



ZEMENTIS for Datameer

User Guide

10.5.0.2

ZEMENTIS for Datameer

User Guide

Software AG

Copyright © 2004 - 2016 Zementis Inc.

Copyright © 2016 - 2019 Software AG, Darmstadt, Germany and/or Software AG USA Inc., Reston, VA, USA, and/or its subsidiaries and/or its affiliates and/or their licensors.

This document applies to ZEMENTIS 10.5.0.2 and to all subsequent releases.

Specifications contained herein are subject to change and these changes will be reported in subsequent release notes or new editions.

The name Software AG and all Software AG product names are either trademarks or registered trademarks of Software AG and/or Software AG USA Inc. and/or its subsidiaries and/or its affiliates and/or their licensors. Other company and product names mentioned herein may be trademarks of their respective owners.

Detailed information on trademarks and patents owned by Software AG and/or its subsidiaries is located at <http://softwareag.com/licenses/>.

Use of this software is subject to adherence to Software AG's licensing conditions and terms. These terms are part of the product documentation, located at <http://softwareag.com/licenses/> and/or in the root installation directory of the licensed product(s).

This software may include portions of third-party products. For third-party copyright notices, license terms, additional rights or restrictions, please refer to "License Texts, Copyright Notices and Disclaimers of Third Party Products". For certain specific third-party license restrictions, please refer to section E of the Legal Notices available under "License Terms and Conditions for Use of Software AG Products / Copyright and Trademark Notices of Software AG Products". These documents are part of the product documentation, located at <http://softwareag.com/licenses/> and/or in the root installation directory of the licensed product(s).

Table of Contents

1. Introduction	1
2. Overview	2
2.1. Predictive Model Markup Language (PMML)	2
2.2. ZEMENTIS Predictive Analytics (ZEMENTIS)	2
3. Installation	5
3.1. Requirements	5
3.2. Packaging	5
3.3. New Installation	5
3.4. Upgrade Installation	6
4. From PMML to Datameer®	7
4.1. Configure ZEMENTIS for Datameer®	7
5. Scoring Models in Datameer®	14
5.1. Scoring Models with a single output	14
5.2. Scoring Models with Multiple Outputs	16

List of Figures

2.1. Overview of the ZEMENTIS for Datameer®	3
4.1. Configure ZEMENTIS for Datameer®	7
4.2. Uploading the Tax Audit SVM Model to Datameer®	8
4.3. Viewing Tax Audit SVM Model Details	9
4.4. Configuration error in Datameer®	10
4.5. Uploading the Iris NN Model to Datameer®	11
4.6. Error during model deployment	12
4.7. Inspecting errors and warnings	13
5.1. Datameer® Formula Builder	14
5.2. Scoring the Tax Audit SVM Model in ZEMENTIS for Datameer®	15
5.3. Tax Audit SVM Model inputs	15
5.4. Tax Audit SVM Model output	16
5.5. Scoring the Iris NN Model in ZEMENTIS for Datameer®	17
5.6. Multiple outputs with JSON_OUTPUT	18
5.7. Selecting a specific output with JSON_VALUE	18

List of Tables

3.1. Directory Structure of the ZEMENTIS for Datameer [®] package.	5
--	---

Chapter 1. Introduction

As advanced analytics becomes pervasive across the enterprise to drive better business decisions, the need for efficient execution of predictive models is paramount. An ever growing array of data mining tools and, all too often, custom specialized software is used to mine data and derive statistical models from a wealth of collected data. The ultimate goal is to turn these models into business value by incorporating them into day to day business operations. This necessitates the ability to integrate them into the IT infrastructure where the outcomes can easily flow into the finger-tips of the decision makers. At the same time, the accelerating growth rate of data collected implies that only the most scalable database architectures will be able to meet storage, and more importantly, processing requirements.

In the era of big data, more and more organizations are turning to the scalable architecture of Hadoop to meeting this growing challenge. To bring the power of predictive models into this architecture, [Software AG](#) and [Datameer®](#) have joined forces to deliver the ZEMENTIS Predictive Analytics (ZEMENTIS) for [Datameer®](#). ZEMENTIS offers [Datameer®](#) users the best combination of open standards and scalability for the application of predictive analytics. With the Predictive Modeling Mark-up Language ([PMML](#)) as the bridge between the model development environment and the IT data warehousing infrastructure, ZEMENTIS for [Datameer®](#) offers standards-based deployment of predictive models and execution on a highly scalable analytics platform. This joint solution combines the power of ZEMENTIS Predictive Analytics server, the flagship product of Software AG, with that of [Datameer®](#), the only analytics application natively built on Hadoop. As a result, a wide range of predictive models, possibly developed with different tools in different environments, can fully leverage the scalability and flexibility of Hadoop to bring game changing data integration and analytics to end users. Practically, [PMML](#) becomes yet another analytical function available in [Datameer®](#) offering execution performance that can meet the volume and performance requirements of the most demanding environments.

This document serves as a guide for installing and using ZEMENTIS for [Datameer®](#). It first gives a brief overview of the plug-in. It then presents the simple installation process. Finally, it illustrates the use of ZEMENTIS with two [PMML](#) models showing how to upload, deploy and score them using ZEMENTIS for [Datameer®](#).

Note

With respect to the EU General Data Protection Regulation (GDPR), our product does not collect or store any personally identifiable information. However, as the input data might contain sensitive personal information, please anonymize any such data to ensure that the processing of personal data is in accordance with the GDPR.

Chapter 2. Overview

2.1. Predictive Model Markup Language (PMML)

As the de-facto standard for data mining models, [PMML](#) provides tremendous benefits for business, IT, and the data mining industry in general. Developed by the [Data Mining Group \(DMG\)](#), an independent, vendor-led consortium, [PMML](#) increases business agility by eliminating the need for proprietary solutions or custom code development.

Today, it is supported by all major data mining tools, commercial and open source. As an open standard, it enables project stakeholders to standardize on one common representation for data mining models. It practically eliminates the barriers and gaps between development and production deployment of predictive analytics. In effect, it minimizes the complexity, cost, and time to turn predictive models into operational IT and business assets.

As the lingua franca for predictive analytics, data mining models can be easily exchanged between [PMML](#)-compliant applications. In this way, a model may be built in one statistical tool and easily moved to another for production deployment or visualization. [PMML](#) also serves as a bridge between all the teams involved in the data mining process inside a company since it can be used to disseminate knowledge and best practices, thereby stimulating cross-team and inter-organizational collaboration. In a world in which data-driven decisions are becoming more and more pervasive, predictive analytics and standards such as [PMML](#) make it possible for organizations to benefit from smart solutions that will truly revolutionize their business.

Besides offering a rich set of structures for describing all the intricate details of a predictive algorithm, [PMML](#) also provides information about the inputs and outputs of a model. This includes names and types of all input and output data fields, often along with the set of permissible values. In addition, a model expressed in [PMML](#) typically includes information about how to handle invalid or missing input values.

Note

A variety of sample [PMML](#) models are included with the ZEMENTIS distribution package. In addition, a wealth of resources on [PMML](#) can be found from the [PMML in Action](#).

