Lessons learned and opportunities with Predictive Analytics at Takeda Pharmaceuticals

Discover breakthroughs industry leaders are making with analytics

20 September 2022 | Munich

Better Health, Brighter Future

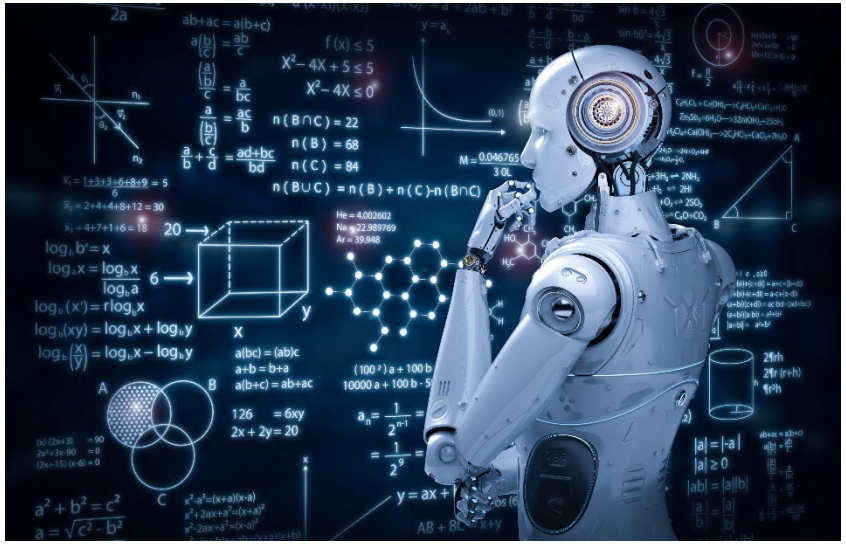
Agenda

- Predictive Analytical Module principle
- Advantages and Benefits coming from Predictive Analytical Module (PAM)
- Fields to be used and case study (Advate)
- PAM Strategy deployment

Predictive Analytical Module principle

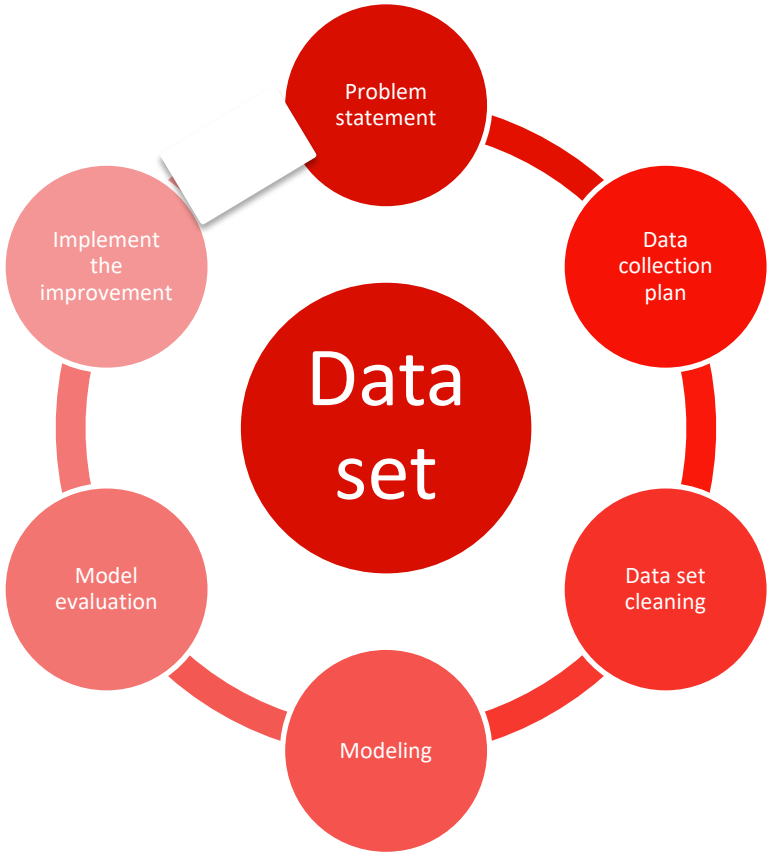
What is machine learning ?

- Technology used to implement machine learning as algorithms. An algorithm is a series of step-by-step operations, usually computations, that can solve a defined problem in a finite number of steps. In machine learning, the algorithms use a series of finite steps to solve the problem by learning from data.



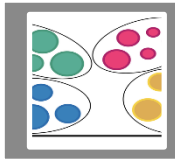
DATA mining process

- Problem statement : define the business problem to be solved
- Data collection plan : determine what data we need
- Data set cleaning : organize the data set to be used in Minitab
  - Missing data
  - Noisy data
  - Inconsistent data
- Modeling : by using CART construct models
- Model evaluation : evaluate the robustness of the model
- Implement the improvement : define a strategy to deploy the solution to fix the issue



How to choose the right methodology

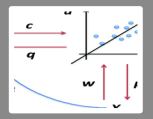
Machine learning



Unsupervised learning

- No response
- Want to find grouping in our Data only (Xs)

- CLUSTERING
- PCA*



Supervised learning

- Response / Target / variable (Y)
- Multiple predictors Xs

Categorical response

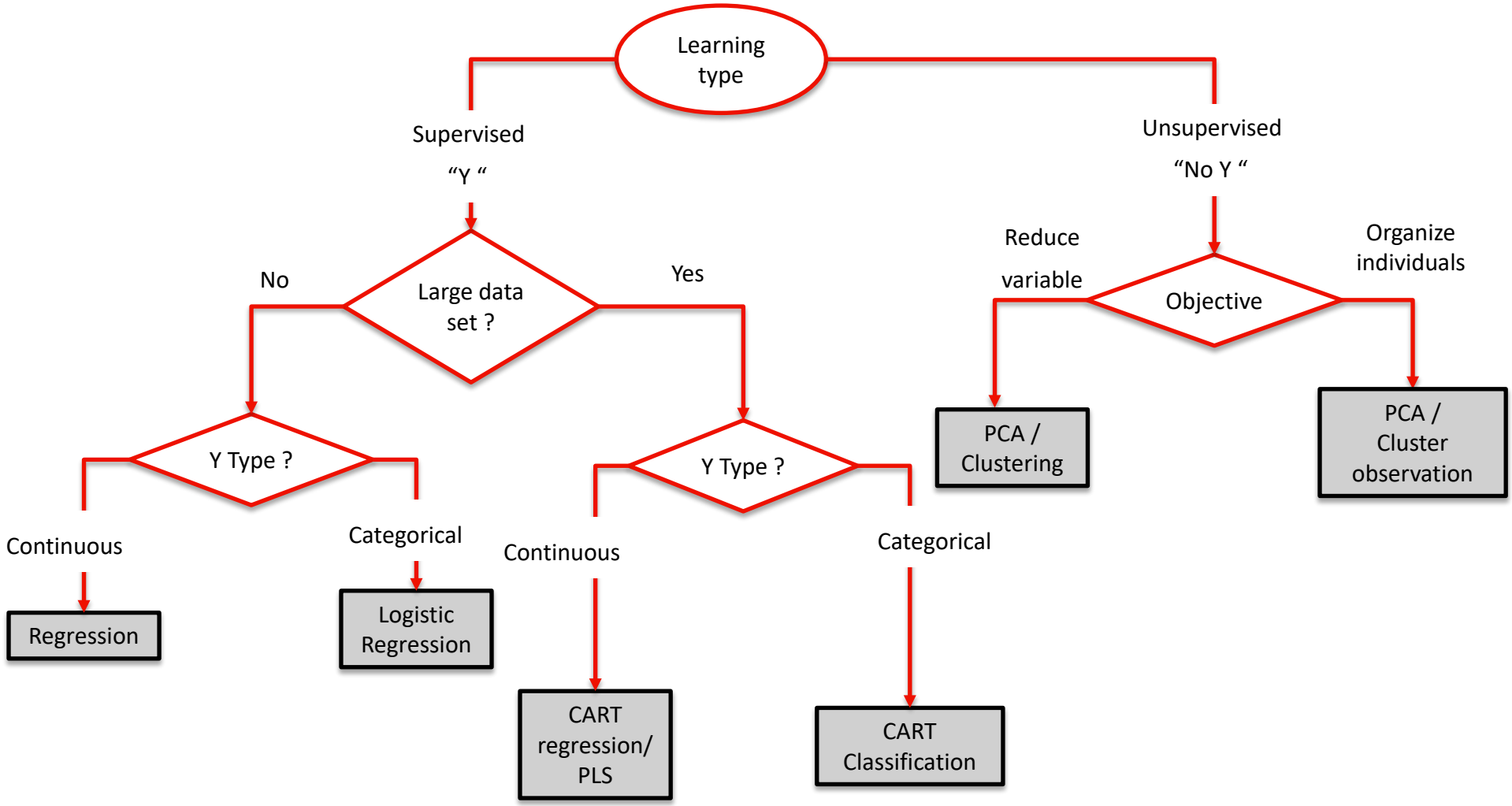
- CART Classification
- Logistic Regression

