What does SAS® Visual Data Mining and Machine Learning do?
It provides a comprehensive, visual interface for accomplishing all steps related to the analytical life cycle. In addition to innovative machine-learning techniques for analyzing structured and unstructured data, it integrates all other tasks in your analytical processes. From data preparation and exploration to model development and deployment, everyone works in the same, integrated environment. Scalable and elastic processing provide flexibility and speed for faster answers to complex questions.

Why is SAS® Visual Data Mining and Machine Learning important?
SAS Visual Data Mining and Machine Learning is the first solution that combines the very best advanced analytics, data prep, visualization, model assessment and model deployment in a single environment. It also supports programming from popular open source languages. This consistent, collaborative environment produces repeatable results, helping improve organizational processes and uncover new opportunities for growth.

For whom is SAS® Visual Data Mining and Machine Learning designed?
It is designed for anyone who needs to analyze large, complex data and build predictive models. This includes data scientists, statisticians, data miners, business analysts, citizen data scientists, data engineers and researchers.

Benefits
- Boost the productivity of your analytical teams. With support for the entire machine-learning pipeline, this solution enables a variety of users to build and expand upon sophisticated models to get highly accurate results – all in a single, collaborative environment.
- Reduce latency between data and decisions. Interactive visual and programming interfaces dramatically shorten the time it takes to prepare data and build models. High-speed processing delivers rapid results.
- Explore multiple approaches to find optimal solutions with confidence. Superior performance from distributed processing and the feature-rich building blocks for machine-learning pipelines let numerous users quickly explore and compare multiple approaches. Automated tuning tests different scenarios to find the best-performing model. Reproducibility in every stage of the analytical life cycle delivers answers and insights you everyone can trust.
- Solve complex analytical problems faster. This solution runs on SAS® Viya®, the latest addition to the SAS Platform, delivering predictive modeling and machine-learning capabilities at breakthrough speeds. In-memory data persistence eliminates the need to load data multiple times during iterative analyses. Analytical model processing time is measured in seconds or minutes rather than hours so you can find solutions to difficult problems faster than ever.
- Quickly deploy your predictive models with automatically generated SAS score code. Shorten the time to value even more with easy-to-implement score code that is automatically generated in multiple programming languages for all your machine-learning models.
- Empower users with language options. Python, R, Java and Lua programmers can experience the power of this solution without having to learn how to program in SAS. Give them access to trusted and tested SAS machine-learning algorithms they can use from other languages.
Overview

SAS Visual Data Mining and Machine Learning offers an exciting, new end-to-end visual environment that covers all aspects of machine learning and deep learning – from data access and data wrangling to sophisticated model building and deployment. In-memory, distributed processing handles large data and complex modeling, providing faster answers and efficient use of resources.

Flexible and approachable visual environment for analytics

Multiple users can currently analyze any amount of structured and unstructured data with the easy-to-use visual interface. Each project (goal) is defined by visual pipelines that break the analytics life cycle into a series of steps presented in a logical sequence. Branching can execute asynchronously. The visual interface (Model Studio) provides an integrated environment for the most common machine-learning steps: data prep, feature engineering, exploration, model building and deployment. Interactive tasks provide an easy way to apply sophisticated algorithms to large and complex data. These interactions also generate SAS code that can be save for later automation of tasks. In addition, code snippets and best practice templates are easily shared. Model Studio provides a highly collaborative environment for building, expanding and sharing models.

Highly scalable, in-memory analytical processing

This solution provides a secure, multiuser environment for concurrent access to data in memory. Data and analytical workloads operations are distributed across nodes, in parallel, and are multithreaded on each node for very fast speed. All data, tables and objects are held in memory as long as required, allowing for efficient processing. With built-in fault tolerance and memory management, advanced workflows can be applied to data, ensuring that processes always finish.

You get dramatically reduced runtimes for large data and analytical processing, reduced network traffic and can take full advantage of modern, multicore architectures to find solutions much faster.

