEL FOR PREDICTING MORTALITY CAUSED BY HEART FAILURE.



# ABOUT DATASET

- Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide.  
Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.
- Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.
- People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

# OBJECTIVE

- Clean the dataset to remove any existing outliers and extremes.
- Train the model to predict whether the patient is likely to face mortality due to heart failure considering his age, health and other lifestyle choices.
- Train the model to determine patterns amongst the various lifestyle and health aspects that are more likely to cause mortality due to heart failure.
- Perform exploratory analysis to find patterns and insights.



# DATA DESCRIPTION REPORT

KNOW YOUR DATA



# DATA DESCRIPTION ANALYSIS

- Age: Age of patients (years)
- Anaemia: Condition where red blood cells are decreased (Boolean: 0 for false, 1 for true)
- Creatinine\_phosphokinase: Level of CPK enzyme in blood (mcg/L)
- Diabetes: Whether the patient has diabetes (Boolean: 0 for false, 1 for true)
- Ejection\_fraction: Percentage of blood leaving from heart at each contraction (percentage)
- High\_blood\_pressure: Whether the patient has high blood pressure (Boolean: 0 for false, 1 for true)
- Platelets: Platelets in blood (kiloplatelets/mL)
- Serum\_creatinine: Level of serum creatinine in blood (mg/dL)
- Serum\_sodium: Level of serum sodium in blood (mEq/L)
- Sex: Man or Woman (Boolean: 0 for man, 1 for woman)
- Smoking: Whether the patient smokes or not (Boolean: 0 for false, 1 for true)
- Time: Follow-up period (days)
- Death\_event: If the patient dies during follow-up period (Boolean: 0 for false, 1 for true)

# DATA DESCRIPTION ANALYSIS

- **What is the format of the data?**

The data is of .csv format.

- **Which method is used to capture the data?**

Here, the method to capture the data is unknown since we are not a part of end-to-end process.

- **How large is the database?**

The database has 299 rows and 13 columns.

- **Does the data include characteristics relevant to the business perspective?**

Yes, the data does include characteristics relevant to the business perspective which can help us achieve our aim.

- **What data types are present?**

We have integer and real data types and Flag and Continuous measurement levels.

# DATA DESCRIPTION ANALYSIS

- **Did you compute basic statistics for the key attributes? What insight did this provide into the business question?**

Yes, I did compute basic statistics for the key attributes. We can see that some of the data is skewed. We will see what this provides to our business questions further.

Field	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
age	Continuous	40	95	60.829	11.895	0.424	--	299
anaemia	Flag	0	1	--	--	--	2	299
cpk	Continuous	23	7861	581.839	970.288	4.463	--	299
diabetes	Flag	0	1	--	--	--	2	299
ejection	Continuous	14	80	38.084	11.835	0.555	--	299
bp	Continuous	0	1	0.351	0.478	0.627	--	299
platelets	Continuous	25100.000	850000.000	263358.029	97804.237	1.462	--	299
serum creatinine	Continuous	0.500	9.400	1.394	1.035	4.456	--	299
serum sodium	Continuous	113	148	136.625	4.412	-1.048	--	299
sex	Flag	0	1	--	--	--	2	299
smoking	Flag	0	1	--	--	--	2	299
time	Continuous	4	285	130.261	77.614	0.128	--	299
death	Flag	0	1	--	--	--	2	299



# DATA DESCRIPTION ANALYSIS

- **Are you able to prioritize relevant attributes? If not, are business analysts available to provide further insights?**

Yes, I did prioritize relevant attributes. All the fields can somehow be the cause of death. Hence, we are assigning the role of Input to all the attributes.