Continue response

- CART Regression
- PLS **

*Principal Components Analysis
** Partial Least Squares

Machine learning Roadmap



Fields to be used and Advantages and Benefits coming from Predictive Analytical Module (PAM)

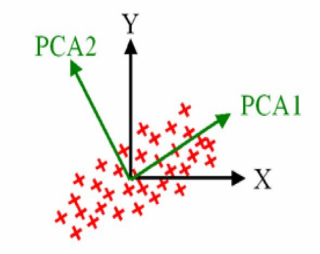
Fields to be used

Recently, due to the significance of Industry 4.0, the manufacturing industry is developing globally. Conventionally, the manufacturing industry generates a large volume of data that is often related to process, line and products.

Use decision tree (CART) to answer question like :

- What is the root cause of process defects ?
- What factors contribute to excess variation in my process?
- What are the conditions that could cause machine degradation ?
- What parameters could increase my Yield, my efficiency ?

Advantages and Benefits coming from Predictive Analytical



STATS

Advantages

- Well known by the health authorities
- Very well deployed at Takeda
- User-friendly graphic interface designed
- Suitable for use with most data analyzing
- Reduce overfitting

Disadvantages

- Difficult to use with large data set
- Treatment of the missing data

PAM

Advantages

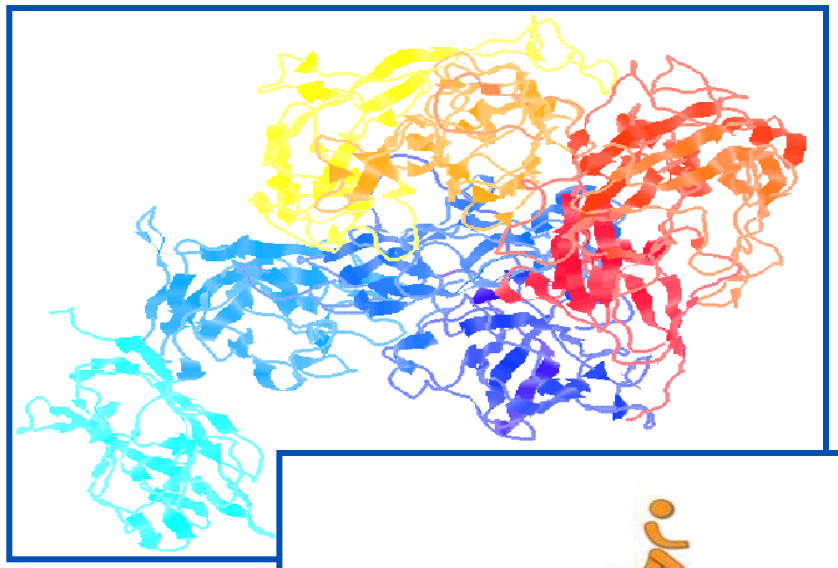
- Less effort
- Not require normalization and scaling of the data
- Very visual and straightforward.
- Very intuitive
- Relative variable importance
- Additional tool to the regular stats
- Treat well missing data

Disadvantages

- Small change in the settings can cause a large change in the structure of the tree

Fields to be used and case study (Advate)

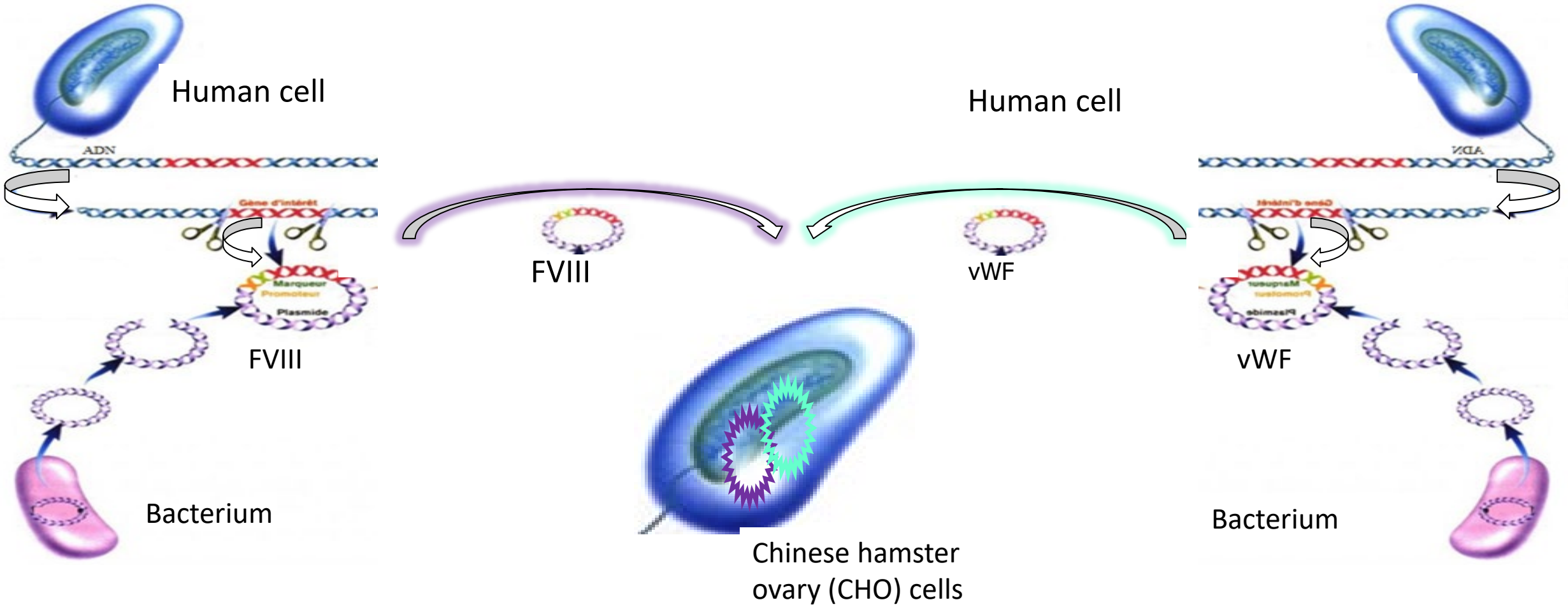
About ADVATE



ADVATE temporarily replaces the clotting factor VIII that's missing or low in the blood.
In 2003, ADVATE became the first recombinant factor VIII treatment free of blood-based additives.