2.2. ZEMENTIS Predictive Analytics (ZEMENTIS)

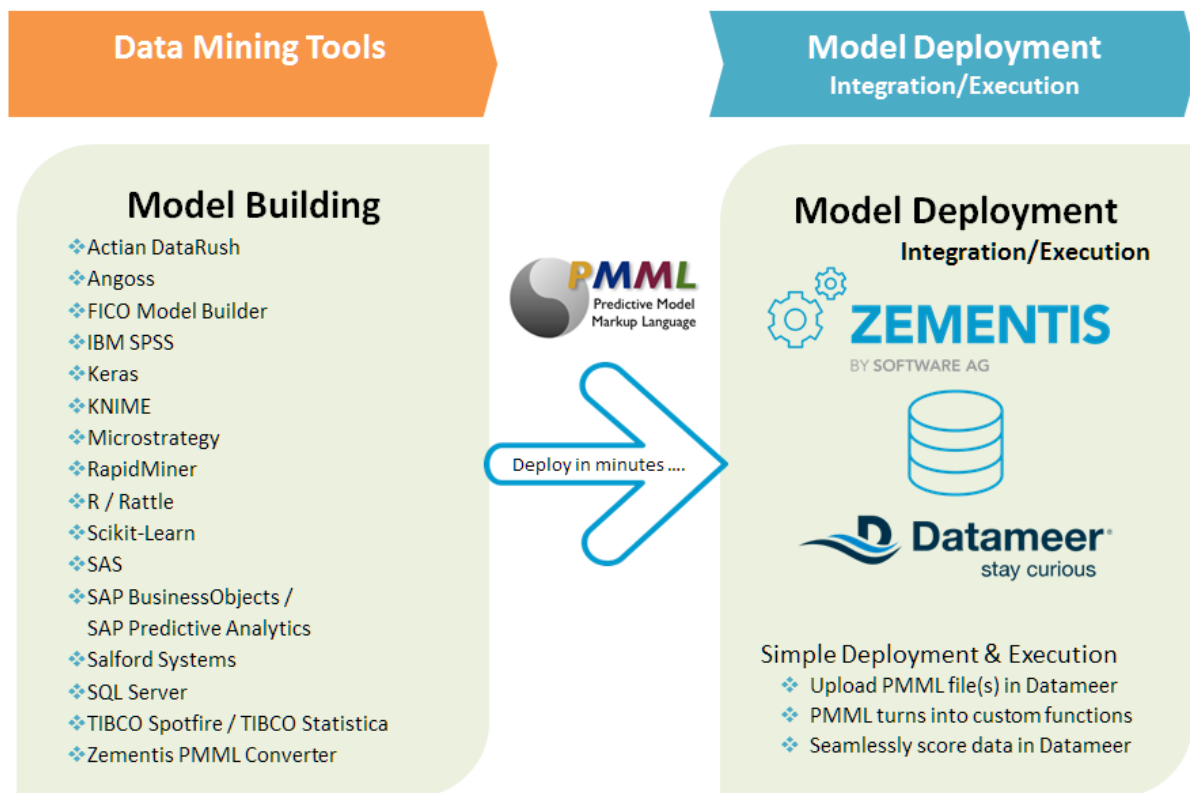
ZEMENTIS enables execution of standards-based predictive analytics directly within Datameer[®]. It shares the [PMML](#) execution core with the ZEMENTIS server offered by Software AG.

In addition, the plug-in takes on the responsibility of bridging the [PMML](#) and the Datameer[®]/Hadoop world. This means that it presents each loaded [PMML](#) model as a Datameer[®] function. The name, input arguments and outputs

of each function match the name, input fields, and output fields of the corresponding model as defined in the [PMML](#) file. This way, scoring a data set requires nothing more than invoking the analytics function correspondent to the appropriate predictive analytics model. Predictions (scores, probabilities, categories, clusters, etc.) can be just as easily become part of further analytics inspection, or plotted with the use of dashboards.

Like the ZEMENTIS server, ZEMENTIS plugin accepts [PMML](#) models of all versions (2.0, 2.1, 3.0, 3.1, 3.2, 4.0, 4.1, 4.2 and 4.3) generated by any of the major commercial and open source data mining tools.

Figure 2.1. Overview of the ZEMENTIS for Datameer®



At a high level, the process of using [PMML](#) models in Datameer® is presented in [Figure 2.1](#). It starts after the predictive models have been created and have been exported in [PMML](#) format from the data mining and model development tool. With the [PMML](#) files at hand, all it takes is to upload and deploy them into ZEMENTIS for Datameer® where they are translated into functions.

The plug-in supports a wide range of [PMML](#) elements that implement the following predictive analytic techniques:

- Anomaly Detection Model (Isolation Forest and One-Class SVM)
- Association Rules Models (Rectangular or Transactional format) (mining schema attribute `usageType = "group"` is not supported)

- Clustering Models (Distribution-Based, Center-Based, and 2-Step Clustering)
- Decision Trees (for classification and regression) together with multiple missing value handling strategies (Default Child, Last Prediction, Null Prediction, Weighted Confidence, Aggregate Nodes)
- K-Nearest Neighbors (for regression, classification and clustering)
- Naive Bayes Classifiers (with continuous or categorical inputs)
- Neural Networks (Back-Propagation, Radial-Basis Function, and Neural-Gas)
- Regression Models (Linear, Polynomial, and Logistic) and General Regression Models (General Linear, Ordinal Multinomial, Generalized Linear, Cox)
- Ruleset Models (Each rule contains a predicate and a predicted class value)
- Support Vector Machines (for regression and multi-class and binary classification)
- Scorecards (point allocation for categorical, continuous, and complex attributes as well as support for reason codes)
- Multiple models (model ensemble, segmentation, chaining, composition and cascade), including Random-Forest Models and Stochastic Boosting Models

In addition, it implements all the [PMML](#) built-in functions and transformations for data pre- and post-processing.

Note

- Note that in addition to association rules which use `usageType = "group"`, ZEMENTIS for Datameer[®] does not currently support time series, text models, and sequence models.
- ZEMENTIS for Datameer[®] does not support BlockIndicator elements, which might be used as part of Lag expressions.

Chapter 3. Installation

This section describes how to install the ZEMENTIS for Datameer[®].

3.1. Requirements

In order to install ZEMENTIS for Datameer[®] on your system, you will need an existing Datameer[®] installation. Uploading new PMML models with ZEMENTIS for Datameer[®] requires a valid Product License Key which can be obtained by contacting [Software AG](#).

3.2. Packaging

ZEMENTIS for Datameer[®] is distributed as a compressed archive file: `uppi-datameer6-10.5.0.2.zip`. The distributed package consists of several files, including this documentation and several sample files. When uncompressed, the package reveals a number of directories as described in [Table 3.1](#).

Table 3.1. Directory Structure of the ZEMENTIS for Datameer[®] package.

Directory	Contents
<code>docs</code>	Documentation in HTML and PDF format.
<code>pmm1</code>	A number of sample PMML files along with data files in CSV format. These include the examples described in this document.
<code>classes, lib</code>	Runtime code and library used by ZEMENTIS for Datameer [®] .
<code>license</code>	License information for ZEMENTIS and 3rd party libraries.

3.3. New Installation

To install ZEMENTIS for Datameer[®] navigate your browser to "Administration" tab and select "Plugins" option from left-side menu. On the bottom of the page you will find "Upload Plug-in" panel. Click on "Choose File" button, select `uppi-datameer6-10.5.0.2.zip` file, and click on "Upload" button. If installation was successful, a green message box will be displayed, and if it failed, a red message box will be displayed. Also, on the same page in the "Plug-ins" panel which displays table of all installed plugins you will find new row entry with "ZEMENTIS Plug-In" in "Name" column. Lastly, place the Product License Key file, `zementis.license`, in the Datameer[®] root directory. Please note that execution of existing models will not be interrupted when the license expires.

Note

Note that if you have installed Datameer® in a computer running Mac OS X, you will need to right-click on top of the Datameer® application package and choose "Show Package Contents" in order to have access to its directory structure.

Once the ZEMENTIS for Datameer® is installed, you can go ahead and upload and deploy your predictive models. This is described in [Chapter 4](#).

3.4. Upgrade Installation

Upgrading existing ZEMENTIS for Datameer® installation involves few steps to ensure that Datameer® worksheets with currently deployed [PMML](#) functions will remain operational with newer version of ZEMENTIS for Datameer®. First, you have to make sure that copies of all [PMML](#) files corresponding to [PMML](#) functions are backed up and available for re-upload. If this is the case, remove all [PMML](#) functions using ZEMENTIS configuration interface. Details on [PMML](#) function removal and deployment operations can be found in [Chapter 4](#). Next, remove ZEMENTIS for Datameer® by navigating to "Administration" tab, selecting "Plugins" option from left-side menu, and then clicking on trash icon in "ZEMENTIS Plug-In" entry of "Plug-Ins" panel. Install new ZEMENTIS for Datameer® by uploading `uppi-datameer6-10.5.0.2.zip` file through "Upload Plug-in" panel at the bottom of the same page. Finally, restore all [PMML](#) functions by uploading each [PMML](#) file using ZEMENTIS configuration interface.