Field	Measurement	Values	Missing	Check	Role
age	Continuous	[40,95]		None	Input
anaemia	Flag	1/0		None	Input
cpk	Continuous	[23,7861]		None	Input
diabetes	Flag	1/0		None	Input
ejection	Continuous	[14,80]		None	Input
bp	Continuous	[0,1]		None	Input
platelets	Continuous	[25100.0,850000.0]		None	Input
serum creatinine	Continuous	[0.5,9.4]		None	Input
serum sodium	Continuous	[113,148]		None	Input
sex	Flag	1/0		None	Input
smoking	Flag	1/0		None	Input
time	Continuous	[4,285]		None	Input
death	Flag	1/0		None	Input



# DATA EXPLORATION REPORT

EXPLORE YOUR DATA



# DATA EXPLORATION

## Measurement levels

- Age, cpk, ejection level, blood pressure, platelets, serum creatinine, serum sodium, time and death are assigned Continuous measurement level since they all have values from a continuous range.
- Anaemia, diabetes, sex, smoking and death have been assigned Flag measurement level since they have 0/1 as their values.

# DATA EXPLORATION

## Roles

- Currently, all fields have been assigned Input as their roles since we want to derive all the patterns among the attributes that will help build a connection between lifestyle and heart failure in order to prevent it rather than just predicting if the patient is likely to die or not.
- However, we will set death as a target in another Type node later to merely predict whether the patient is likely to die of heart failure or not.

Field	Measurement	Values	Missing	Check	Role
age_transfor...	Continuous	[-1.760536...		None	Input
creatinine_ph...	Continuous	[-0.757990...		None	Input
ejection_fracti...	Continuous	[-2.054378...		None	Input
platelets_tran...	Continuous	[-2.653651...		None	Input
serum_creati...	Continuous	[-1.171258...		None	Input
serum_sodiu...	Continuous	[-2.999999...		None	Input
time_transfor...	Continuous	[-1.647126...		None	Input
anaemia_tra...	Flag	1/0		None	Input
diabetes_tran...	Flag	1/0		None	Input
high_blood_p...	Flag	1/0		None	Input
sex_transfor...	Flag	1/0		None	Input
smoking_tran...	Flag	1/0		None	Input
DEATH_EVE...	Flag	1/0		None	Target

# DATA EXPLORATION

## Format

- ##### indicates that there is maximum 4 digits in that field.

Field	Format	Justify	Column Width
age_transformed	#####	Auto	Auto
creatinine_phosphokinase_transformed	#####	Auto	Auto
ejection_fraction_transformed	#####	Auto	Auto
platelets_transformed	#####	Auto	Auto
serum_creatinine_transformed	#####	Auto	Auto
serum_sodium_transformed	#####	Auto	Auto
time_transformed	#####	Auto	Auto
anaemia_transformed	####	Auto	Auto
diabetes_transformed	####	Auto	Auto
high_blood_pressure_transformed	####	Auto	Auto
sex_transformed	####	Auto	Auto
smoking_transformed	####	Auto	Auto
DEATH_EVENT_transformed	####	Auto	Auto

# DATA EXPLORATION

## Statistics

Data Audit of [13 fields] #76



**Audit** Quality Annotations

Field	Graph	Measurement	Min	Max	Sum	Range	Mean	Mean Std. Err.	Std. Dev	Variance	Skewness	Skewness Std. Err.	†
age		Continuous	40	95	18188	55	60.829	0.688	11.895	141.491	0.424	0.141	
anaemia		Flag	0	1	--	--	--	--	--	--	--	--	
creatinine_phosphokinase		Continuous	23	7861	173970	7838	581.839	56.113	970.288	941458.571	4.463	0.141	
diabetes		Flag	0	1	--	--	--	--	--	--	--	--	
ejection_fraction		Continuous	14	80	11387	66	38.084	0.684	11.835	140.063	0.555	0.141	
high_blood_pressure		Flag	0	1	--	--	--	--	--	--	--	--	
platelets		Continuous	25100.000	850000.000	78744050.750	824900.000	263358.029	5656.165	97804.237	9565668749.449	1.462	0.141	
serum_creatinine		Continuous	0.500	9.400	416.770	8.900	1.394	0.060	1.035	1.070	4.456	0.141	
serum_sodium		Continuous	113	148	40851	35	136.625	0.255	4.412	19.470	-1.048	0.141	
~		--	-	-	-	-	-	-	-	-	-	-	