What does recombinant mean?
Today, clotting factor VIII concentrates can be made without human plasma. These types of infusions are called recombinant clotting factors. Recombinant factor VIII was one of the first treatments approved for hemophilia that didn't pose the threat of transmitting a blood-borne virus.

Genetic Modification of Chines Hamster Ovary cells

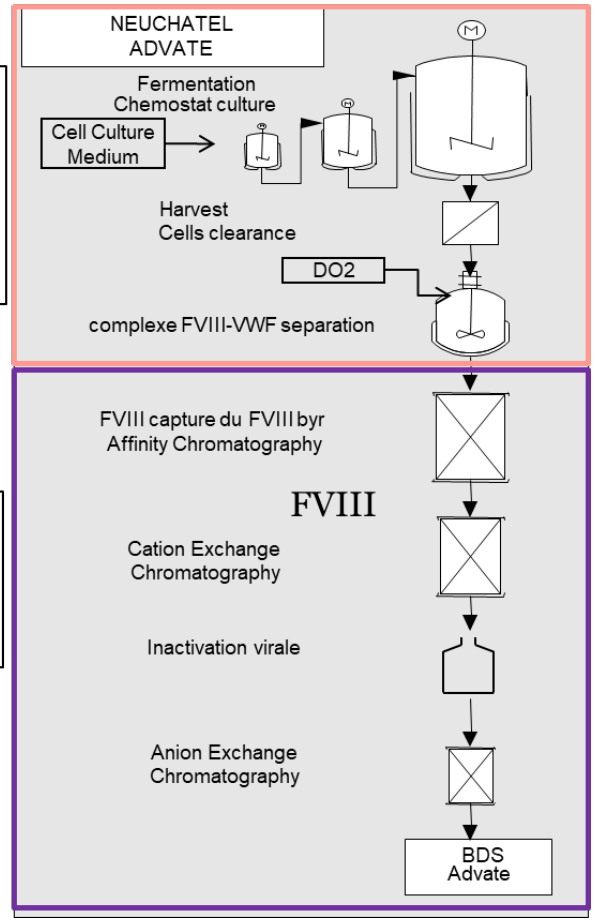


ADVATE is containing recombinant human antihaemophilic factor VIII, which is synthesised by **a genetically engineered Chinese hamster ovary (CHO) cell line**.

Manufacturing parameters

The biopharmaceutical industry, however, has accrued greater benefits as evidenced by the use of MVDA tools both in upstream (Culture cell) as well as downstream processes (Purification).

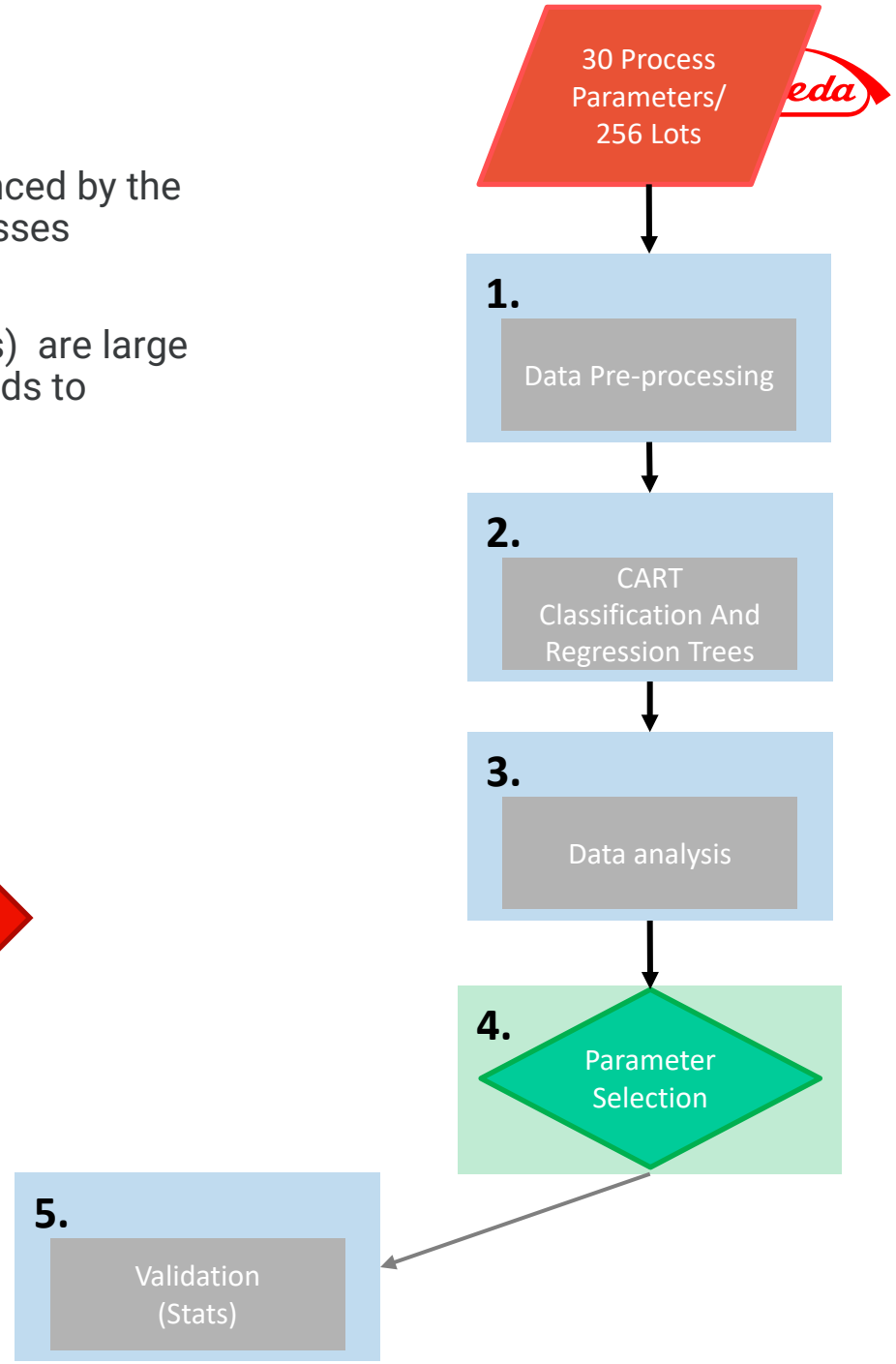
The datasets downloaded from the MES system (productions factors and probes) are large and complex with multivariate interactions and require the use of adapted methods to understand the main parameters on the Yield



30 Parameters

- 3 Column MAB, SOS, MOQ
  - Elution Profile
  - Alignment
  - Process duration
  - Storage duration
  - Resin N° of use
  - Volume loaded
  - Column ID
  - Elution gradient slope
  - Conductivity
  - Hold time
  - Mixing duration
  - ……..