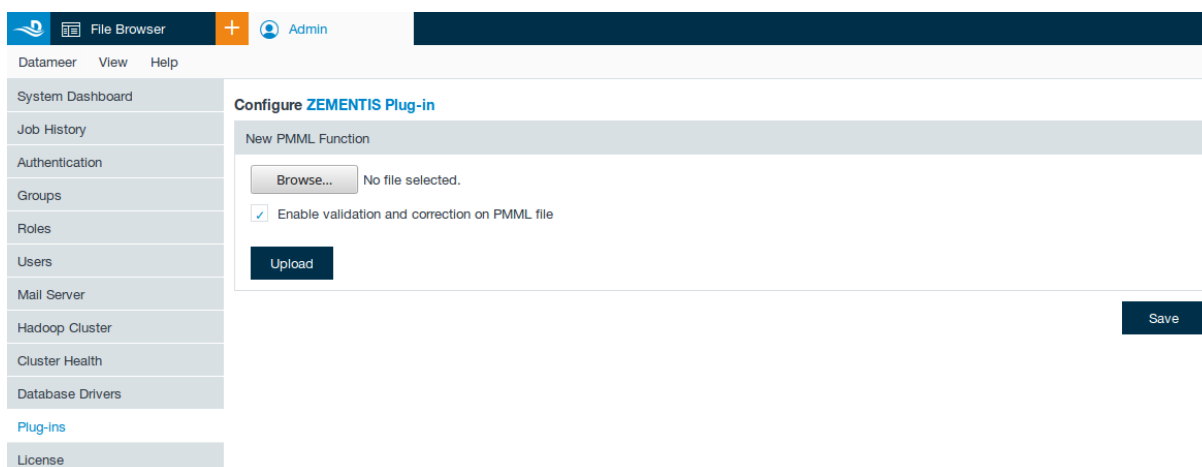
Chapter 4. From PMML to Datameer®

This section will describe in detail how PMML models are made available to be used directly in Datameer®. As explained earlier, ZEMENTIS turns predictive models into functions which can be readily used from inside Datameer® as any other function available through the "Formula Builder".

4.1. Configure ZEMENTIS for Datameer®

To make your PMML models available in Datameer®, go to the "Administration" tab and select the "Plug-ins" option from the left-side menu. In the table of available plug-ins, locate the ZEMENTIS Plug-In and click on the "Configure" hyperlink. Once you do that, you will be faced with a screen as shown in Figure 4.1. It is here that you will upload and deploy your PMML models. The "Enable validation and correction on PMML file" checkbox indicates if comprehensive semantic checks and corrections are to be performed on the provided PMML file. The checkbox is checked by default. Disabling the checkbox will improve upload time, but this is only recommended for PMML files that are generated by ZEMENTIS certified exporters.

Figure 4.1. Configure ZEMENTIS for Datameer®



In this section, we describe the process of uploading and deploying PMML files in ZEMENTIS for Datameer®. These files represent two distinct predictive models.

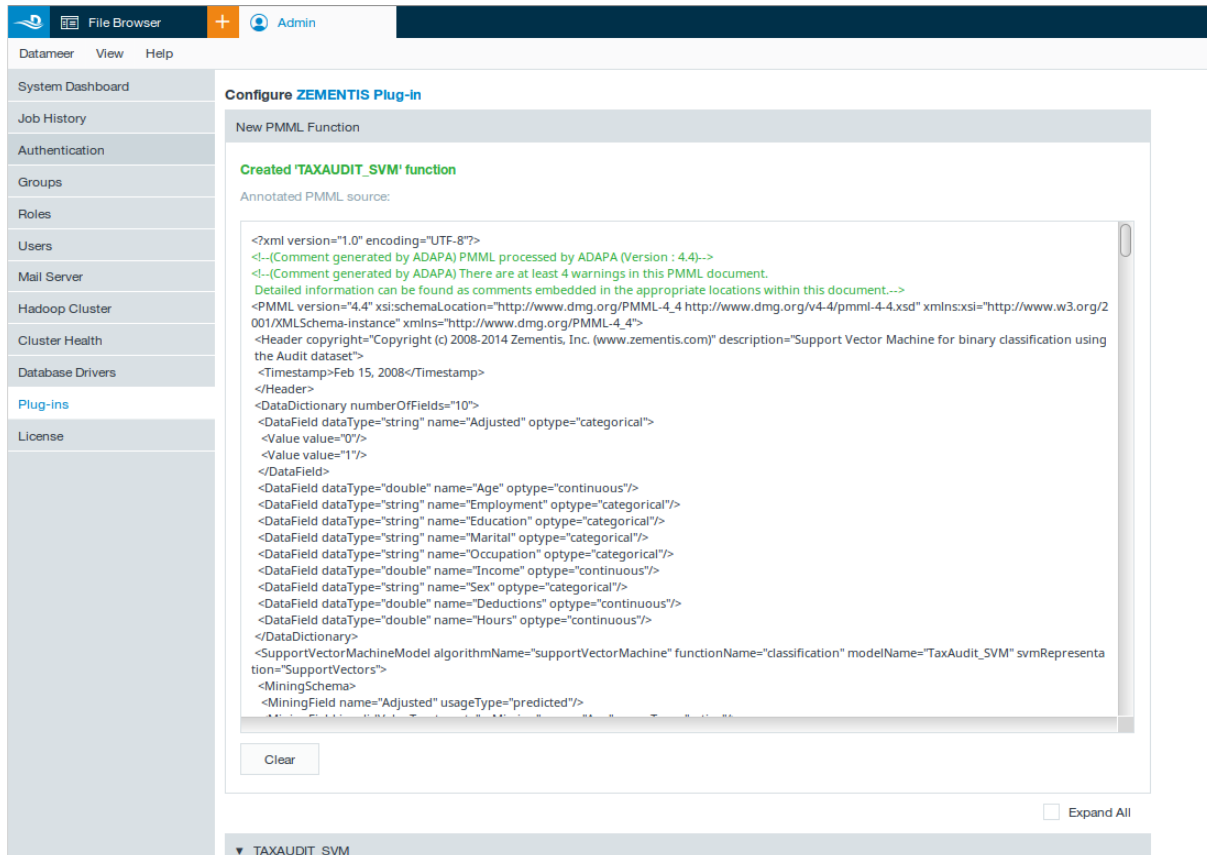
The first model is a Support Vector Machine (SVM) trained with the Audit dataset. This model will produce a single output field which will contain "0" in case the taxpayer's claim does not need to be adjusted or "1" in case it does. The audit dataset is supplied as part of the R Rattle package - <http://rattle.togaware.com> (it is also available for download as a CSV file from <http://rattle.togaware.com/audit.csv>). It consists of fictional clients who have been

audited, perhaps for tax refund compliance. For each case an outcome is recorded (whether the taxpayer's claims had to be adjusted or not).

The second model is a Neural Network (NN) model trained with the Iris dataset. This model will produce four output fields, the class or type of Iris plant (setosa, virginica, or versicolor) with the computed probabilities for the three different flower types. The Iris dataset is perhaps the best known dataset to be found in the pattern recognition literature. It contains three classes representing different types of the Iris plant. Each class is represented by 50 records. For more information on the Iris dataset, please check the Iris page at the UCI Repository of Machine Learning Databases - <http://archive.ics.uci.edu/ml/datasets/Iris> (Asuncion, A. and Newman, D.J. (2007). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science).