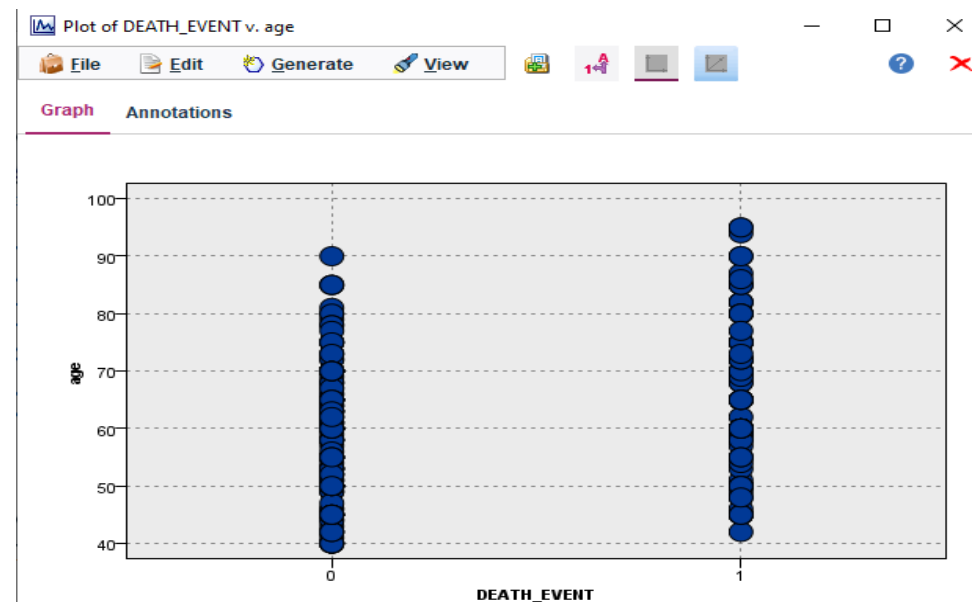
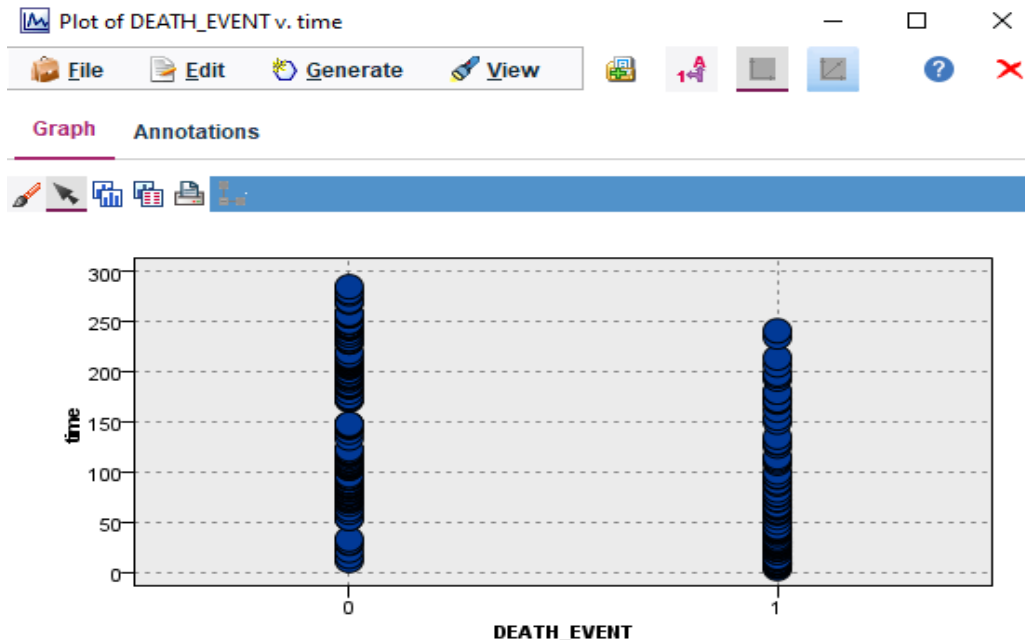
† Indicates a multimode result ‡ Indicates a sampled result

Activate Windows  
Go to Settings to activate Windows.