256 batches

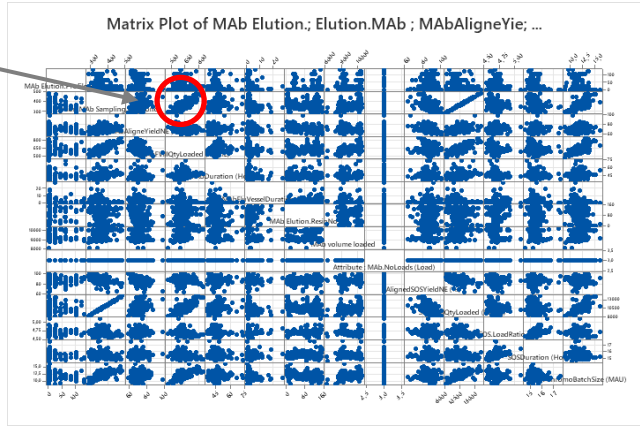


PRE-PROCESSING prepares data for the models

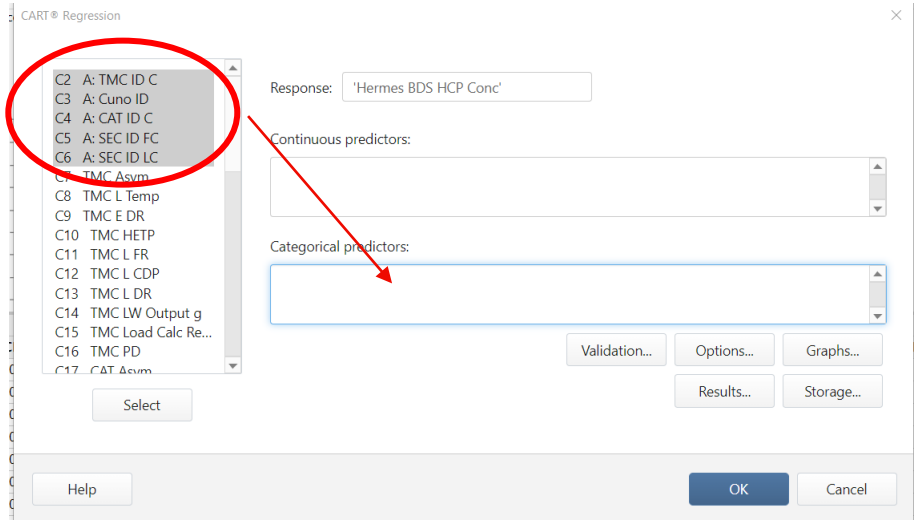
- Feature engineering : Sit with the SME to review the dataset (few rounds): normalize parameters, add features describing the process, etc…
- Check missing values
- Identify the Input and Output
- Remove obvious correlations between factors and the Y
- Look at your data!

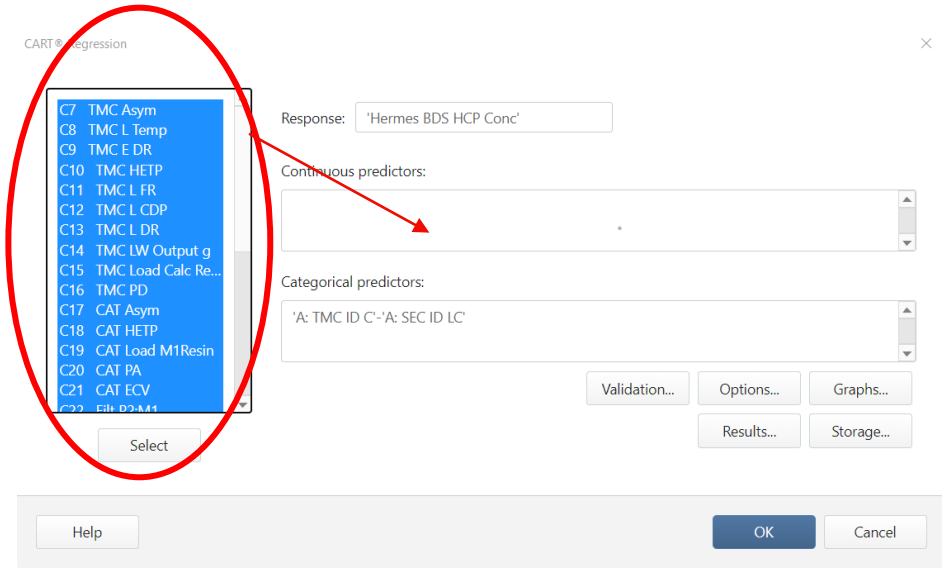
AJ13 | 18,6

	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ
1	MOQ WashConduc (mS/cm)	Y : ChromoBatchSize (MAU)	Y: .ClotBatchSize (MAU)	Y : MOQ Sampling. CHO/mL	Y: MOQ Sampling. CHO ELISA (µg/mL)	Y: MOQ Sampling. ChromoFVIIIConc (IU/ml)	Y : Sampling. ChromoSpecAct (IU/mg Protein)	Y : CLOTCHO ELISA (µg/1000 IU FVIII)	Y : ClotCHO (µg/1000 IUrAHF)	Y : MOQ Sampling. ClotFVIII Conc (IU/ml)	Y : MOQ ClotSpec Act (IU/mg Protein)	Y : MOQ ClotvWFUnitConc	Y : MOQ Sampling. DegPolymerisation (%)	Y : MOQ Sampling. HPLC.43KD (%)	Y : MOQ HPLC.50KD (%)	Y: MOQ HPLC.73KD (%)	Y : MOQ HPLC.B-domain (%)
2	18,7	11,1248	11,50138	6,849	9,95	9601	6008	1	0,7	9926	6212	0,3	101	0,201	0,232	0,391	0,176
3	18,5	12,63869	13,21343	7,101	12,25	10929	6314	1,1	0,6	11426	6601	0,2	101	0,202	0,233	0,393	0,172
4	18,6	13,84485	14,36987	7,673	14,2	11972	6633	1,1	0,6	12426	6884	0,2	102	0,202	0,234	0,391	0,173
5	18,6	15,29543	16,46225	7,305	18,85	13161	6451	1,3	0,5	14165	6944	0,2	101	0,2	0,231	0,39	0,179
6	18,9	13,6463	14,44464	6,736	15,3	11743	6488	1,2	0,5	12430	6867	0,2	103	0,201	0,234	0,396	0,17
7	19	13,96064	15,33667	7,352	17,25	11079	6169	1,4	0,6	12171	6777	0,3	104	0,202	0,239	0,398	0,161
8	19,2	12,87202	13,98996	7,917	13,1	10236	6137	1,2	0,7	11125	6670	0,4	105	0,203	0,234	0,397	0,165
9	18,4	13,73571	14,66831	8,387	19,7	11827	5934	1,6	0,7	12630	6337	0,3	103	0,2	0,233	0,402	0,166
10	18,8	13,59868	14,68588	8,371	17,3	11720	6349	1,4	0,7	12657	6856	0,3	101	0,203	0,235	0,402	0,16
11	18,8	15,00544	15,83069	8,56	19,6	12928	6503	1,4	0,6	13639	6861	0,3	99	0,2	0,235	0,404	0,162
12	18,5	13,47003	14,70648	8,264	15,9	11624	6454	1,3	0,7	12691	7047	0,3	99	0,207	0,238	0,398	0,158
13	18,6	13,98812	14,72574	8,151	17,55	12099	6470	1,4	0,6	12737	6811	0,3	100	0,204	0,239	0,397	0,161
14	19	13,40621	14,65373	7,143	11,85	11563	6774	0,9	0,6	12639	7404	0,3	100	0,2	0,232	0,404	0,164
15	18,7	14,30264	15,29034	7,729	17,8	12352	6296	1,3	0,6	13205	6730	0,3	101	0,199	0,236	0,408	0,158
16	18,6	13,95772	14,13494	7,141	13,4	12050	6728	null	0,6	12203	6814	0,4	101	0,202	0,236	0,411	0,152
17	18,9	13,93375	14,30715	7,556	16,1	12053	6456	1,3	0,6	12376	6629	0,3	99	0,203	0,235	0,411	0,152
18	18,8	12,50109	13,10195	6,381	12,7	10798	6736	1,1	0,6	11317	7060	0,3	104	0,204	0,236	0,411	0,148
19	18,4	13,25099	14,23124	7,367	15,35	10544	6383	1,4	0,7	11324	6855	0,3	100	0,203	0,231	0,405	0,161
20	18,5	12,16332	12,57433	5,903	11,7	9677	6530	1,2	0,6	10004	6750	0,4	100	0,206	0,237	0,407	0,15
21	18,2	8,91936	9,251478	6,81	13,7	7090	6269	1,9	0,9	7354	6502	0,3	100	0,198	0,229	0,407	0,165
22	19	8,37998	8,726347	6,495	12,45	7234	5998	1,7	0,9	7533	6246	0,4	101	0,202	0,232	0,405	0,161