Key Features

Interactive programming in a web-based development environment
- Visual interface for the entire analytical life cycle process.
- Drag-and-drop interactive interface requires no coding, though coding is an option.
- Supports automated code creation at each node in the pipeline.
- Best practice templates (basic, intermediate or advanced) to get users started quickly with machine-learning tasks.
- Collaborative environment for easy sharing of data, code snippets and best practices between different personas.

Highly scalable, distributed in-memory analytical processing
- Distributed, in-memory processing of complex analytical calculations on large data sets provides low-latency answers.
- Analytical tasks are chained together as a single, in-memory job without having to reload the data or write out intermediate results to disks.
- Concurrent access to the same data in memory by many users improves efficiency.
- Data and intermediate results are held in memory as long as required, reducing latency.
- Built-in workload management ensures efficient use of compute resources.
- Built-in failover management guarantees submitted jobs always finish.

The visual pipeline approach provides a collaborative, efficient environment for creating and deploying sophisticated machine-learning and deep learning models.

SAS Visual Data Mining and Machine Learning gives users the ability to rapidly develop and easily interpret advanced machine-learning algorithms.
Innovative statistical, data mining and machine-learning techniques

SAS Visual Data Mining and Machine Learning delivers an incredibly broad set of modern statistical, machine learning, deep learning and text analytics algorithms within a single environment.

Analytical capabilities include clustering, different flavors of regression, random forests, gradient boosting models, support vector machines, natural language processing, topic detection and more. These powerful methods drive the identification of new patterns, trends and relationships between data attributes in structured and unstructured data. The solution also provides matrix factorization for building customized recommendation systems.

With its ability to process high velocity and high-volume data sets, SAS Visual Data Mining and Machine Learning is uniquely suited for deep learning techniques. Deep learning algorithms include deep neural networks, convolution neural networks for image classification and recurrent neural networks for improved text analysis.

Complex learning algorithms, such as neural networks, gradient boosting and random forest, can be automatically tuned for optimal performance, saving time and resources.

Integrated data preparation, exploration and feature engineering

To overcome usually time-consuming analytical data preparation activities, the drag-and-drop interface enables data engineers to quickly build and run transformations, augment data and join data within the integrated visual pipeline of activities. All actions are performed in memory to maintain a consistent data structure. Discover data issues and fix them with advanced analytical techniques. Quickly identify potential predictors, reduce the dimensions of large data sets and easily create new features from your original data.

---

**Model development with modern machine-learning algorithms**

- **Random forests:**
  - Automated ensemble of decision trees to predict a single target.
  - Automated distribution of independent training runs.
  - Supports intelligent autotuning of model parameters.
  - Automated generation of SAS code for production scoring.

- **Gradient boosting:**
  - Automated iterative search for optimal partition of the data in relation to selected label variable.
  - Automated resampling of input data several times with adjusted weights based on residuals.
  - Automated generation of weighted average for final supervised model.
  - Supports binary, nominal and interval labels.
  - Ability to customize tree training with variety of options for numbers of trees to grow, splitting criteria to apply, depth of subtrees and compute resources.
  - Automated stopping criteria based on validation data scoring to avoid overfitting.
  - Automated generation of SAS code for production scoring.

- **Neural networks:**
  - Automated intelligent tuning of parameter set to identify optimal model.
  - Supports modeling of count data.
  - Intelligent defaults for most neural network parameters.
  - Ability to customize neural networks architecture and weights.
  - Ability to use an arbitrary number of hidden layers to support deep learning.
  - Automatic standardization of input and target variables.
  - Automatic selection and use of a validation data subset.
  - Automatic out-of-bag validation for early stopping to avoid overfitting.
  - Supports intelligent autotuning of model parameters.
  - Automated generation of SAS code for production scoring.

- **Support vector machines:**
  - Models binary target labels.
  - Supports linear and polynomial kernels for model training.
  - Ability to include continuous and categorical in/out features.
  - Automated scaling of input features.
  - Ability to apply the interior-point method and the active-set method.
  - Supports data partition for model validation.
  - Supports cross-validation for penalty selection.
  - Automated generation of SAS code for production scoring.
  - Supports warm restart (update models with new transactions without full retraining).