# DATA EXPLORATION

## What sort of hypothesis have you formed about the data?

- Death rate caused by heart failure might decrease with the increase number of days the patient will be under observation since he would have discharged only after showing improvements.
- Deaths due to heart failure might be high in people aged 40-60 due to their unhealthy lifestyle.
- Deaths due to heart failure are related to all the health conditions in the database.
- Deaths due to heart failure might not be related to sex of the person.



# DATA EXPLORATION

## **Which attributes seem promising for further analysis?**

Age, anaemia, creatinine\_phosphokinase, diabetes, ejection\_fraction, high\_blood\_pressure, serum\_creatinine, serum\_sodium and smoking are seeming to be the most promising for further analysis.

## **Have your explorations revealed new characteristics about your data?**

Explorations helped us understand the data better but haven't revealed any new characteristics yet. It will probably happen during modeling.

## **How have your explorations changed your initial hypothesis?**

Explorations have definitely helped us understand the data better and gave us the clarity of attributes which will be helpful during modeling. However, the initial hypothesis still stands the same.

## **Can you identify particular subsets of data for later use?**

I am planning to segment each of my attributes with death to understand the individual patterns properly and to not miss out on anything.

## **Take another look at your data mining goals. Has this exploration altered the goals?**

While the goals remain the same, I definitely believe that additional data might really help to predict deaths caused by heart failures.





# DATA QUALITY REPORT

EVALUATE THE QUALITY OF YOUR DATA



# DATA QUALITY

Using data audit node, we can see there are some outliers and extremes that are skewing our data. Let us handle them. Here, I have coerced the outliers and discarded the extremes and then connected the super node (newly generated) to auto data prep node.

Complete fields (%):

Complete records (%):

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space
age	Continuous	0	0	None	Never	Fixed	100	299	0	0	
anaemia	Flag	--	---		Never	Fixed	100	299	0	0	
creatinine_p...	Continuous	4	3	Coerce outliers / discard extremes	Never	Fixed	100	299	0	0	
diabetes	Flag	--	---		Never	Fixed	100	299	0	0	
ejection_fract...	Continuous	1	0	Coerce	Never	Fixed	100	299	0	0	
high_blood_...	Flag	--	---		Never	Fixed	100	299	0	0	
platelets	Continuous	2	1	Coerce outliers / discard extremes	Never	Fixed	100	299	0	0	
serum_creati...	Continuous	3	3	Coerce outliers / discard extremes	Never	Fixed	100	299	0	0	
serum_sodi...	Continuous	2	1	Coerce outliers / discard extremes	Never	Fixed	100	299	0	0	
sex	Flag	--	---		Never	Fixed	100	299	0	0	
smoking	Flag	--	---		Never	Fixed	100	299	0	0	
time	Continuous	0	0	None	Never	Fixed	100	299	0	0	
DEATH_EVE...	Flag	--	---		Never	Fixed	100	299	0	0	

# DATA QUALITY

- Let us optimize for accuracy rather than speed.

Objectives

Fields

Settings

Analysis

Annotations

Automated Data Preparation can recommend data preparation steps that will speed up model building and transformed.

What is your objective?

☐ Balance speed and accuracy

Transform the data with an emphasis on building models with a balance of speed and accuracy.

☐ Optimize for speed

Transform the data with an emphasis on building models as quickly as possible.

☒ Optimize for accuracy

Transform the data with an emphasis on building models with the greatest predictive power.

☐ Custom analysis

Choose this option to fine tune the algorithm on the Settings tab.

# DATA QUALITY


- We have taken all fields as input and none as target.