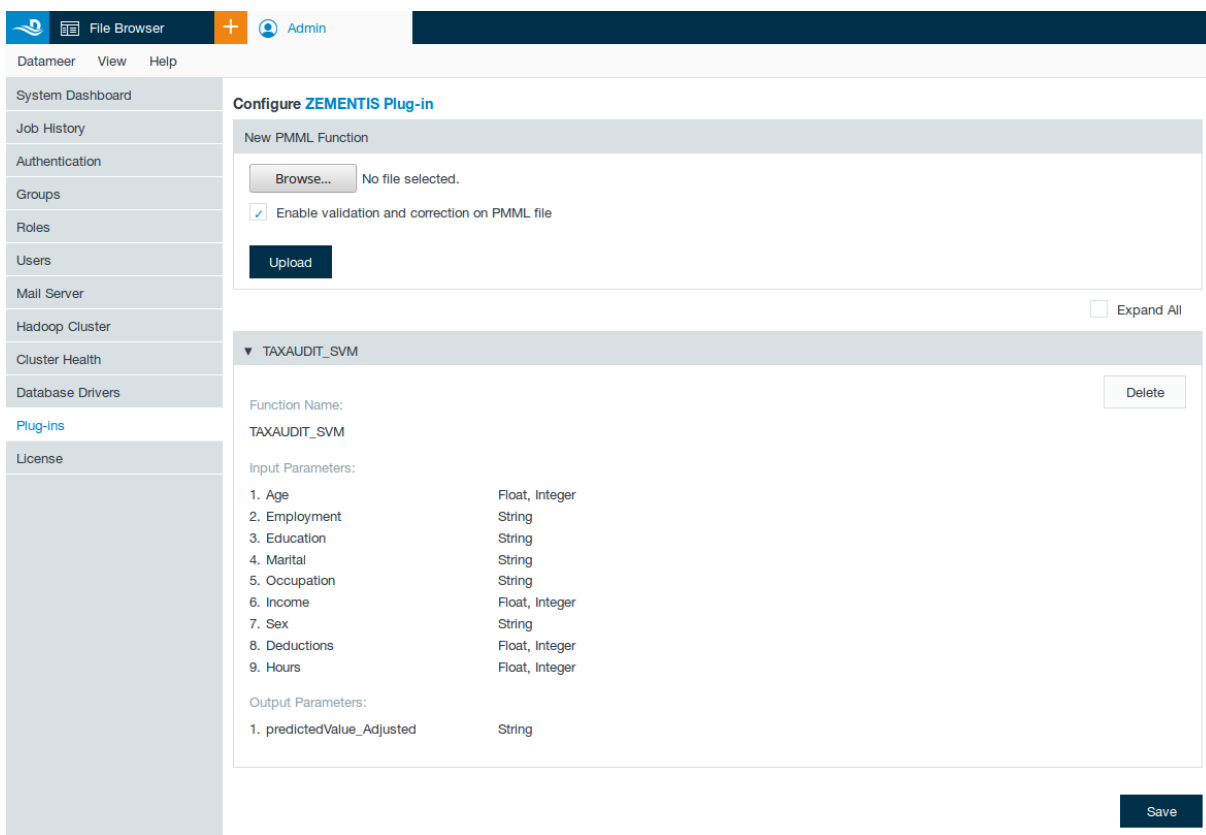
To make the Tax Audit SVM model available in Datameer[®], simply click on the "Choose File" button and locate the corresponding PMML file, and click on the "Upload" button. Once the file is uploaded and model is successfully deployed, you should see the screen as shown in [Figure 4.2](#). You can inspect PMML file for warnings highlighted in green by scrolling down text in the window under "Annotated PMML source" heading. Clicking "Clear" button causes this message panel to revert back to PMML file upload panel.

Figure 4.2. Uploading the Tax Audit SVM Model to Datameer[®]



You can view a list of all deployed PMML functions under "PMML Functions" section. It contains expandable bars, each with function name on the left and delete button on the right. You can view parameters used by each model, such as list of inputs and outputs, by clicking on expandable bar or by clicking "Expand All" checkbox. The function name corresponding to a PMML file is derived from the PMML attribute "modelName". The input parameters reflect all the input fields necessary for the model to be executed. In case of the Tax Audit SVM model, these are: "Age", "Employment", "Education", "Marital", "Occupation", "Income", "Sex", "Deductions", and "Hours". The output parameters represent fields that the model returns as output. For the Tax Audit SVM model, the output parameter is field "predictedValue_Adjusted". The details of Tax Audit SVM model are shown in [Figure 4.3](#)

Figure 4.3. Viewing Tax Audit SVM Model Details

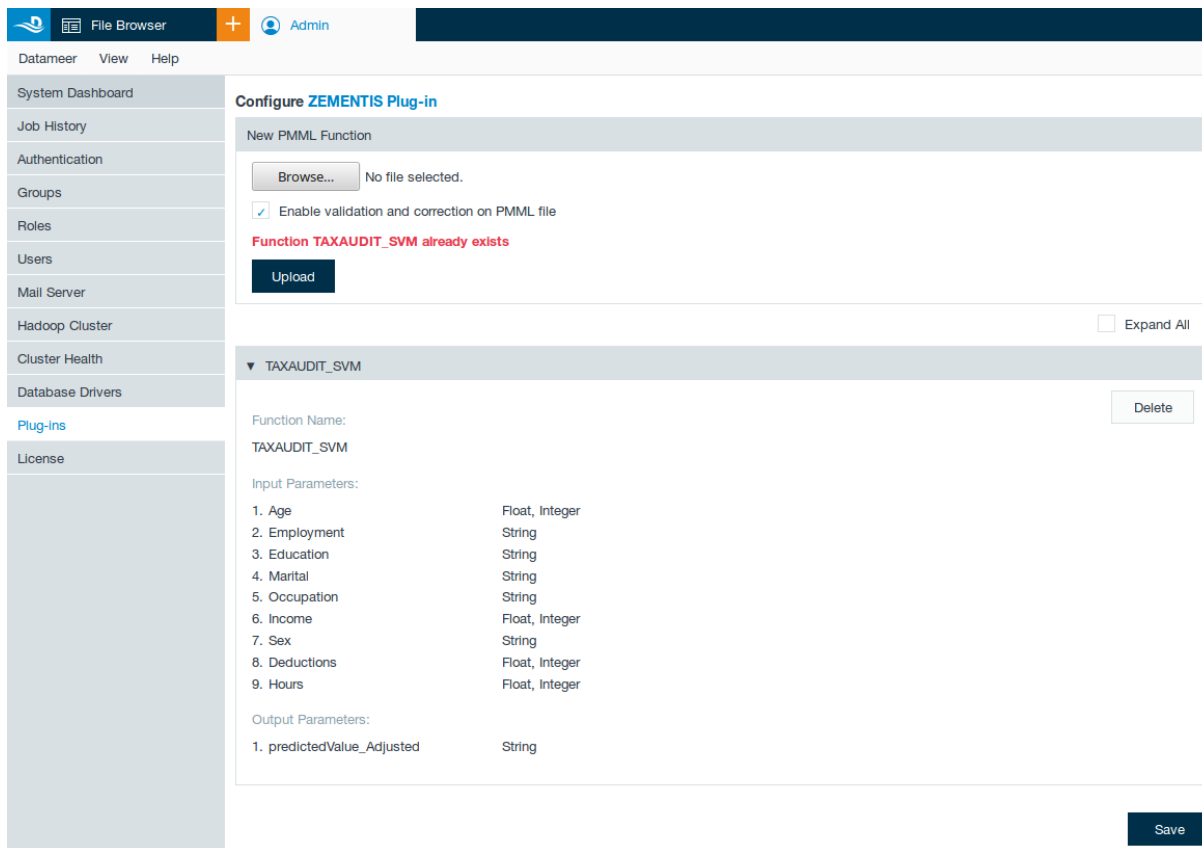


If you try to upload and deploy a model in ZEMENTIS for Datameer[®] that has already been uploaded, you will get an error notifying you that a Datameer[®] function with the same name already exists as shown in [Figure 4.4](#) . To avoid such an error, we recommend you delete the existing model before attempting to upload the new one.

