Paper ML14

Is Artificial Intelligence our destiny? Data-driven Decision-making in a Dynamic Landscape using SAS® Viya®

Aman Bahl, Syneos Health, Burlington, Canada
Hrideep Antony, Syneos Health, North Carolina, USA

ABSTRACT

Health is not something that happens a couple of times a year when you visit doctors. Health is something that happens every day. It is happening right now. It is cumulative of choices, lifestyles.

Artificial Intelligence (AI) is making its way into the realm of clinical trials. While a lot of discussions seems to be centered on clinical trial recruitment and using AI to mine electronic medical records (EHRs), but it seems to only scratch the surface. Experts predict monitoring drug adherence, pre-emptive risk monitoring, decision-making, diagnostics, and process optimization are some of the other areas where the technology is expected to make an impact in future.

In this paper, we will explore the benefits of SAS Viya from a machine learning and AI perspective. We will briefly cover various tasks in the analytics life cycle and show how to create a decision tree model as an example using SAS Viya Studio. The various types of output created will be described along with some of the SAS Viya machine learning procedures such as decision trees, boosting trees and random forest models.This paper will summarise a few key points that are important to consider while implementing Artificial Intelligence in healthcare and how it will revolutionize the future of healthcare.

INTRODUCTION

Why have medications that are only for the general population, when it can be customized for each individual? Artificial Intelligence algorithms will enable physicians and hospitals to better analyze data and customize their health care to the genetics, lifestyle, and surroundings of each patient. From diagnosing tumors to deciding which cancer treatment will work best for an individual, AI will drive the personalized medicine revolution in the future.

As noted by computer scientist Kai-Fu Lee Artificial Intelligence (AI) is going to change the world more than anything in the history of mankind, more than electricity (Reference [1])

Artificial Intelligence is making its mark in healthcare. From early detection to improved diagnosis, AI is positively contributing to the betterment of humanity. In some areas, it is already being used, and there are areas where we can see the progression of AI in the future.

AI adoption in healthcare is very important due to the following reasons:

- ✓ Increase in digital data and difficulties in handling a large amount of patients' data.
- ✓ Increase in the number of diseases and the need to understand and diagnose better.
- ✓ Speedy diagnosis of patient's conditions to help the providers in administering effective and efficient treatments.
- ✓ Reduction in the costs and improvement in efficiency by treating a larger number of patients.

SAS Visual data mining and machine learning come equipped with an incredibly broad set of modern machine learning, deep learning, and text analytics algorithms that are all accessible within a single environment. This makes the AI implementation using the SAS Viya ideal for all organizations, given its diverse analytics capabilities that include clustering, different modes of regression, random forests, gradient boosting models, support vector machines, natural language processing, and topic detection etc. SAS users not only gain access to a platform that is highly functional but one that is equipped with powerful predictive and decision-making capabilities.

SAS Viya provides all the necessary components in an absolute and intuitive manner, thus managing the full analytics life cycle (as shown in Figure 1) with less effort.

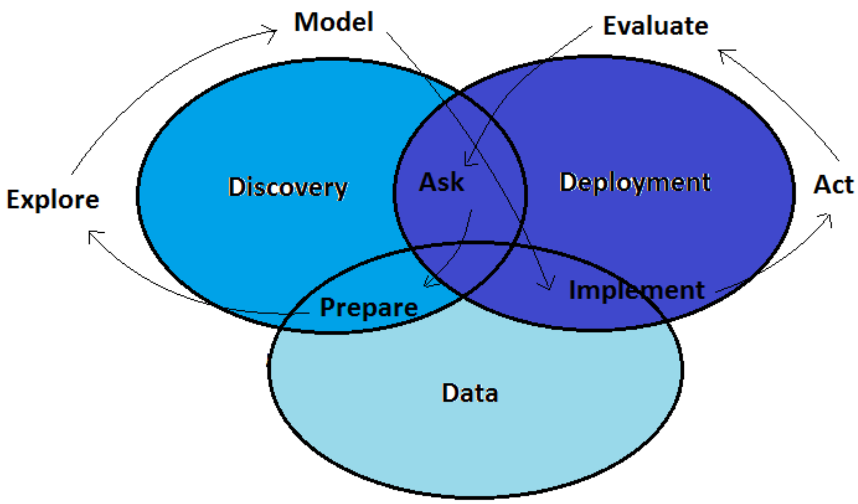


Figure 1: Analytics life cycle

The initial step in the analytics life cycle is to prepare the data which might involve combining data from multiple sources into a single source table, cleansing and augmenting the data, partitioning the data, etc. Once the data is prepared, it can be then explored and visualized by using SAS® Visual Analytics.

We can then take models created in SAS Visual Analytics and create analytics pipelines in SAS Visual Data Mining and Machine Learning. These pipelines can be then augmented with additional models, continually tuning and comparing them until we have built the best model for the data.

The model can be then be deployed from SAS Visual Data Mining and Machine Learning by publishing the model directly to a location that contains the data we want to score. The model can also be registered in SAS Model Manager to take advantage of model versioning and governance.