**Objectives** **Fields** **Settings** **Analysis** **Annotations**


☐ Use predefined roles ☒ Use custom roles


Target (optional):


Inputs:


 age


 anaemia


 creatinine\_phosphokinase


 diabetes


 ejection\_fraction


 high\_blood\_pressure


 platelets


 serum\_creatinine

 serum\_sodium

 sex

 smoking

 time

 DEATH\_EVENT

# DATA QUALITY

- I have made a few changes in settings according to my dataset and requirements as shown.

**Objectives** **Fields** **Settings** **Analysis** **Annotations**

---

Settings

**Field Settings**  
Prepare Dates & Times  
Exclude Input Fields  
Prepare Inputs & Target  
Construct & Select Features  
Field Names

Field settings are not affected if you change your objective.

☐ Use frequency field

☐ Use weight field

How to handle fields that are excluded from modeling:

☐ Filter out unused fields

☒ Set the direction of unused fields to "None"

If the incoming fields do not match the existing analysis:

☒ Stop execution and keep the existing analysis

☐ Clear the existing analysis and analyze the new data

- I do not need this because I have no dates and times in my data.

Objectives   Fields   **Settings**   Analysis   Annotations

# DATA QUALITY

- I unselected nominal and ordinal settings because I do not have data with those measurement levels.

Objectives Fields **Settings** Analysis Annotations

## Settings

Field Settings

Prepare Dates & Times

Exclude Input Fields

Prepare Inputs & Target

Construct & Select Features

Field Names

Constant fields will always be excluded.

☒ Exclude low quality input fields

### Exclude Input Fields

☒ Exclude fields with too many missing values

Maximum percentage of missing values:  %

☐ Exclude nominal fields with too many unique categories

Maximum number of categories:

☐ Exclude categorical fields with too many values in a single category

Maximum percentage in a single category:  %

# DATA QUALITY

- Again, I only kept the ones having my measurement levels in it. I have kept the remaining settings unchanged since they did not require to be changed.

Objectives Fields **Settings** Analysis Annotations

Settings

- Field Settings
- Prepare Dates & Times
- Exclude Input Fields
- Prepare Inputs & Target**
- Construct & Select Features
- Field Names

☒ Prepare the input and target fields for modeling

Adjust Type and Improve Data Quality

Inputs	Target
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Adjust the type of numeric fields (ordinal and continuous)
<input type="checkbox"/>	<input type="checkbox"/> Reorder nominal fields to have smallest category first, largest last
<input checked="" type="checkbox"/>	<input type="checkbox"/> Replace outlier values in continuous fields (recommended for input fields if they will be put on a common scale)
<input checked="" type="checkbox"/>	<input type="checkbox"/> Continuous fields: replace missing values with mean
<input type="checkbox"/>	<input type="checkbox"/> Nominal fields: replace missing values with mode
<input type="checkbox"/>	<input type="checkbox"/> Ordinal fields: replace missing values with median

Maximum number of values for ordinal fields:

Minimum number of values for continuous fields:

Outlier cutoff value:  (standard deviations)

Method for replacing outliers: ☒ Replace with cutoff value ☐ Delete value

Transform Continuous Field

☒ Put all continuous input fields on a common scale (highly recommended if feature construction will be performed)

Rescaling method:  Final mean:  Final standard deviation:

☐ Rescale a continuous target with a Box-Cox transformation to reduce skew

Final mean:  Final standard deviation:

Activate Windows  
Go to Settings to activate Windows.



# DATA QUALITY

- Now connect a data audit node to this auto data prep node.
- We can see how the records having extremes have been removed from the data and how the outliers have been coerced.
- This shows that the data has been cleaned according to the requirements.

