|  |
| --- |
| Department of Software Engineering  Mehran University of Engineering and Technology, Jamshoro |

|  |  |  |  |
| --- | --- | --- | --- |
| Course: SWE324 - Data Warehousing and Data Mining | | | |
| Instructor | Rabeea Jaffari | **Practical/Lab No.** | 10 |
| Date | 2nd Aug 2018 | **CLOs** | CLO-4: P3 & P4 |
| Signature |  | **Assessment Score** | 1 Marks |

|  |  |
| --- | --- |
| Topic | Data mining technique: association rule analysis |
| Objectives | * To learn Data mining using Microsoft Azure Machine Learning Studio. * To perform association rule analysis using Apriori algorithm in Microsoft Azure ML Studio. |

|  |
| --- |
| Lab Discussion: Theoretical concepts and Procedural steps |

**Data Mining:** Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining has an overall goal to extract information (with intelligent method) from a data set and transform the information into a understandable structure for further use Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, complexity considerations, post-processing of discovered structures, visualization, and online updating.

**Microsoft Azure Machine Learning Studio:** Microsoft Azure Machine Learning Studio is a simple browser-based, drag-and-drop tool for building, testing, and deploying Machine Learning (predictive analytics) solutions on our data. Machine Learning Studio publishes models as web services that can easily be consumed by custom apps or tools such as Excel.

It should be noted that Machine Learning can be used to replace Data Mining or vice versa. Data Mining is simpler and very easy to learn and very practical. Machine Learning is a more complex and complete tool. Azure ML Studio allows us to perform both from the simpler data mining to the complex machine learning experiments.

[**https://studio.azureml.net/**](https://studio.azureml.net/)

**Features:**

1. Serverless, drag and drop development
2. Code free intuitive experimentation
3. Deploy your machine learning models as web services in minutes- a web service that can be called from any device, anywhere, and that can use any data source.
4. Azure Machine Learning Studio includes hundreds of built-in packages and support for custom code such as for R and Python language users.

**Azure ML Studio Interactive Workspace:** To develop a predictive analysis model (ML model), you typically use data from one or more sources (data collection), transform and analyze that data through various data manipulation and statistical functions (data cleaning), and generate a set of results. Developing a model like this is an iterative process.

**Azure Machine Learning Studio** gives you an interactive, visual workspace to easily build, test, and iterate on a predictive analysis model. Just drag-and-drop datasets and analysis modules onto an interactive canvas, connecting them together to form an experiment and then run it with the help of “run” button. To iterate on your model design, you edit the experiment, save a copy if desired, and run it again. When you're ready, you can convert your **training experiment** to a **predictive experiment**, and then publish it as a **web service** so that your model can be accessed by others through any custom application.

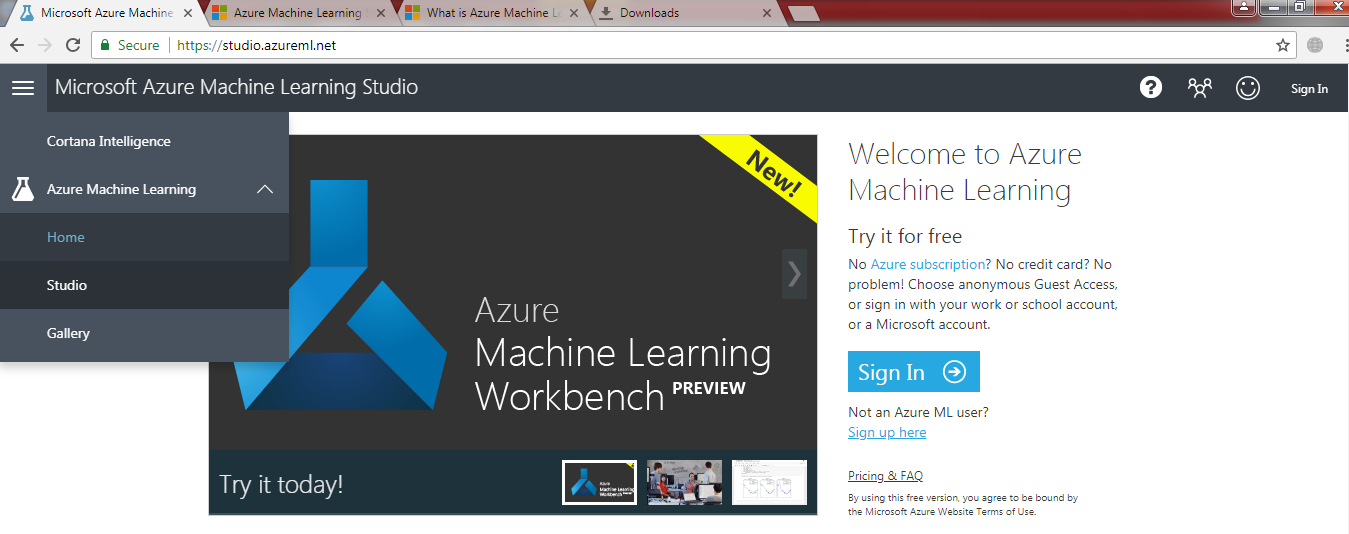
There is no programming required, just visually connecting datasets and modules to construct your predictive analysis model.