Note

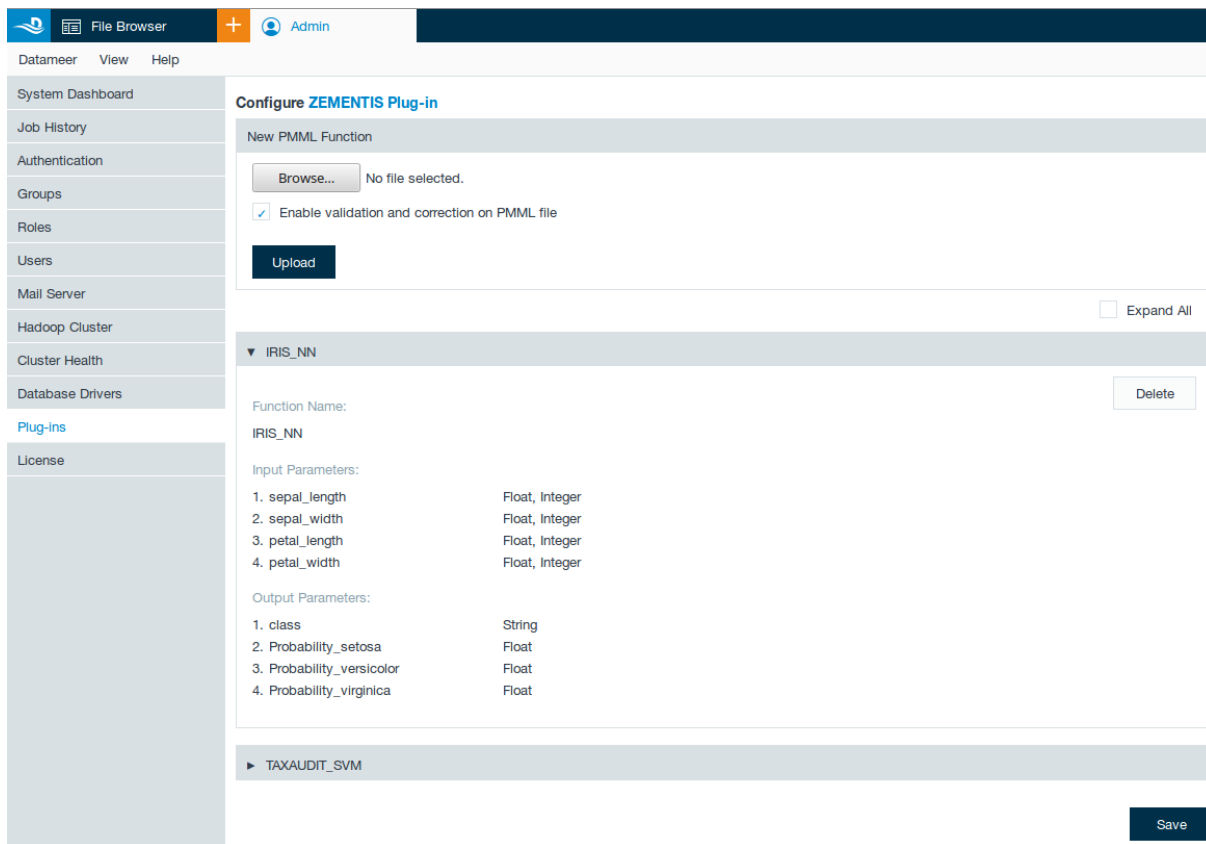
ZEMENTIS will also notify you of any errors if the PMML file you are trying to upload has any errors. In that case, we recommend you correct the issue(s) pointed out and upload the model again.

Figure 4.4. Configuration error in Datameer®



To make the Iris NN model available in Datameer®, select the corresponding PMML file and click on the "Upload" button. Once the file is uploaded and model is successfully deployed, you should see a successful upload panel with the annotated source. After clicking on "Clear" button on the upload panel, you should see the screen as shown in [Figure 4.5](#).

Figure 4.5. Uploading the Iris NN Model to Datameer®



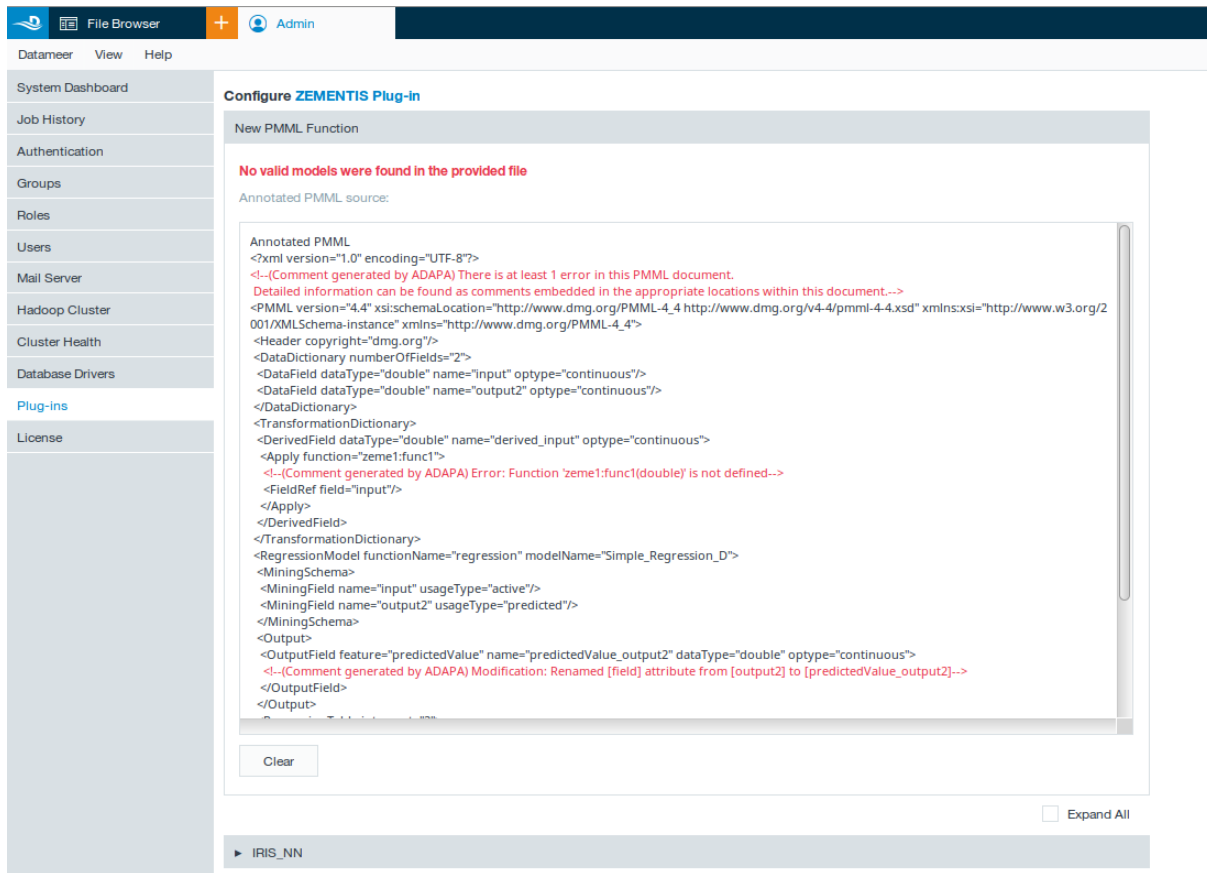
To continue uploading and deploying models in ZEMENTIS for Datameer® simply repeat the process described above. There is no limit to the number of models that can be uploaded. However, we recommend you delete models whenever they are no longer required. You can do that by clicking on "X" button located on the right side of each PMML function bar as shown in the screenshots above.

Note

Note that models are listed in alphabetical order in Datameer®, not the order in which they were uploaded. In this case, even if the Iris NN model had been uploaded first, it would still be shown after the Tax Audit SVM model.