Complete fields (%):  Complete records (%):

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space
age_transfor...	Continuous	0	0	None	Never	Fixed	100	291	0	0	
creatinine_p...	Continuous	0	0	None	Never	Fixed	100	291	0	0	
ejection_fract...	Continuous	0	0	None	Never	Fixed	100	291	0	0	
platelets_tra...	Continuous	0	0	None	Never	Fixed	100	291	0	0	
serum_creati...	Continuous	0	0	None	Never	Fixed	100	291	0	0	
serum_sodi...	Continuous	0	0	None	Never	Fixed	100	291	0	0	
time_transfor...	Continuous	0	0	None	Never	Fixed	100	291	0	0	
anaemia_tra...	Flag	--	--		Never	Fixed	100	291	0	0	
diabetes_tra...	Flag	--	--		Never	Fixed	100	291	0	0	
high_blood_...	Flag	--	--		Never	Fixed	100	291	0	0	
sex_transfor...	Flag	--	--		Never	Fixed	100	291	0	0	
smoking_tra...	Flag	--	--		Never	Fixed	100	291	0	0	
DEATH_EVE...	Flag	--	--		Never	Fixed	100	291	0	0	

# DATA QUALITY

- **Have you identified missing attributes and blank fields? If so, is there meaning behind such missing values?**

No, there are no missing attributes and blank fields. However, the labels in sex attribute were not mentioned but I did my research and added them.

- **Are there spelling inconsistencies that may cause problems in later merges or transformations?**

No, there are no spelling inconsistencies that may cause problems in later merges or transformations since the data has no alphabets or characters.

- **Have you explored deviations to determine whether they are "noise" or phenomena worth analysing further?**

There is some skewness in a few attributes which can later be reduced to clean the noise.

# DATA QUALITY

- **Have you conducted a plausibility check for values? Take notes on any apparent conflicts (such as teenagers with high income levels).**

Yes, I did conduct a plausibility check for values. There is no apparent conflict as such.

- **Have you considered excluding data that has no impact on your hypothesis?**


Yes, I have. It concluded with the fact that all data is important for the analysis since each and every attribute might add up to the death caused by heart failure. So, I will not exclude any data.


- **Are the data stored in flat files? If so, are the delimiters consistent among files? Does each record contain the same number of fields?**

Yes. The delimiters are consistent among files. Each record contains the same number of fields.

# SEGMENTATION MODELING

- We will use K-Means model to create 5 clusters of our data.

 K-Means ×

 ? □ □

Fields

**Model**

Expert

Annotations

Model name:

☒ Auto ☐ Custom

☒ Use partitioned data

Number of clusters:

☐ Generate distance field

Cluster label:

☒ String ☐ Number

Label prefix:

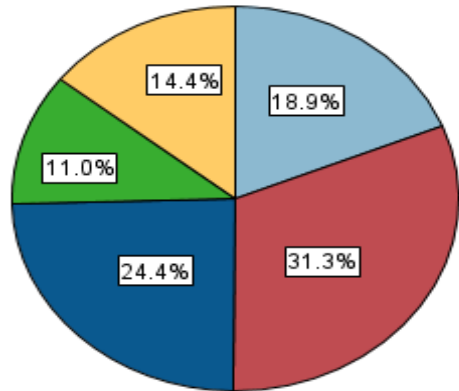
Optimize:

☐ Speed ☒ Memory

# SEGMENTATION MODELING

- Clusters have been created. These clusters will group the data for us, as to which group of lifestyle patterns is more likely to experience a heart failure.

Cluster Sizes



Cluster

- cluster-1
- cluster-2
- cluster-3
- cluster-4
- cluster-5

Size of Smallest Cluster	32 (11%)
Size of Largest Cluster	91 (31.3%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	2.84