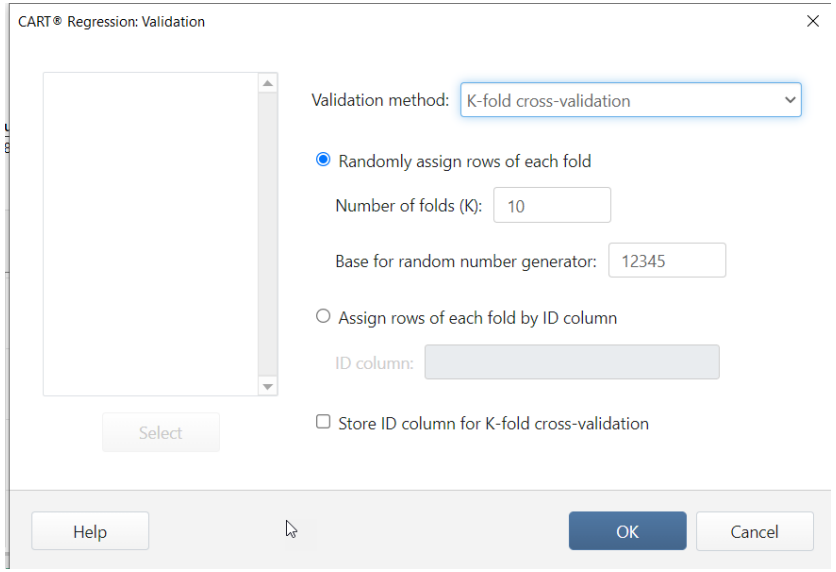


CART Regression - I





- Response : Hermes BDS HCP Conc
- Categorical predictors : Predictors start by A
- Continuous predictors : from C7 up to C37
- Options > Node Splitting Method: Least Squared Error



- Validation > Validation Method: K-fold cross validation

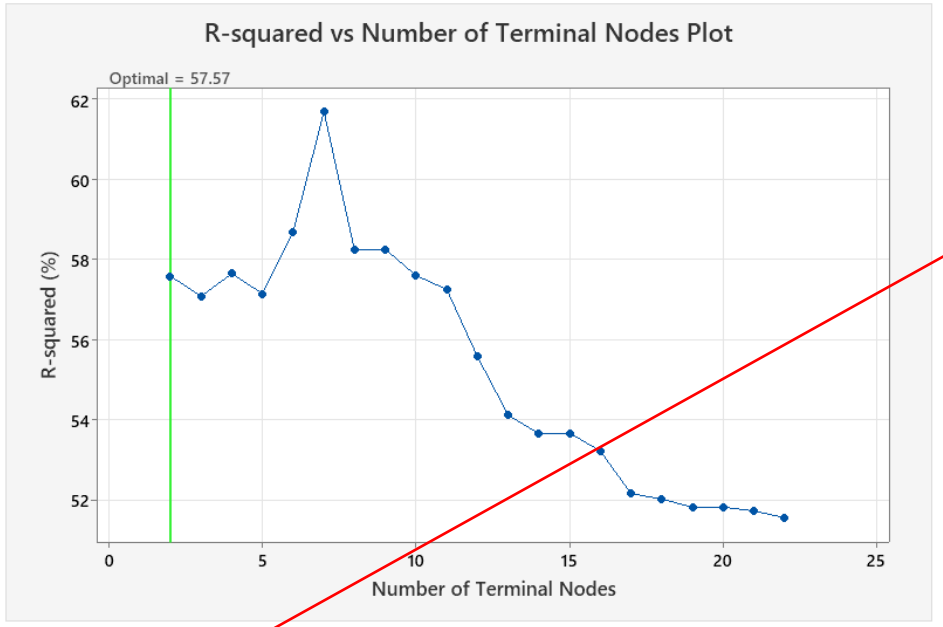
CART Regression - II

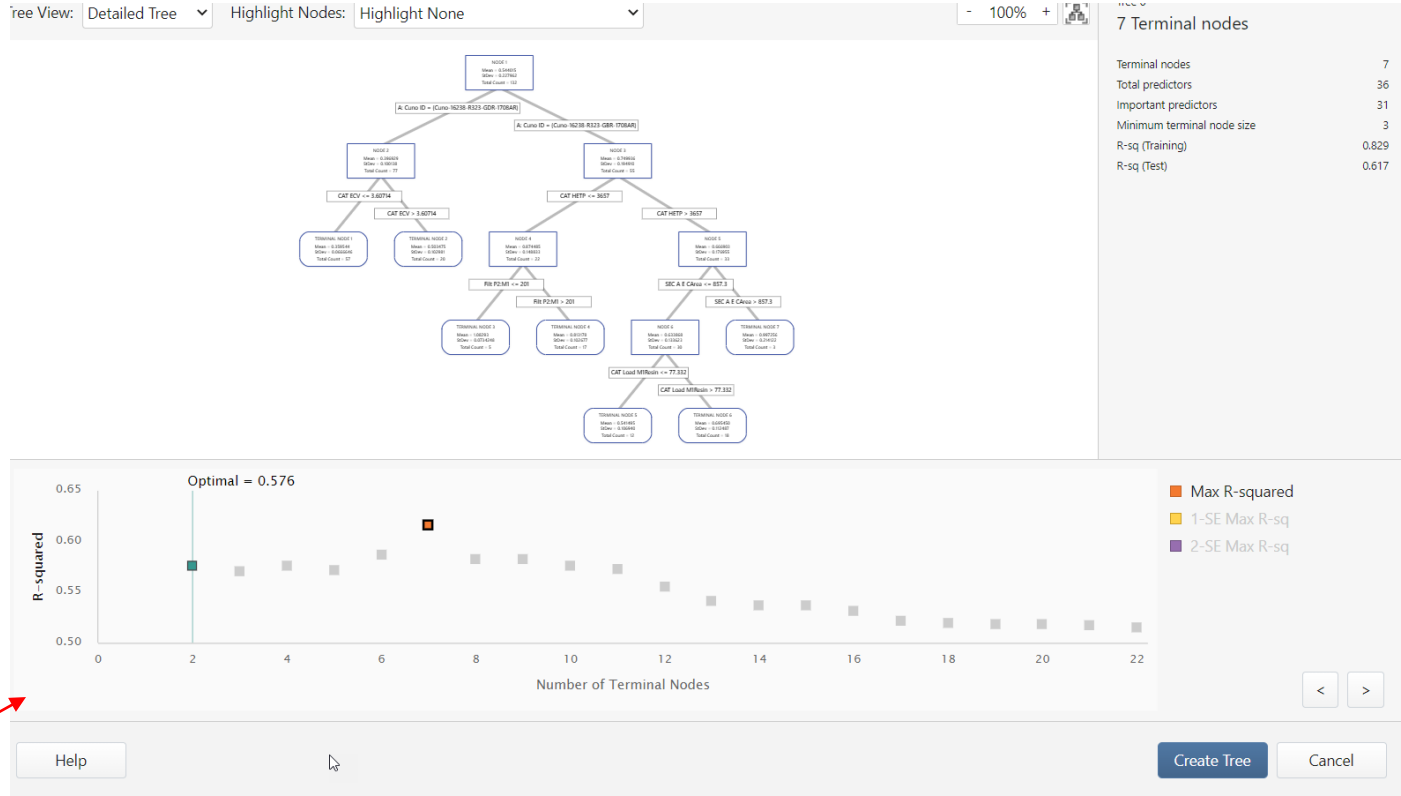
Method

Node splitting	Least squared error
Optimal tree	Within 1 standard error of maximum R-squared
Model validation	10-fold cross-validation
Rows used	132