As the data evolves, we can use the SAS Model Studio to monitor our model's performance over time so that we can make decisions about whether we should rebuild the models. If we decide to rebuild our model, we can go back to the beginning of the analytics life cycle. All these steps are contained within one unified analytics environment. Additional details on analytics life cycle of SAS Viya can be found at reference [2].

This paper will describe some of the discovery tasks in the analytics life cycle i.e. selecting an Algorithm, improvise the model, optimize the complexity of the model, regularize and tune the hyperparameters of the model, build ensemble models etc.

What is SAS Visual Data Mining and Machine Learning?

SAS® Visual Data Mining and Machine Learning consist of data mining tools that enable us to build and customize predictive models. The access to its features is enabled by a Model Studio which is a web-based visual interface. The overall process flow of SAS Visual Data Mining and Machine Learning is shown in Figure 2 below.

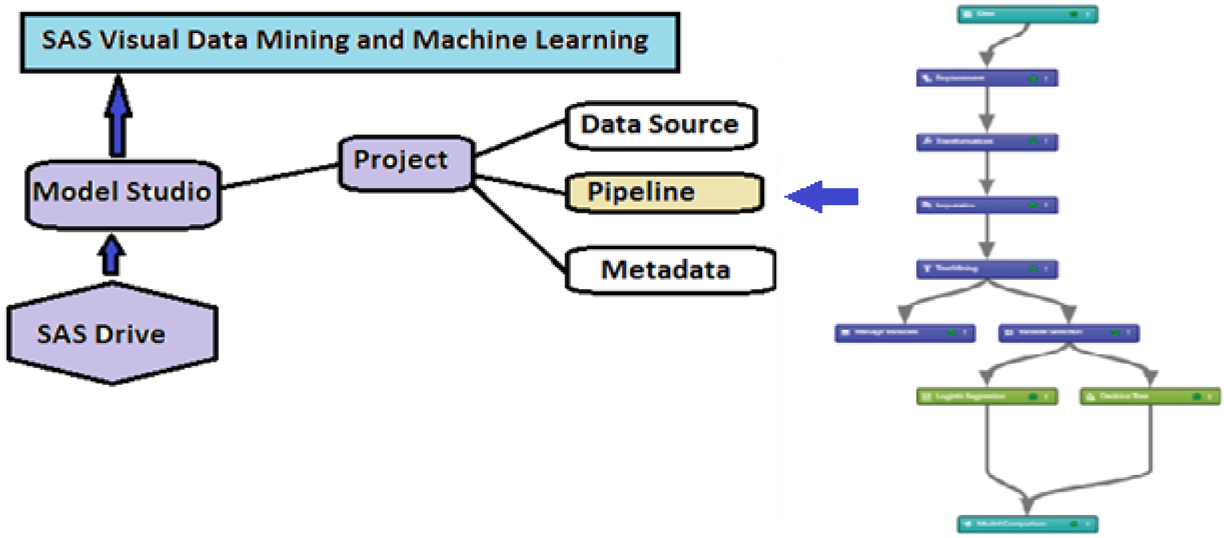


Figure 2: SAS Visual Data Mining and Machine Learning process flow

The first thing you notice is that all the working projects are well organized in one place, in SAS Drive.

One can create or open Model Studio projects from the SAS Drive. The nodes panel allows the user to drag and drop a Node into the project pipeline to include a new analytical task in the machine learning predictive model.

Some of the components of the model studio will be further explained in this paper.

The initial task of the discovery phase of the analytics life cycle which is to select an algorithm is explained below using a decision tree-based algorithm.

Create a Decision Tree model using SAS Viya Studio

As noted by Niklaus Wirth, Swiss computer scientist "The possible solutions to a given problem emerge as the leaves of a tree, each node representing a point of deliberation and decision" (Reference [3]).

Methods like decision trees, random forest, gradient boosting are being popularly used in all kinds of data science problems. Tree-based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree-based methods empower predictive models with higher accuracy, stability and ease of interpretation.

Following are some of the advantages of using decision tree models in an organization:

- ✓ Trees require less data preparation than other machine learning algorithms.
- ✓ They are also easy to explain and interpret.
- ✓ Trees follow a decision split if-then logic and can be represented in a tree like graphical representation.

We selected one of the SAS Viya Studio pipeline templates in the below Figure 3 and added a decision tree model based on default parameters.