Table (27 fields, 291 records) #3

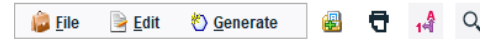









Table Annotations

	rmed	serum_sodium_transformed	time_transformed	anaemia_transformed	diabetes_transformed	high_blood_pressure_transformed	sex_transformed	smoking_transformed	DEATH_EVENT_transformed	\$KM-K-Means
1	0.847	-1.655	-1.647	1	1	0	1	1	1	0 cluster-1
2	0.018	-1.902	-1.608	1	1	1	1	0	0	0 cluster-4
3	0.847	0.071	-1.608	0	1	1	1	1	1	0 cluster-1
4	2.000	-3.287	-1.595	0	0	1	0	1	1	0 cluster-2
5	1.135	-1.162	-1.595	0	1	0	1	0	0	0 cluster-5
6	0.162	0.071	-1.570	0	1	1	1	1	1	0 cluster-1
7	0.306	-1.409	-1.570	0	0	1	1	0	0	0 cluster-4
8	0.270	0.318	-1.570	1	1	1	0	1	1	0 cluster-1
9	3.874	-1.409	-1.570	0	1	0	1	0	0	0 cluster-5
10	0.595	0.811	-1.570	1	1	0	1	0	0	0 cluster-1
11	0.306	0.071	-1.557	0	1	1	1	1	1	0 cluster-1
12	0.306	0.071	-1.557	0	1	0	1	1	1	0 cluster-5
13	0.451	0.318	-1.544	0	1	0	0	0	1	1 cluster-5
14	0.018	-0.175	-1.531	0	1	1	1	1	1	0 cluster-1
15	0.595	0.811	-1.518	0	1	1	1	1	1	0 cluster-1
16	0.739	-2.396	-1.518	1	1	1	1	1	1	0 cluster-1
17	0.451	0.811	-1.505	0	1	0	0	0	1	0 cluster-5
18	0.847	-3.287	-1.505	0	0	1	0	0	1	0 cluster-2
19	0.018	0.071	-1.492	0	1	0	0	0	1	1 cluster-5
20	0.414	-0.175	-1.441	0	0	0	0	0	1	0 cluster-5
21	0.595	0.811	-1.441	0	1	0	0	1	0	0 cluster-5
22	0.739	-0.422	-1.415	1	0	1	1	1	1	1 cluster-2
23	0.746	-0.669	-1.402	1	0	0	0	0	1	0 cluster-1
24	0.847	1.798	-1.402	1	0	1	1	0	0	0 cluster-4
25	0.451	0.318	-1.389	0	1	0	0	0	1	0 cluster-5
26	0.018	-0.175	-1.363	1	0	0	1	0	0	0 cluster-4
27	4.591	-0.669	-1.363	0	1	1	1	1	1	0 cluster-1
28	0.162	-1.162	-1.363	1	0	1	1	0	0	0 cluster-4
29	0.746	-0.669	-1.350	1	0	0	1	1	1	0 cluster-1
30	2.432	-1.162	-1.337	1	1	1	1	1	1	0 cluster-1
31	0.451	-2.149	-1.337	0	0	0	0	0	1	0 cluster-5
32	0.162	0.318	-1.324	0	0	1	0	0	1	1 cluster-2
33	0.451	0.811	-1.324	1	0	0	1	1	1	0 cluster-1
34	3.153	-0.669	-1.311	1	0	1	1	1	1	0 cluster-4

# SUPERVISED MODELING

- We will partition the data into 80% training and 20% testing.

 Partition >

  Generate  Preview   

Settings Annotations

Partition field:

Partition

Partitions:

☒ Train and test ☐ Train, test and validation

Training partition size:

80

Label: Training

Value = "1\_Training"

Testing partition size:

20

Label: Testing

Value = "2\_Testing"

Validation partition size:

0

Label: Validation

Value = "3\_Validation"

Total size:

100%

Values:

☐ Use system-defined values ("1", "2" and "3")