- **Factorization machines:**
  - Supports the development of recommender systems based on sparse matrices of user IDs and item ratings.
  - Ability to apply full pairwise-interaction tensor factorization.
  - Includes additional categorical and numerical input features for more accurate models.
  - Supercharge models with timestamps, demographic data and context information.
  - Supports warm restart (update models with new transactions without full retraining).
  - Supports cross-validation for penalty selection.
  - Automated generation of SAS code for production scoring.

- **Bayesian networks:**
  - Learns different Bayesian network structures, including naive, tree-augmented naive (TAN), Bayesian network-augmented naive (BAN), parent-child Bayesian networks and Markov blanket.
  - Performs efficient variable selection through independence tests.
  - Selects the best model automatically from specified parameters.
  - Generates SAS code or an analytics store to score data.
  - Loads SAS code from multiple nodes and performs computations in parallel.

---

**Analytical data preparation**

- Distributed data management routines in a visual front end.
- Large-scale data exploration and summarization.
Key Features (continued)

- Cardinality profiling:
  - Large-scale data profiling of input data sources.
  - Intelligent recommendation for variable measurement and role.
- Sampling: Supports random and stratified sampling, oversampling for rare events and indicator variables for sampled records.

Data exploration, feature engineering and dimension reduction
- Feature binning.
- High-performance imputation of missing values in features with user-specified values, mean, pseudo median and random value of nonmissing values.
- Feature dimension reduction.
- Large-scale principal components analysis (PCA), including moving windows and robust PCA.
- Unsupervised learning with cluster analysis and mixed variable clustering.

Integrated text analytics
- Supports 30 native languages out of the box: English, Arabic, Chinese, Croatian, Czech, Danish, Dutch, Farsi, Finnish, French, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Slovak, Slovenian, Spanish, Swedish, Tagalog, Turkish, Thai and Vietnamese.
- Automatically identifies term part of speech (more than 15 are system-defined).
- Extracts standard entities such as location, time, date and address from predefined options.
- Detects noun groups and multiterm lists, and treats as single terms in machine-learning processing.
- Detects different term stems without manual intervention.
- Finds term variants automatically with synonym detection.
- Ability to select frequency-term weighting to dampen term-occurrence effect in a document.
- Term weights can be used to distinguish more important terms in a document collection.
- Uses default start and stop lists to manage terms for parsing and downstream processing.
- Add, delete and edit terms, including multiword terms, to refine start and stop lists.
- Machine-learned topics represent the term-by-document, matrix-generated text processing as a structured numeric representation of the document collection.
- Resultant semantically related generated topics can be used as input into machine-learning models.
- Automatically generates code for production document scoring in SAS, including text preprocessing and parsing for accurate results.

Model assessment
- Automatically calculates supervised learning model performance statistics.
- Produces output statistics for interval and categorical targets.
- Creates lift table for interval and categorical target.
- Creates ROC table for categorical target.

Model scoring
- Automatically generates SAS DATA step code for model scoring.
- Applies scoring logic to training, holdout data and new data.

Integrated text analytics
Designed with big data in mind, you can examine extremely large collections of text documents. Explore all of your textual data, not just a subset, to gain new insights about unknown themes and connections. Combining structured data with text data uncovers previously undetected relationships and adds even more predictive power to analytical models.

Model assessment and scoring
Test different modeling approaches in a single run and compare results of multiple supervised learning algorithms with standardized tests to quickly identify champion models. Then, operationalize analytics in distributed and traditional environments with automatically generated SAS score code.

Accessible and cloud-ready
Whether it's Python, R, Java or Lua, modelers and data scientists can access SAS capabilities from their preferred coding environment. And with SAS Viya REST APIs, you can add the power of SAS to other applications.

You can also deploy SAS Visual Data Mining and Machine Learning where it makes the most sense for your organization: on-site, in a private cloud via technologies such as Cloud Foundry or in public clouds, including Amazon Web Services and Microsoft Azure. You can also access this software via the predeployed and preconfigured managed software-as-a-service offerings provided by SAS.

To learn more about SAS Visual Data Mining and Machine Learning, view screenshots and see other related materials, please visit sas.com/vdmml.