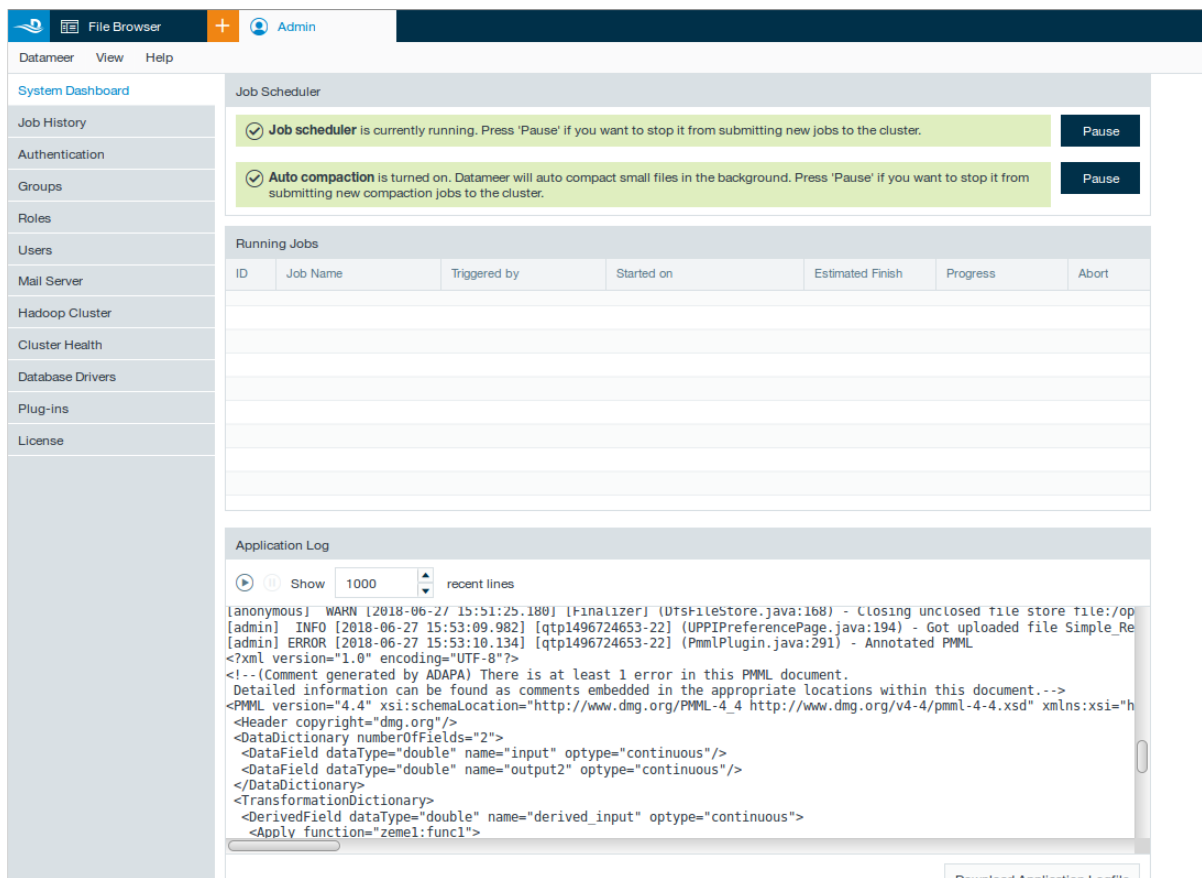
Let's now consider a case where the [PMML](#) file for the Iris NN model had an error which could not be corrected automatically by ZEMENTIS when uploading the model. In this case, the model would not be deployed and instead you would get an error stating that the [PMML](#) file was not valid, as shown in [Figure 4.6](#) .

Figure 4.6. Error during model deployment



Before a **PMML** file is deployed in ZEMENTIS for Datameer[®], it is checked for syntactic and semantic problems. During this process, it may encounter errors, which prevent a model from being deployed, and/or minor issues which are reported as warnings. Errors and warnings are embedded into the uploaded **PMML** file and can be viewed in the configuration panel after model upload is complete. **PMML** models with errors are also available for inspection in Datameer[®] as part of the application log file and can be accessed through the "System Dashboard" as shown in [Figure 4.7](#).

Figure 4.7. Inspecting errors and warnings



Note

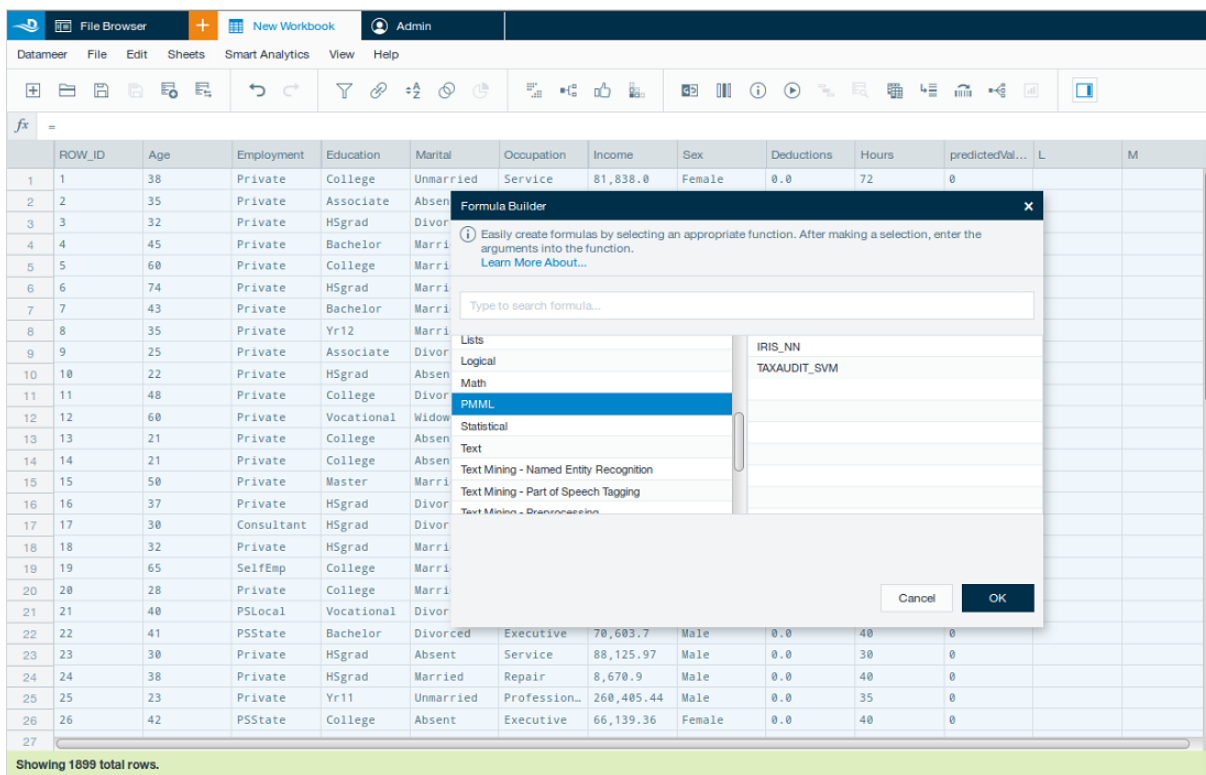
Note that even though a model may be deployed with no errors, ZEMENTIS may still find minor issues with the PMML code that are reported as warnings. Though these warnings typically do not impact the scoring process, they should be reviewed by the data science team.

Chapter 5. Scoring Models in Datameer®

This chapter describes the process of executing predictive models in Datameer®. Models can only be executed if they have already been uploaded and deployed in ZEMENTIS for Datameer®. This process is described in [Chapter 4](#).

Once a predictive model is uploaded and deployed in ZEMENTIS for Datameer®, it becomes yet another function available through the "Formula Builder". To access the formula builder wizard, you will need to create a "Workbook" and import your data into it. The formula builder will be shown whenever you click on an empty cell. Given that we have already uploaded the PMML files for models Tax Audit SVM and Iris NN, these are shown as two distinct functions under the PMML function class as shown in [Figure 5.1](#).