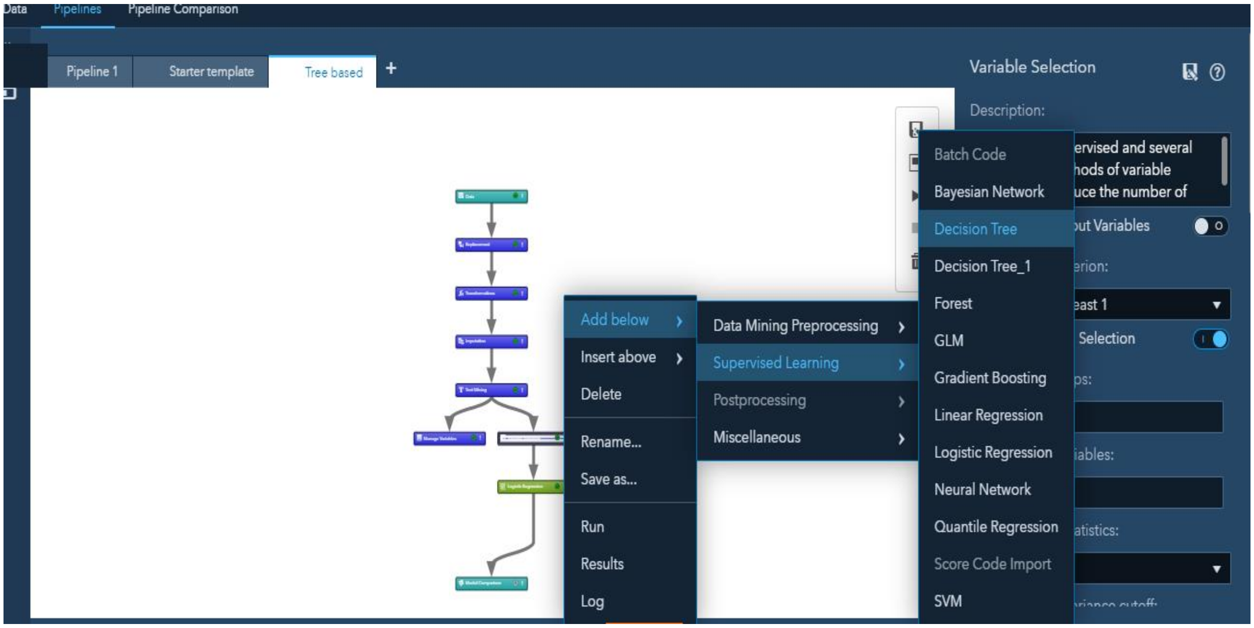


Figure 3: Add Decision Tree model to the tree pipeline

A Decision tree node gets added after variable selection node. It creates parallel path in the pipeline with logistic regression node as shown in Figure 3 above. We can keep all the properties of decision tree node in default state.

A Decision tree model is run by right-clicking on the Decision tree model and selecting 'Run' below. Once the tree has been run successfully, a green checkmark appears on the nodes as shown in Figure 4.

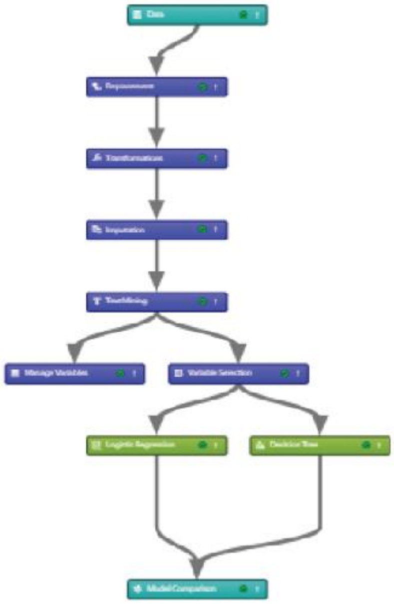


Figure 4: Run the Decision tree node

After the Decision tree model has been run, we can right click on the Decision tree node and select the results as shown in Figure 5 below.