☒ Append labels to system-defined values


☐ Use labels as values

☒ Repeatable partition assignment

Seed: 1234567

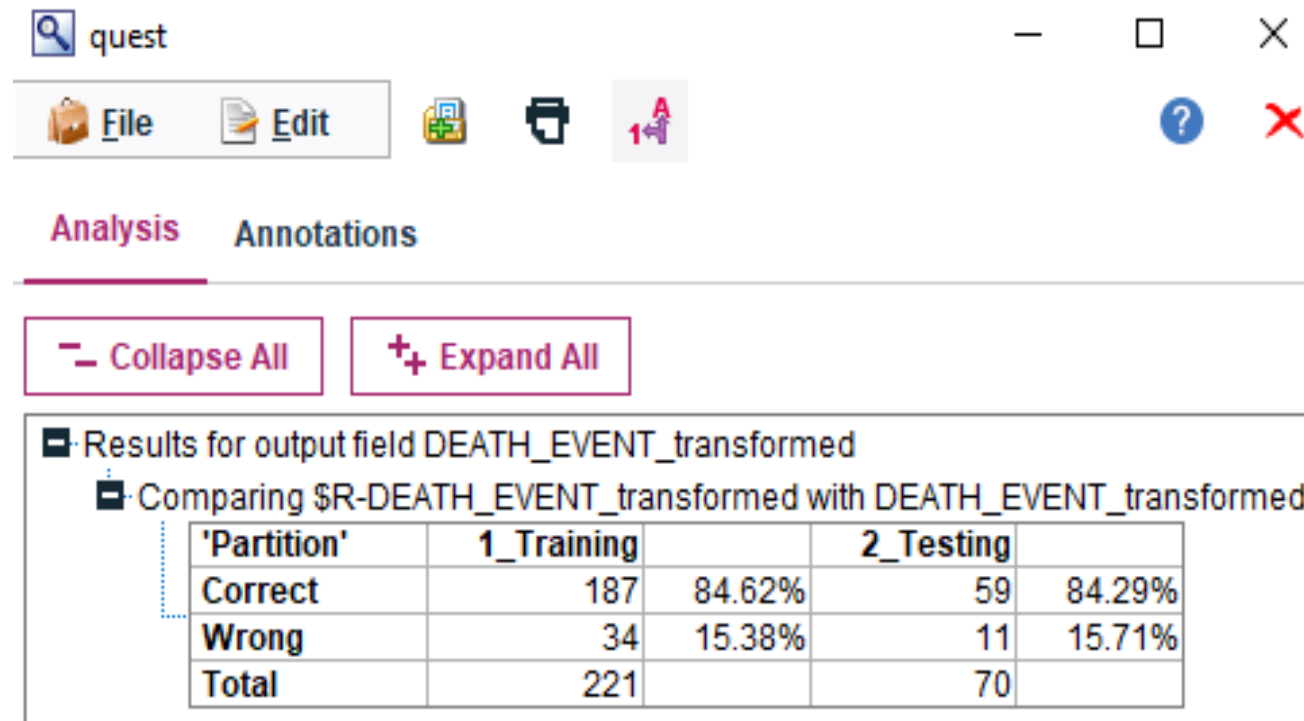
Generate

☐ Use unique field to assign partitions:



# SUPERVISED MODELING

- Now, after checking accuracies of all the models, C5.0 and Quest give us the best accuracy of 84.29%
- We can use this model to predict whether the person is likely to face mortality due to heart failure or not as per his lifestyle data/records.



The screenshot displays the Quest model evaluation interface. At the top, there is a search bar with the text 'quest' and a magnifying glass icon. Below this is a toolbar with icons for File, Edit, and other functions. The main content area is divided into two tabs: 'Analysis' (selected) and 'Annotations'. Under the 'Analysis' tab, there are two buttons: 'Collapse All' and 'Expand All'. The results section shows a tree view with the following structure:

- [-] Results for output field DEATH\_EVENT\_transformed
  - [-] Comparing \$R-DEATH\_EVENT\_transformed with DEATH\_EVENT\_transformed

The comparison results are displayed in a table:

'Partition'	1_Training		2_Testing	
Correct	187	84.62%	59	84.29%
Wrong	34	15.38%	11	15.71%
Total	221		70	

# CONCLUSION

- We have managed to successfully clean the data.
- We segmented our data using K-Means model to divide it in 5 different clusters to predict and study the patterns of lifestyle that contributes to the heart failure.
- We compared all the supervised modelling nodes where C5.0 and Quest models give the best accuracy of 84.29%.
- We can further use these models for deployment and to predict the mortality caused by heart failure.