Response Information

Mean	StDev	Minimum	Q1	Median	Q3	Maximum
0.544015	0.228831	0.252041	0.362348	0.465016	0.687176	1.29841





- Choose 7 Nodes to maximize R-sq (61.7%) and Create Tree

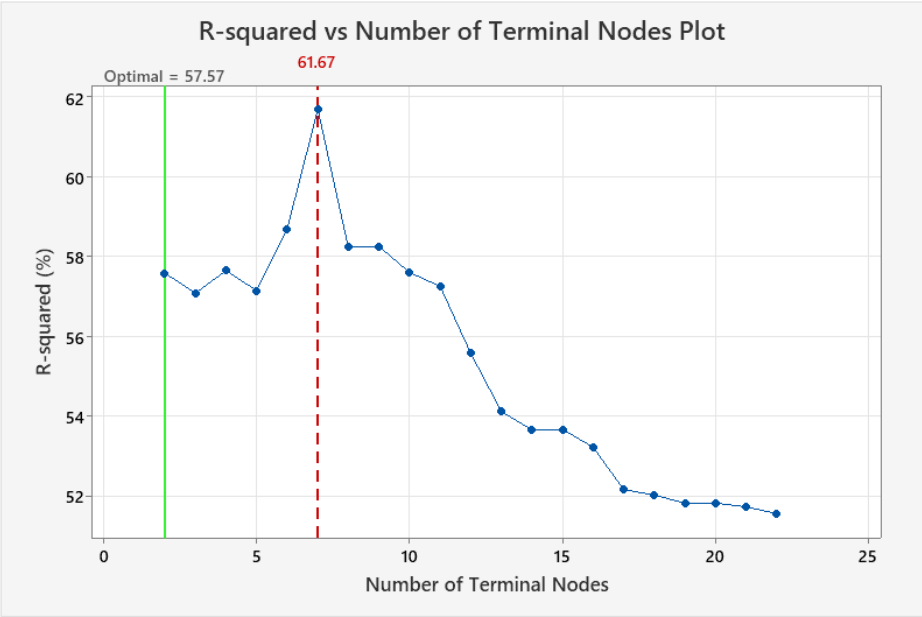
CART Regression - III

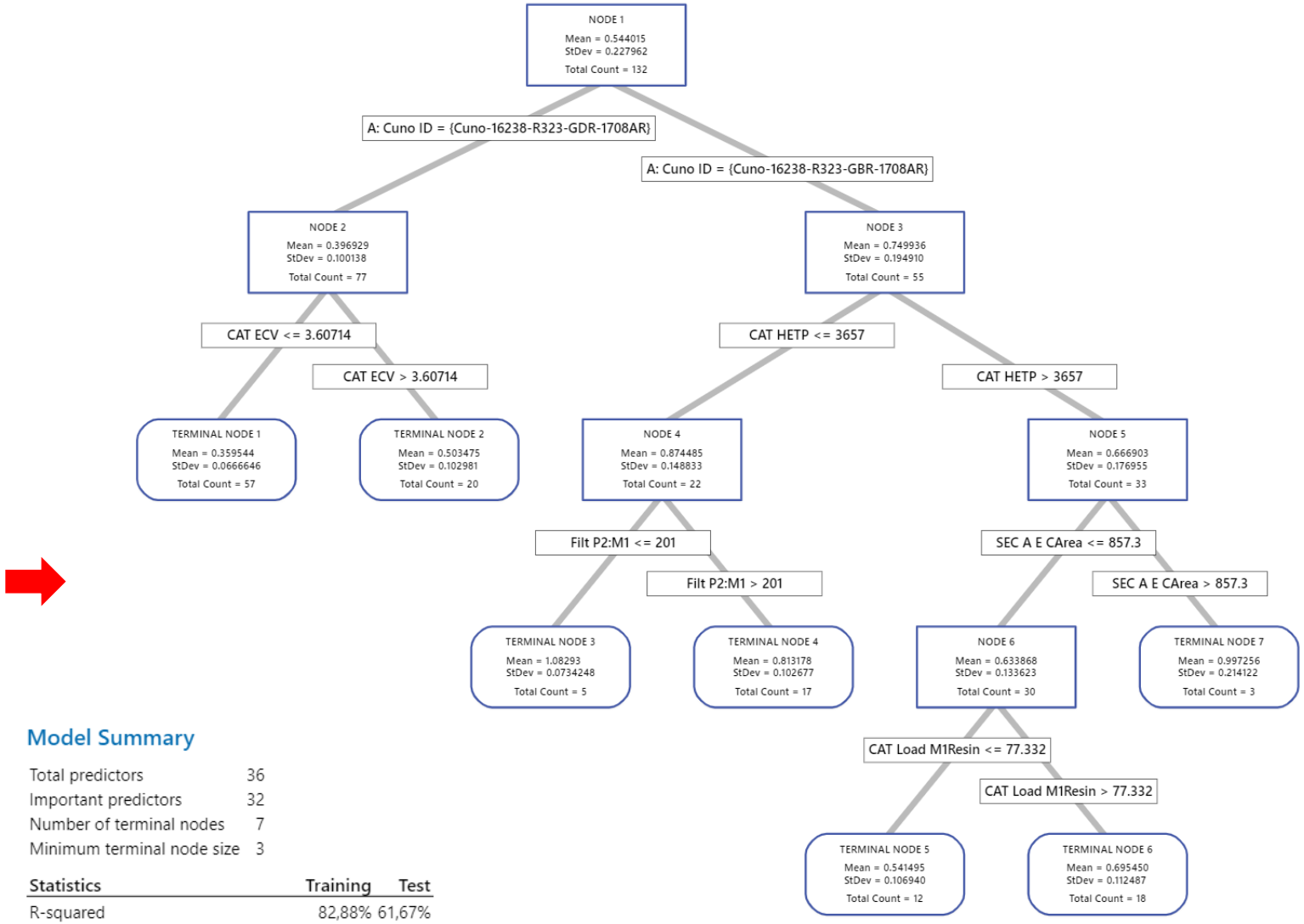
Method

Node splitting	Least squared error
Optimal tree	Within 1 standard error of maximum R-squared
Model validation	10-fold cross-validation
Rows used	132

Response Information

Mean	StDev	Minimum	Q1	Median	Q3	Maximum
0.544015	0.228831	0.252041	0.362348	0.465016	0.687176	1.29841