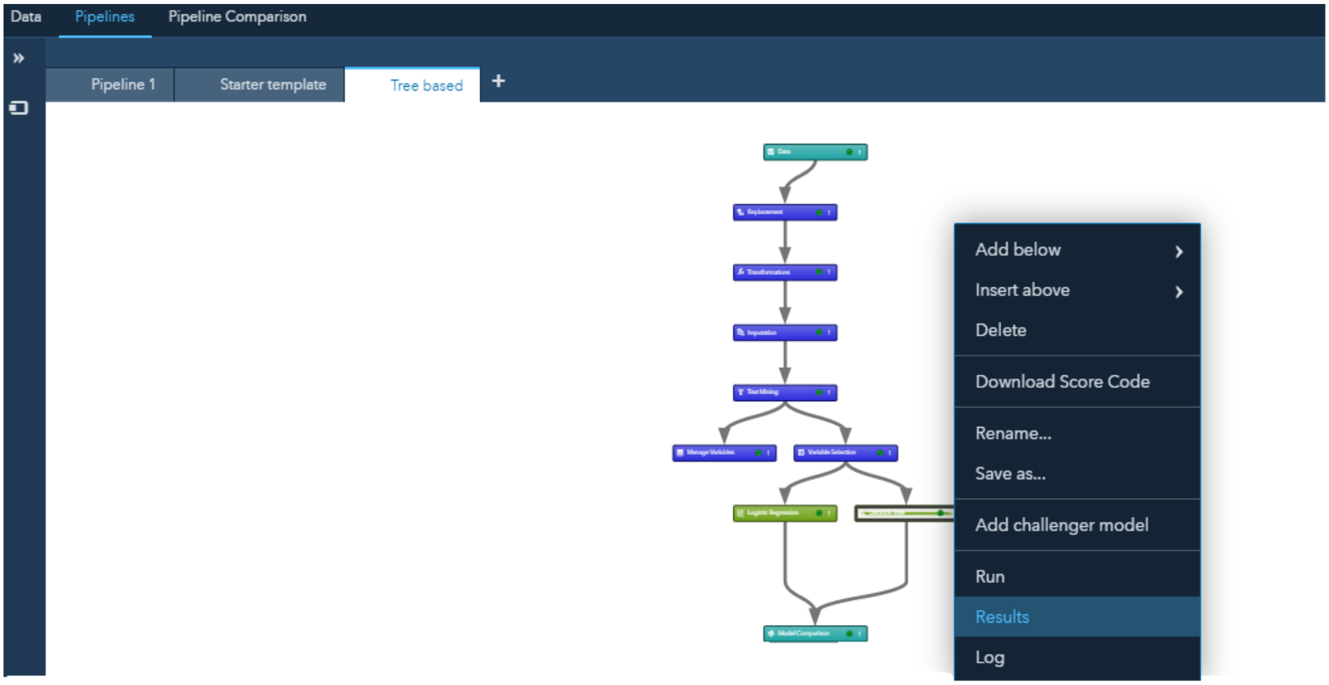
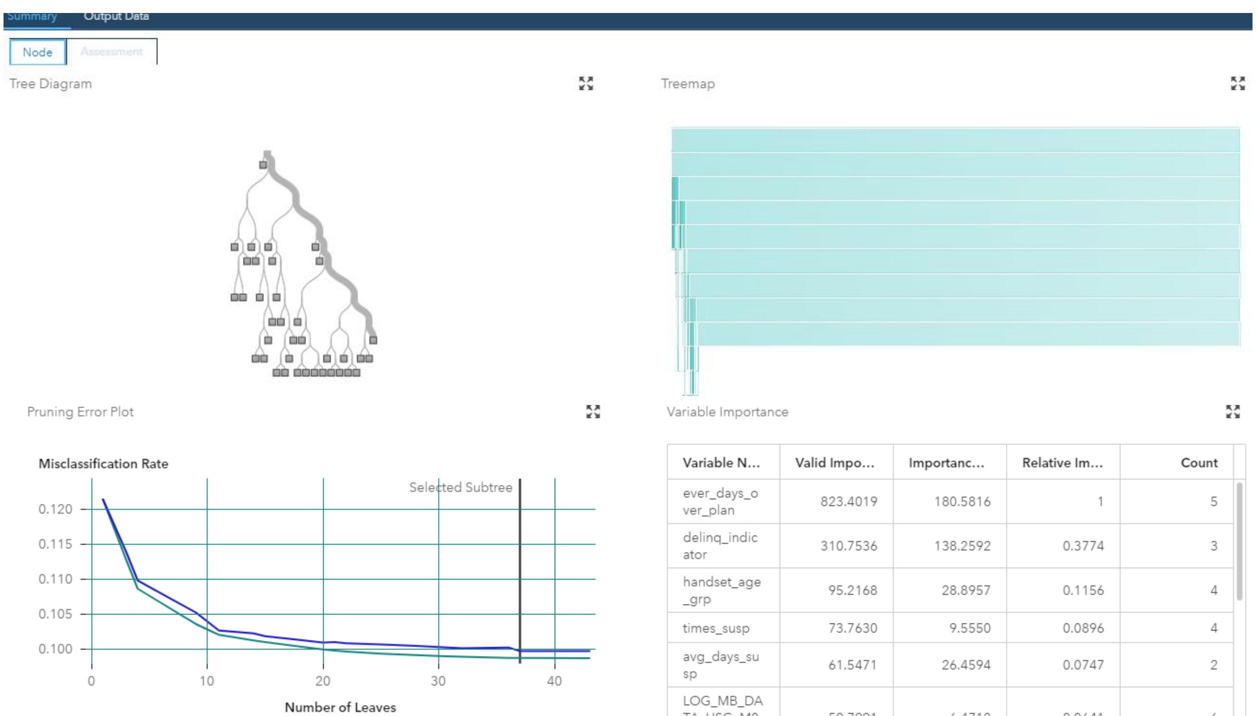


Figure 5: Select Results for the Decision Tree model run

The **Node and assessment** output tabs can be seen in Figure 6 below.

Node tab

The Node tab displays tree diagram, treemap, pruning error plot, variable importance, node score code, score inputs, score outputs, path ep score code, ds2 package code, train code, properties and output.