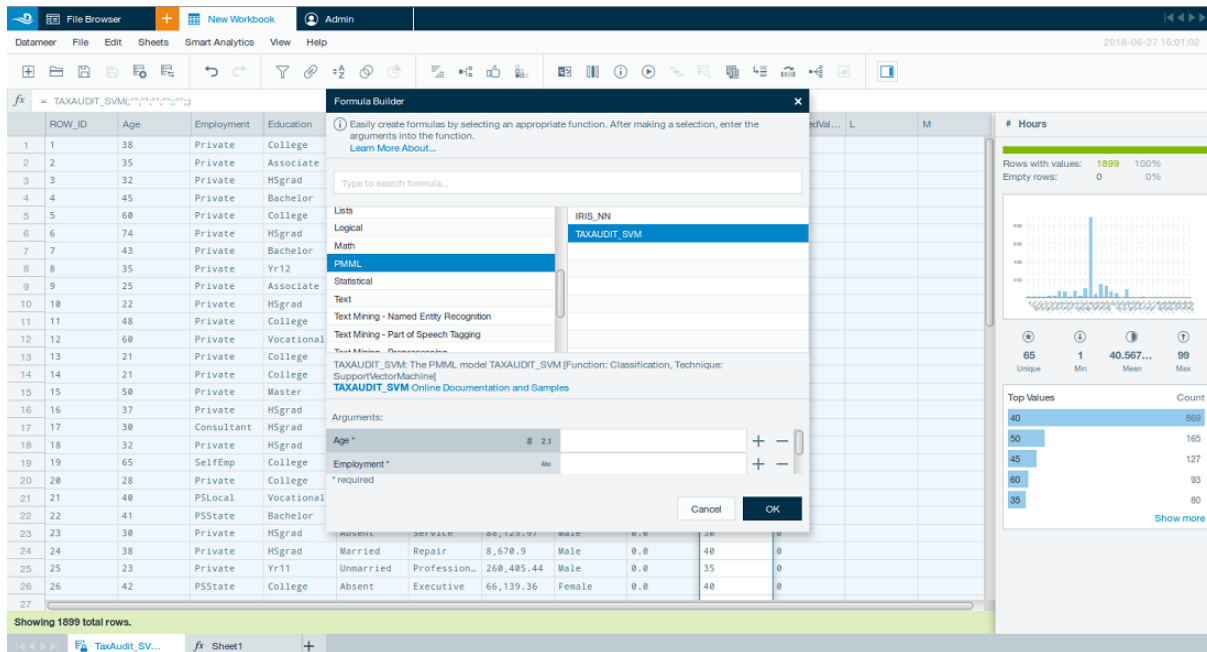
Figure 5.1. Datameer® Formula Builder



5.1. Scoring Models with a single output

To create formula for Tax Audit SVM, simply select the TaxAudit_SVM function in the wizard. Once you do that, it will expand to show all the input arguments required for scoring the Tax Audit SVM model as depicted in [Figure 5.2](#).

Figure 5.2. Scoring the Tax Audit SVM Model in ZEMENTIS for Datameer®



To select the appropriate data columns for the input arguments in the wizard, simply click on any data cell for the column you would like to select. Datameer® automatically assumes the next selection you make is for the next argument in the formula wizard. After all entries have been mapped accordingly, as shown in Figure 5.3, simply click the "OK" button.

Figure 5.3. Tax Audit SVM Model inputs

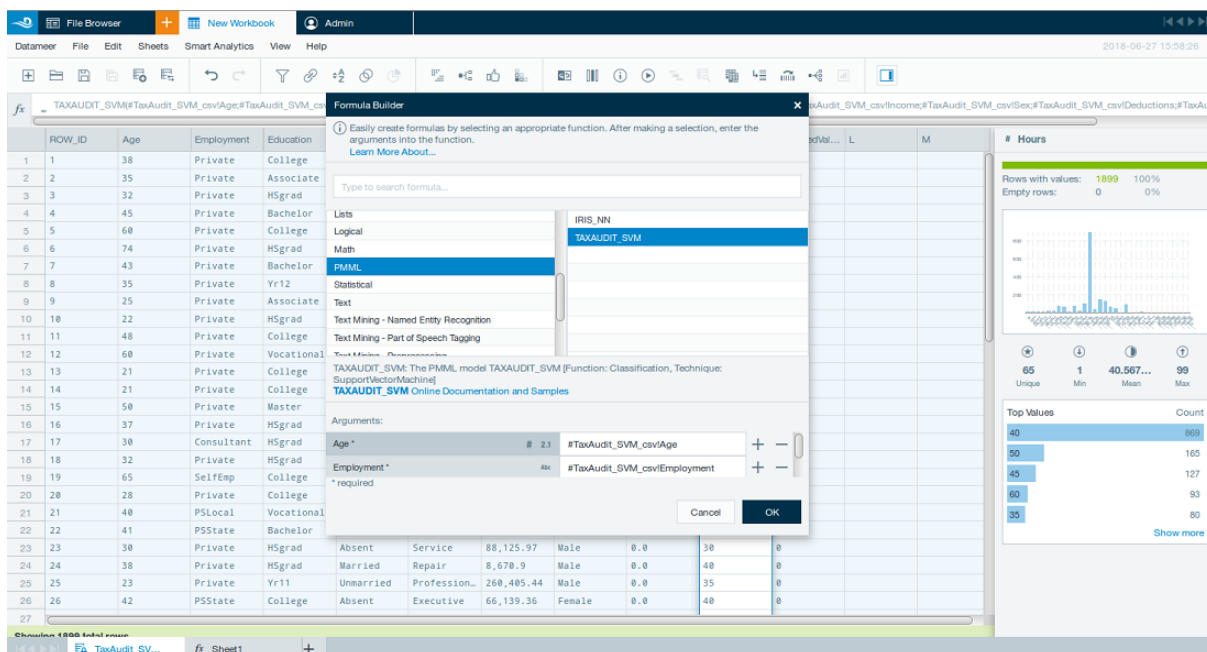
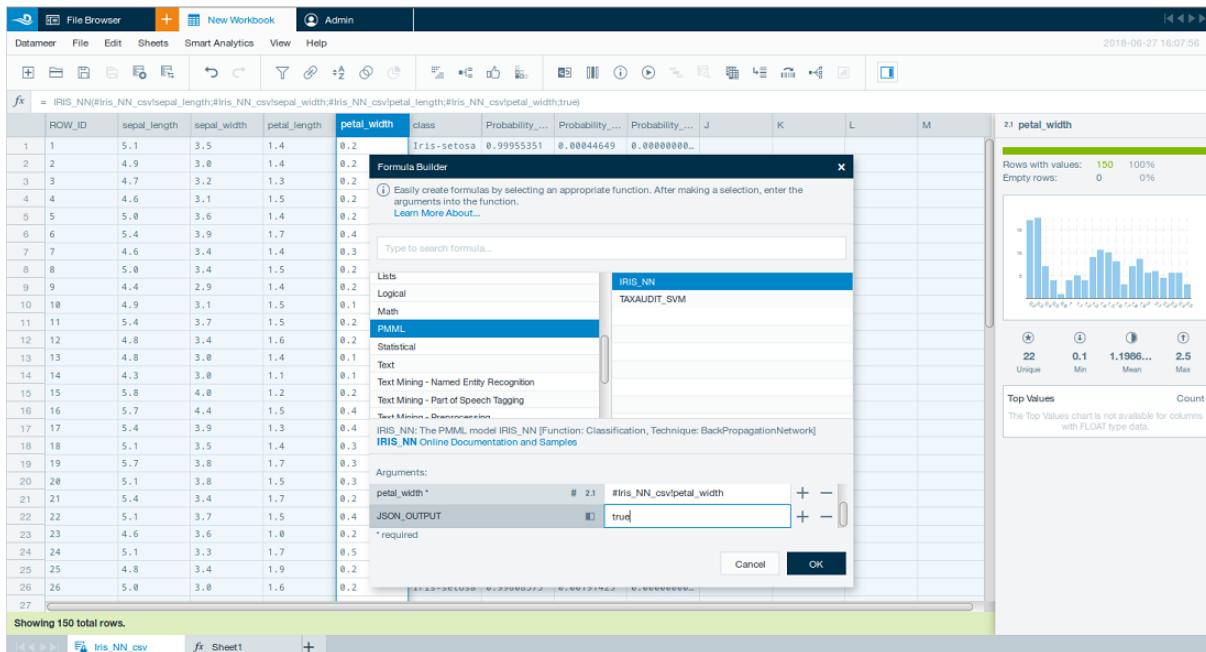
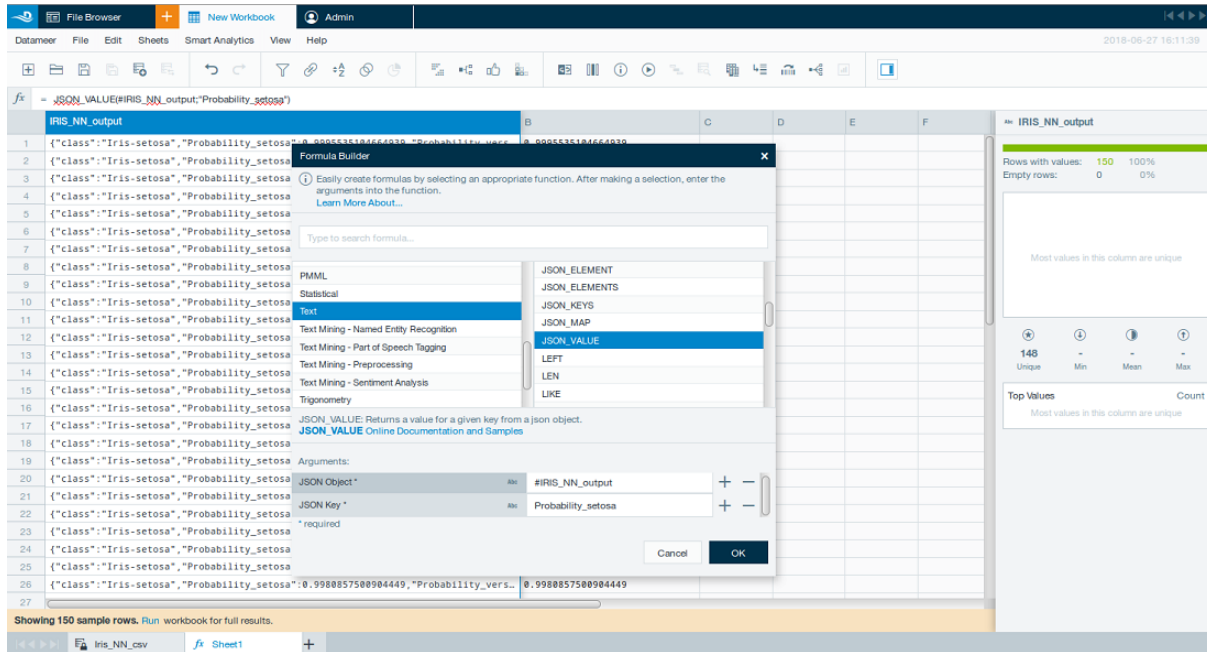