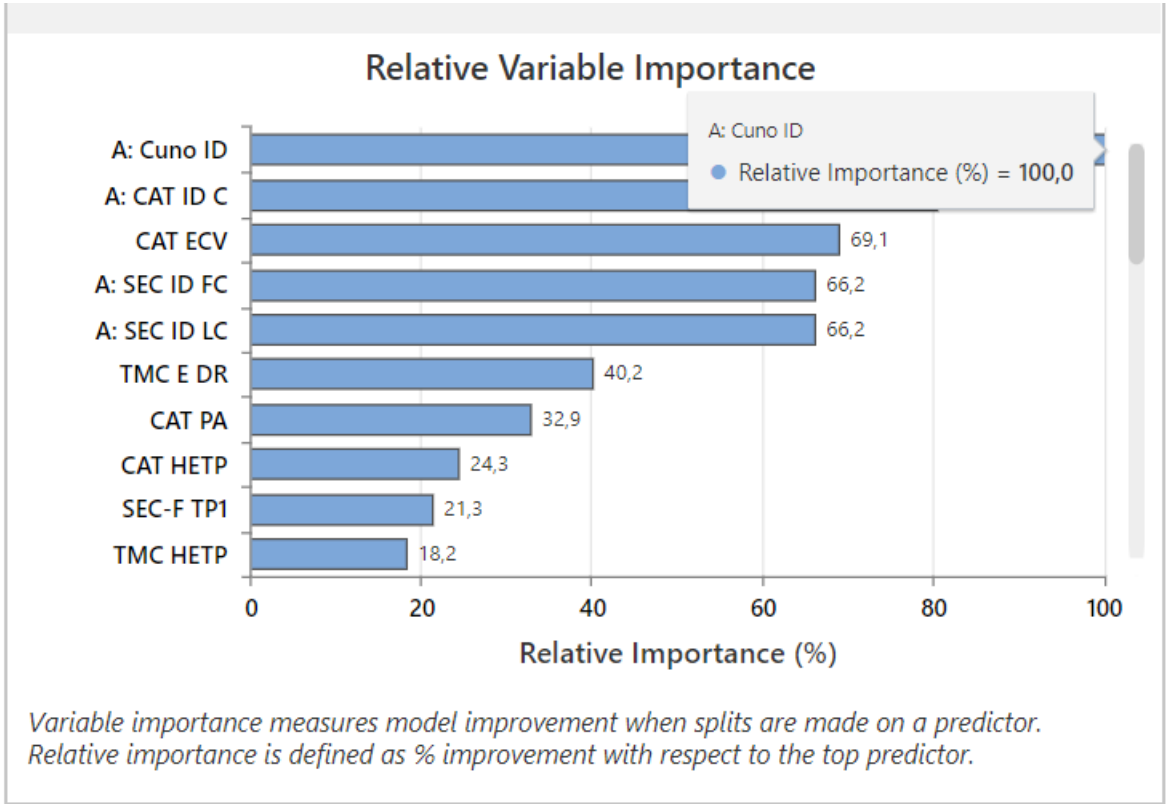


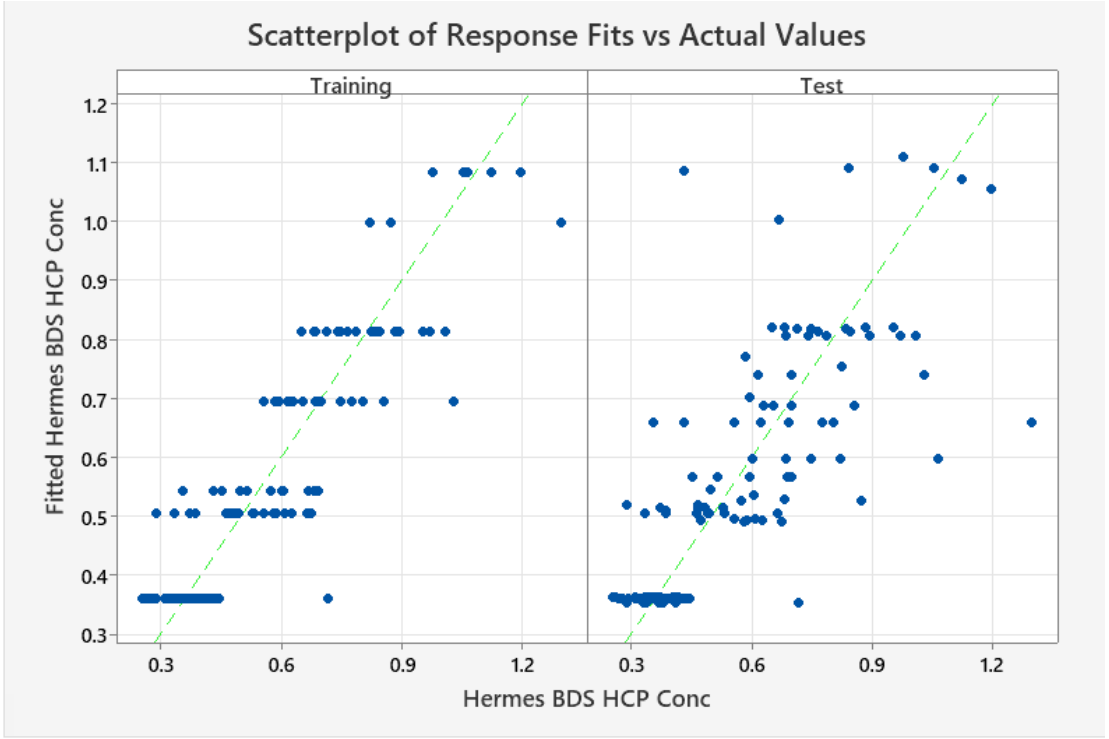
Model Summary

Total predictors	36
Important predictors	32
Number of terminal nodes	7
Minimum terminal node size	3

Statistics	Training	Test
R-squared	82,88%	61,67%
Root mean squared error (RMSE)	0,0943	0,1411
Mean squared error (MSE)	0,0089	0,0199
Mean absolute deviation (MAD)	0,0687	0,0914
Mean absolute percent error (MAPE)	0,1323	0,1676