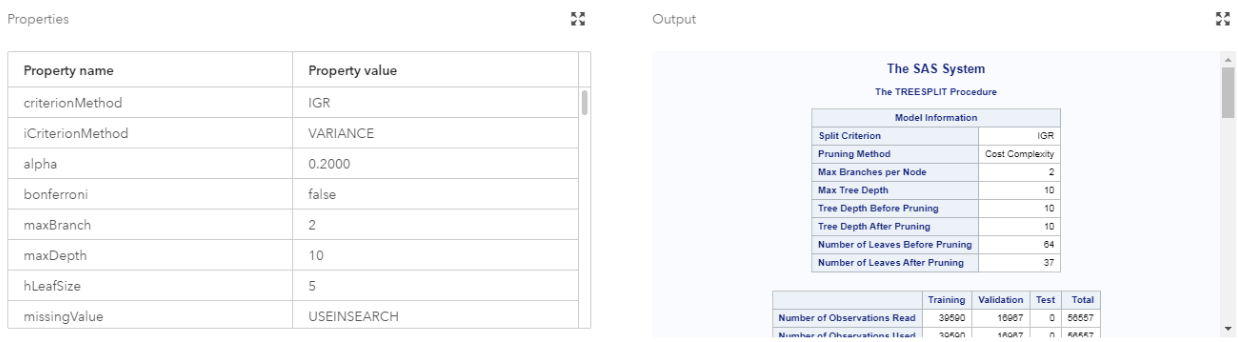


Figure 6: Results - Node and assessment tabs

Due to the nondeterministic behavior of SAS Viya, the results may vary slightly every time the node runs.

The tree diagram shown in Figure 6 displays the complete tree structure along with the depth of the tree and all end leaves.

The **pruning error** plot as shown above is based on the misclassification rate because of the binary target in this example. It shows the change in misclassification rate on training and validation data as the tree grows or when more leaves are added to the tree.

The **variable importance** table shows the final variables selected by the decision tree and their relative importance. The most important variable has its relative importance as 1 and all others are based on the most important input.

The **node score** window shows the final score code that can be deployed in production.

The **train code** window shows the code that can be used to train the model based on different datasets or in different platforms.

Figure 7: Output: TREESPLIT procedure

The output window in Figure 7 above shows that the TREESPLIT procedure is the underlying procedure for this node. Also shows the final decision tree parameters, the variable importance table and pruning iterations.

Assessment tab

The Assessment tab displays Lift Reports, ROC Reports and Fit Statistics. The Figure 8 below shows Assessment tab results.

Data Role	Partition In...	Formatted ...	Sum of Fre...	Average Sq...
TRAIN	1	1	39,590	0.0876
VALIDATE	0	0	16,967	0.0886

Figure 8: Results – Assessment tab

The following three tables are shown in assessment window in Figure 8 above.

- ✓ Lift Reports: A lift value greater than one indicates that the model is an outperforming random model.
.
- ✓ ROC Reports: The receiver operating characteristic curve (ROC) plots display the Sensitivity against 1-Specificity for varying cutoff values. Sensitivity is defined as the true positive rate and 1-Specificity is defined as the false positive rate.

- ✓ Fit Statistics: Compares the performance of the final model in the training and validation dataset.

Once the Decision tree model has been selected the other essential discovery tasks in SAS Viya are as follows:

- ✓ Improve the model: This process involves modifying the Structure Parameters. It needs to be noted that modifying the model does not always result in better performance.

- ✓ Optimize the complexity of the model: The Maximal tree is the starting phase for optimizing the complexity of the model through a process called pruning.

- ✓ Regularize and tune the hyper parameters of the model

- ✓ Build ensemble models. Anyone of the individual models might not be the most effective solutions for a specific problem and dataset. In this case the user can consider building non-ensemble model. An ensemble model is an aggregation of multiple models.

Once the model has been created the next step is to fine-tune the model that will fit the need and select the most fitting one using some of the key beneficial features available in SAS Viya.

A few of the features and benefits of SAS Viya are briefly described below:

Automatic hyperparameter tuning Feature in SAS Viya

The Automatic hyper parameter tuning feature helps the user to find the optimal hyper parameters of a Machine Learning algorithm that can be used to build a model in your pipeline. The auto tune feature saves you lots of time when you search for the best settings to create the best possible model.

Automated assessment visualizations of the predictive models

Once the model has been built, the Model studio automatically creates several interactive, attractive and comprehensive assessment visualizations in output graphs and tables. They provide many fit statistics that can be used as a benchmark to assess and compare models' performance.

Instant choice of the best model in your project.

Model Studio automatically chooses the best fitting model among all the models that have been built and tested in the current project. In a single pipeline, Model studio automatically adds a 'Model Comparison' node as shown in Figure 9 below.

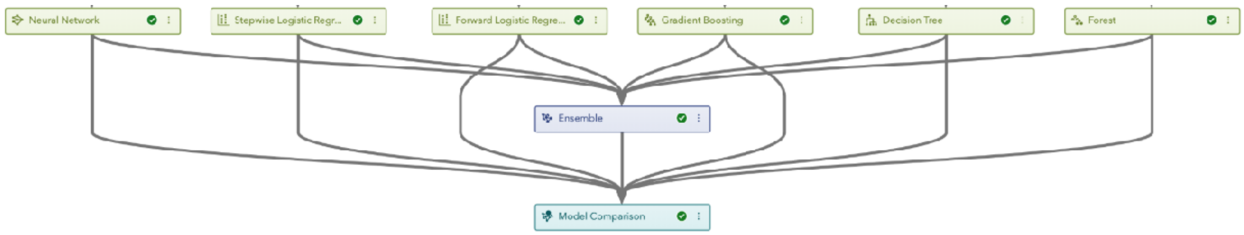


Figure 9: Model comparison node

It compares the performance of all models in the pipeline, ranks them upon various criteria and chooses the champion model as shown in Figure 10 below:

Model Comparison

Champion	Name	Algorithm ...	KS (Youden)	Misclassific...	Root Avera...	Average S...	Sum of Fre...	Multi-Class...	Gini Coeffi...
🔖	Gradient Boosting	Gradient Boosting	0.6049	0.0572	0.2290	0.0524	16,967	0.2106	0.6506
	Forward Logistic Regression	Logistic Regression	0.5909	0.0691	0.2460	0.0605	16,967	0.2366	0.6409
	Ensemble	Ensemble	0.5904	0.0664	0.2443	0.0597	16,967	0.2315	0.6435
	Stepwise Logistic Regression	Logistic Regression	0.5899	0.0693	0.2464	0.0607	16,967	0.2370	0.6369
	Neural Network	Neural Network	0.5882	0.0704	0.2472	0.0611	16,967	0.2378	0.6396
	Forest	Forest	0.5318	0.0951	0.2756	0.0759	16,967	0.2756	0.6096
	Decision Tree	Decision Tree	0.4887	0.0944	0.2924	0.0855	16,967	0.6078	0.5286

Figure 10: Model Comparison showing the champion model chosen

Then, Model studio automatically chooses the champion of the champion models across all pipelines in your project as shown in Figure 11 below:

☐	Champion	Registered	Challenger	Name	Algorithm Name	Pipeline Name	KS (Youden)	Sum of Frequencies
☑	🔖	☑		Gradient Boosting	Gradient Boosting	Advanced template fo...	0.605	16967
☐				Forward Logistic Regr...	Logistic Regression	Intermediate Templat...	0.591	16967
☐			📌	Gradient Boosting	Gradient Boosting	Tree Based	0.586	16967
☐				Neural Network	Neural Network	Neural Network	0.576	16967
☐				SVM	SVM	Support Vector Machine	0.572	16967
☐				Logistic Regression	Logistic Regression	Starter Template	0.553	16967

Figure 11: Champion models across all pipelines in the project

Collaboration in one place

The **Exchange** feature enables the user to quickly and easily create and share node or pipeline templates with other users as shown in Figure 12 below.

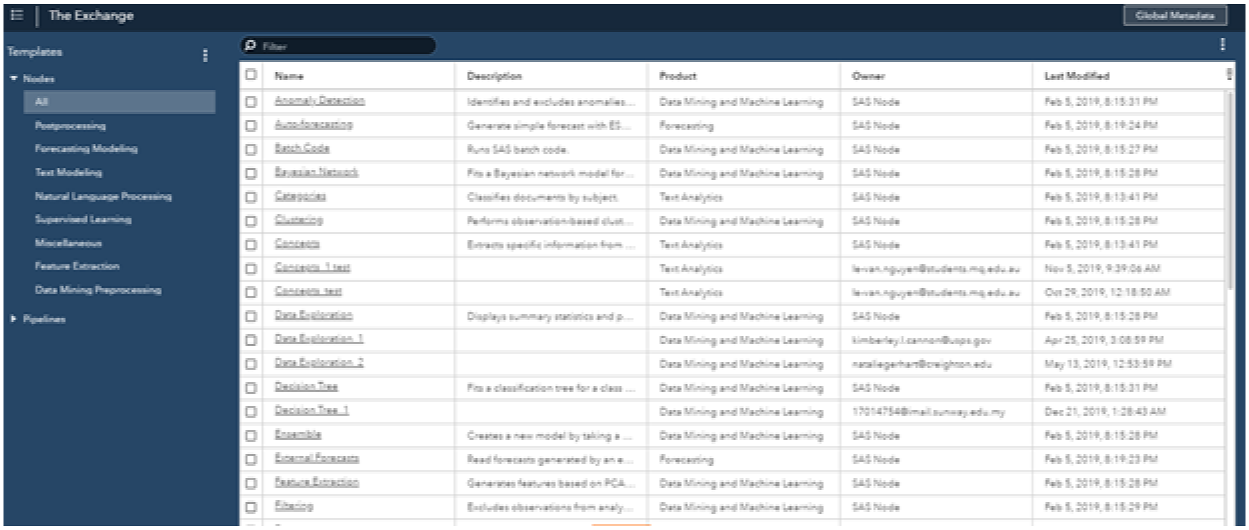


Figure 12: SAS Viya Studio Exchange

For additional SAS Viya features, refer to SAS Viya site reference [4].

SUPPORTED SAS VIYA ALGORITHMS

From a data mining and machine learning perspective, SAS Visual Data Mining and Machine Learning on SAS Viya enables end-to-end analytics - data wrangling, model building, and model assessment.

Table 1 below shows some of the analytic methods available in SAS visual data mining and machine learning on SAS Viya.