Figure 5.5. Scoring the Iris NN Model in ZEMENTIS for Datameer®



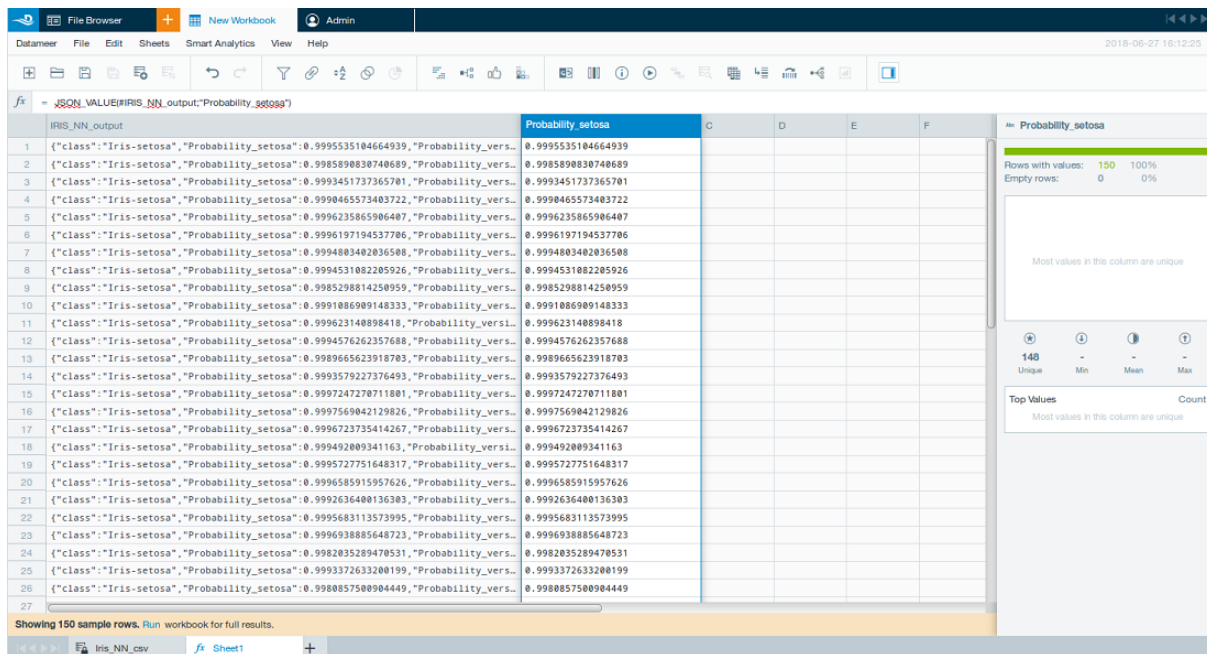
Once you click on the "OK" button, ZEMENTIS for Datameer® will show the preview of the output from the Iris NN model. In our example, the outputs will be written to column "A" of "Result" worksheet as shown in Figure 5.6. In order to separate out a particular output, invoke the "Formula Builder" once again and select function "JSON_VALUE" under "text". Once you do that, the formula wizard will expand to show two arguments: "JSON Array" and "JSON Key". The first, "JSON Array" needs to be mapped to the "JSON_OUTPUT" which is available in column "A". The second argument, "JSON Key" needs to be populated with the output of interest which, in our example, is "Probability_setosa".

Figure 5.6. Multiple outputs with JSON_OUTPUT



By clicking on the "OK" button, the probability for class "Iris_setosa" is then extracted from column "A" and copied to column "B", later manually renamed to "Probability_setosa", as shown in Figure 5.7. Note that the same process could be used to extract the predicted class as well as any of the other class probabilities. After saving our workbook, scoring can be fully processed simply by executing workbook job.

Figure 5.7. Selecting a specific output with JSON_VALUE



Multiple outputs and output types will vary depending on the model type (e.g. SVM, NN, Regression, Decision Tree, Clustering, ...) and function the model implements (e.g. classification, regression, ...). Nonetheless, multiple outputs in PMML are represented in the element "Output". This element determines the types of outputs a model should provide after scoring. The code shown below depicts the "Output" element for the Iris NN model. Note that, as expected, it specifies four outputs, the predicted value "class" together with the three class probabilities.

```
<Output>
  <OutputField name="class" dataType="string" optype="categorical"
    feature="predictedValue"/>
  <OutputField name="Probability_setosa" value="Iris-setosa"
    dataType="double" optype="continuous" feature="probability"/>
  <OutputField name="Probability_versicolor" value="Iris-versicolor"
    dataType="double" optype="continuous" feature="probability"/>
  <OutputField name="Probability_virginica" value="Iris-virginica"
    dataType="double" optype="continuous" feature="probability"/>
</Output>
```

For a list of forums covering PMML and ZEMENTIS products, please make sure to visit [Zementis support site](#). If you need technical support, please visit [Software AG support site](#).