CART Regression - IV

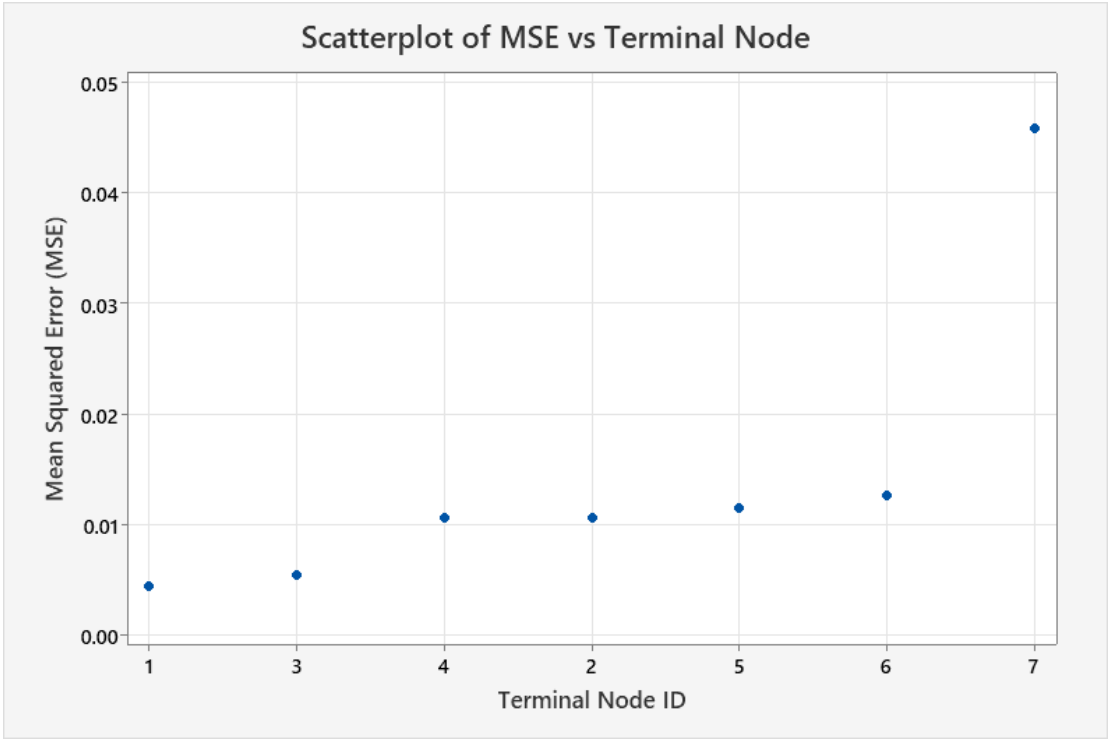




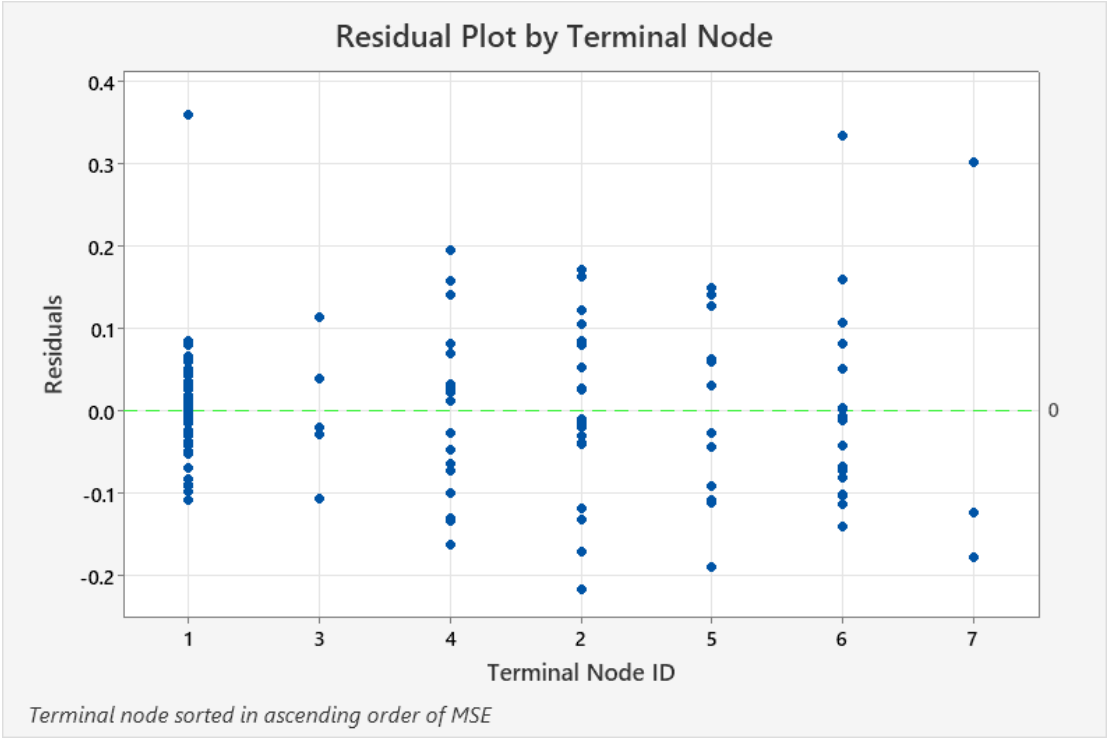
- Top 9 X's Relative importance to Cuno ID (e.g., CAT ID C is 80.4% as important to Cuno ID for predicting Hermes BDC HCP Conc)

- Line goes through middle of training and test data
- R-sq (Training) = 82.9%; R-sq (Test) = 61.7%

CART Regression - V



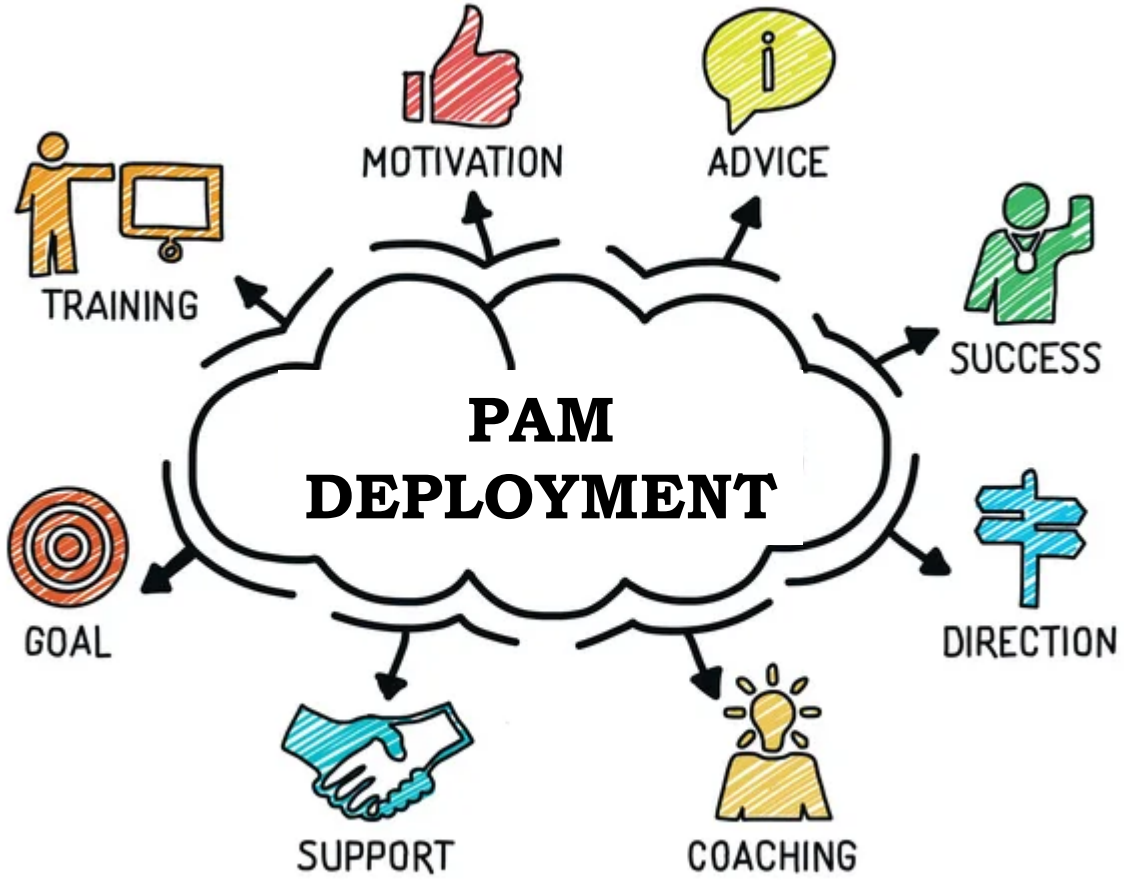
- Large change in MSE between node 6 and node 7



- Residuals are fairly uniform

Predictive Analytical Module principle

PAM Strategy deployment



How to democratize and deploy the usage of PAM

Direction

- Hoshin kanri
- Visual boards
- KPI

Training

- New modules
- Case study
- BB and Manufacturing Science attendees
- Education Hub (Minitab)

Coaching

- No Training Without Coaching, No Coaching Without Training
- Develop adult skills coaching

Motivation

- Individual interests and career goals
- Get examples of their workplace
- Recognition and celebrate

Success

- Community of practice
- Share success on Takeda media (Yammer)
- Get certification