Data Wrangling	**Modeling**
Binning	Logistic Regression
Cardinality	Linear Regression
Imputation	Generalized Linear Models
Transformations	Nonlinear Regression
Transpose	Ordinary Least Squares Regression
SQL	Partial Least Squares Regression
Sampling	Quantile Regression
Variable Selection	Decision Trees
Principal Components Analysis (PCA)	Forest
K-Means Clustering	Gradient Boosting
Moving Window PCA	Neural Network
Robust PCA	Support Vector Machines
	Factorization Machines
	Network / Community Detection
	Text Mining
	Support Vector Data Description

Table 1: SAS Viya analytic methods

Summary of SAS Viya Procedures (VDMML / Visual Statistics)

A few of the SAS Viya procedures that are based on the VDMML version 8.1 and visual statistics version 8.1 is shown in Figure 13 below (Reference [5]).

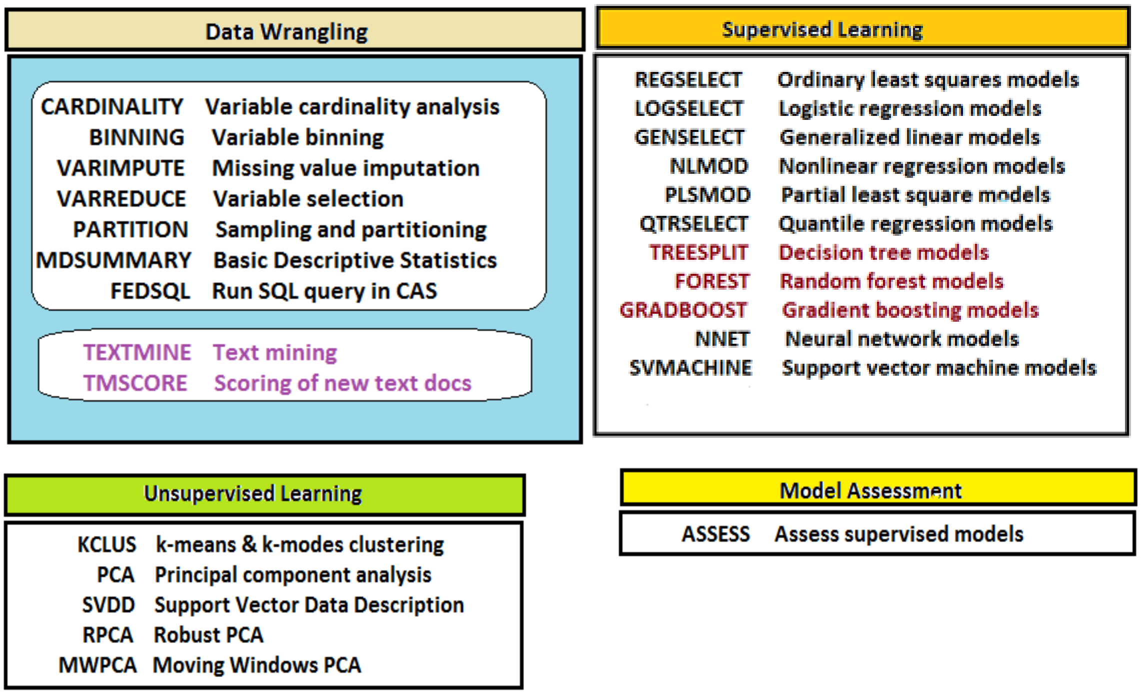


Figure 13: Summary of SAS Viya procedures

For further details on SAS procedures in Figure 13 refer to Exploring SAS® Viya®: Data Mining and Machine Learning (reference [6]).

Table 2 below shows some of the procedures available in both SAS version 9.4 and SAS Viya platforms corresponding to some machine learning analysis.

Analysis	SAS ver 9	SAS VIYA
Regression Modeling	proc reg proc glmselect	proc regselect
Unsupervised Segmentation	proc stdize proc fastclus	proc kclus
Logistic Regression Modeling	proc logistic	proc logselect
Principal Component Analysis	proc princomp	proc pca
Unsupervised Variable Reduction	proc varclus	proc varreduce
Generalized Linear Models	proc genmod	proc genselect
Autotuning : Decision tree		proc treesplit
Autotuning : Neural networks		proc nnet
Ensemble Methods: Gradient Boosting		proc gradboost
Factorization Machines		proc cardinality proc factmac
Community Detection		proc network

Table 2 : Comparison of some SAS 9.4 and SAS Viya machine learning procedures.

Table 3 below shows comparison between decision trees, boosting trees and random forest models (References [5,7])

PROC TREESPLIT	PROC FOREST	PROC GRADBOOST
Easy to Interpret	Forests are harder to interpret then a single tree.	Hard to interpret
ODS Graphics	Bagged ensemble	Overfitting controlled by LASSO and ridge
Overfitting controlled by pruning	High predictive power	Boosted ensemble
		High predictive power

Table 3: Decision Trees, Boosting Trees and Random Forest SAS Viya Comparison

TREESPLIT, FOREST and GRADBOOST procedures have similar code and ODS output tables. The fit statistics and model information table differ among procedures as each procedure has unique properties and information. They have many of the same options and statement. This helps in easy transition from one procedure to another with minimal code changes.

```
proc forest data=mycas.x;
 input xlen xwid ylen ywid;
 target z;
run;
```

```
proc gradboost data=mycas.x;
 input xlen xwid ylen ywid;
 target z;
run;
```

```
proc treesplit data=mycas.x;
 input xlen xwid ylen ywid;
 target z;
 prune none;
run;
```

Figure 14: SAS sample code for FOREST, GRADBOOST and TREESPLIT procedure

PROC TREESPLIT provides useful plots i.e. full and zoomed tree plots and Pruning or cross validation plot.

Table 4 below shows unique procedure options for TREESPLIT, FOREST and GRADBOOST.

PROC TREESPLIT	PROC FOREST	PROC GRADBOOST
CVCC	Inbagfraction	Lasso
Nsurrogates	Loh	Learningrate
Plots	Rbaimp	Ridge
Pruningtable		Samplingrate
Splitonce		

Table 4: Unique procedure options

Table 5 below shows unique procedure statements for TREESPLIT, FOREST and GRADBOOST.

PROC TREESPLIT	PROC FOREST	PROC GRADBOOST
PRUNE		

Table 5: Unique procedure statements

Things that should be kept in mind while implementing AI in healthcare

- ✓ Factor in extra time and cost for early adoption
- ✓ Reduce cost and complexity by leveraging open-source technologies and limiting customization.
- ✓ Build solutions for average transaction length and volumes but with capacity for longer transactions and peak volumes.
- ✓ Involve personnel with a combination of technology and health care backgrounds who have a clearer understanding of end users needs and preferences, as well as options for technology solutions.
- ✓ Carefully select the data used to "train" any AI/ML model: Make sure it accurately represents the production data and does not incorrectly train and bias the model.
- ✓ Since training of models is an ongoing process, expected return on investment (ROI) should include the time period and time frame.

CONCLUSION

SAS Visual Data Mining and Machine Learning enables you to build complex analytics pipelines to determine the best model for your data. Some of the advantages are as follows:

- ✓ **Beautiful and user-friendly interface**
- ✓ **Complex analytical problems solved much faster**
  Since the software runs on the latest edition of the SAS Platform – SAS Viya – it has the ability to deliver predictive modeling and machine-learning capabilities at unprecedented speeds via powerful in-memory processing.
- ✓ **Greater simplicity and improved usability**
  Even users who lack coding expertise and knowledge can leverage the platform by generating advanced machine learning algorithms via the platform's built-in visual drag-and-drop interface without ever having to know or drop a line of code into the system.
  Additionally, business users of the platform would not have to exclusively know how to code in SAS – other languages, such as Python, R, Java, and Lua can be used to code as well. The SAS code is automatically generated behind the scenes!
- ✓ **Leverage machine learning to make the right decision after exploring multiple options**
  Scenario-based decision-making is facilitated via the system's "automated model tuning", which lets users identify the best-performing model.
  The machine learning programs can integrate both structured and unstructured data, enabling users to derive more insights from new data types by adjusting their models accordingly.

As the AI market continues to evolve and new best practices are established, there are challenges and unique considerations for the successful technology adoption.

Solution providers must consider how patient privacy and security will be protected and how to take advantage of unstructured data, deal with limited access to high-quality and unbiased data sets, utilize high-performing and reliable network capabilities, verify performance, Implement data governance strategies, develop and adopt new staffing and training strategies and find a balance between costs and potential benefits

Health is not something that happens a couple of times a year when you visit doctors. Health is something that happens every day. It is happening right now. It is cumulative of choices, lifestyles we make every minute.

REFERENCES

[1] https://www.cbsnews.com/news/60-minutes-ai-facial-and-emotional-recognition-how-one-man-is-advancing-artificial-intelligence/

[2] https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3459-2019.pdf

[3] https://www.brainyquote.com/quotes/niklaus_wirth_306078

[4] SAS Viya features and training : https://support.sas.com/en/software/sas-viya.html

[5] http://dasug.dartmouth.edu/wp-content/uploads/DASUG-MachineLearning.pdf

[6] https://support.sas.com/content/dam/SAS/support/en/books/free-books/exploring-sas-viya-data-mining-machine-learning.pdf

[7] https://support.sas.com/content/dam/SAS/support/en/books/free-books/machine-learning-with-sas-special-collection.pdf

ACKNOWLEDGEMENTS

Sincere thanks to Steve Benjamin, Director Statistical Programming, Biostatistics and Global Contracts for his vision, great leadership, persistent support, and encouragement throughout and for his valuable assistance in reviewing this paper.

Sincere thanks to Gene Vinson, Senior Director, Global Data Technologies, Syneos Health for his wonderful insights, guidance and great support in the Artificial Intelligence and Machine Learning area.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Aman Bahl
Associate Director, Statistical Programming, Clinical Division, Syneos Health
Phone: +1-905-399-6715
E-mail: Aman.Bahl@syneoshealth.com
Web: http://www.syneoshealth.com

Hrideep Antony
Principal Statistical Programmer, Clinical Division, Syneos Health
Work Phone: +1-984 459 4785
Email: hrideep.antony@syneoshealth.com
Web: http://www.syneoshealth.com