**Getting Started With ML Studio:**

When you first enter [Machine Learning Studio](https://studio.azureml.net/) you see the Home page. From here you can view documentation, videos, webinars, and find other valuable resources.

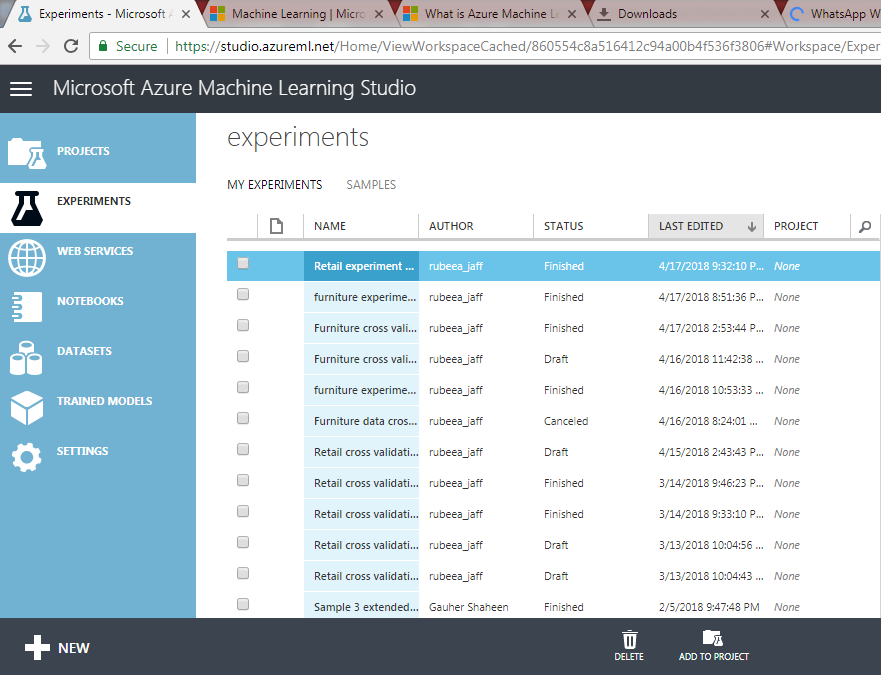
Click the upper-left menu  and you'll see several options. Click on Azure Machine Learning Studio

There are two options here, **Home**, the page where you started, and **Studio** as shown in the screenshot below.



Click **Studio** and you'll be taken to the **Azure Machine Learning Studio**. First you'll be asked to sign in using your Microsoft account, or your work or school account. Once signed in, you'll see the following tabs on the left as shown in the screenshot below:

* **PROJECTS** - Collections of experiments, datasets, notebooks, and other resources representing a single project
* **EXPERIMENTS** - Experiments that you have created and run or saved as drafts that contain Machine learning logic
* **WEB SERVICES** - Web services that you have deployed from your experiments
* **NOTEBOOKS** - Jupyter notebooks for R language that you have created
* **DATASETS** - Datasets that you have uploaded into Studio
* **TRAINED MODELS** - Models that you have trained in ML experiments and saved in Studio
* **SETTINGS** - A collection of settings that you can use to configure your account and resources.



**Experiments:**

Experiments area opens the visual workspace where we’ll easily build, test, and iterate on a ML model. To create a new ML experiment, click on the “Experiments” tab on the left and then the “New” portion in the bottom-left corner of your web page (highlighted by the yellow boundary above)

An ML experiment consists of datasets (sample data) that provide data to analytical modules (ML algorithms), which you connect together to construct a ML model. Specifically, a valid experiment has these characteristics:

* The experiment has at least one dataset and one module
* Datasets may be connected only to modules
* Modules may be connected to either datasets or other modules
* All input ports for modules must have some connection to the data flow
* All required parameters for each module must be set

**Datasets**

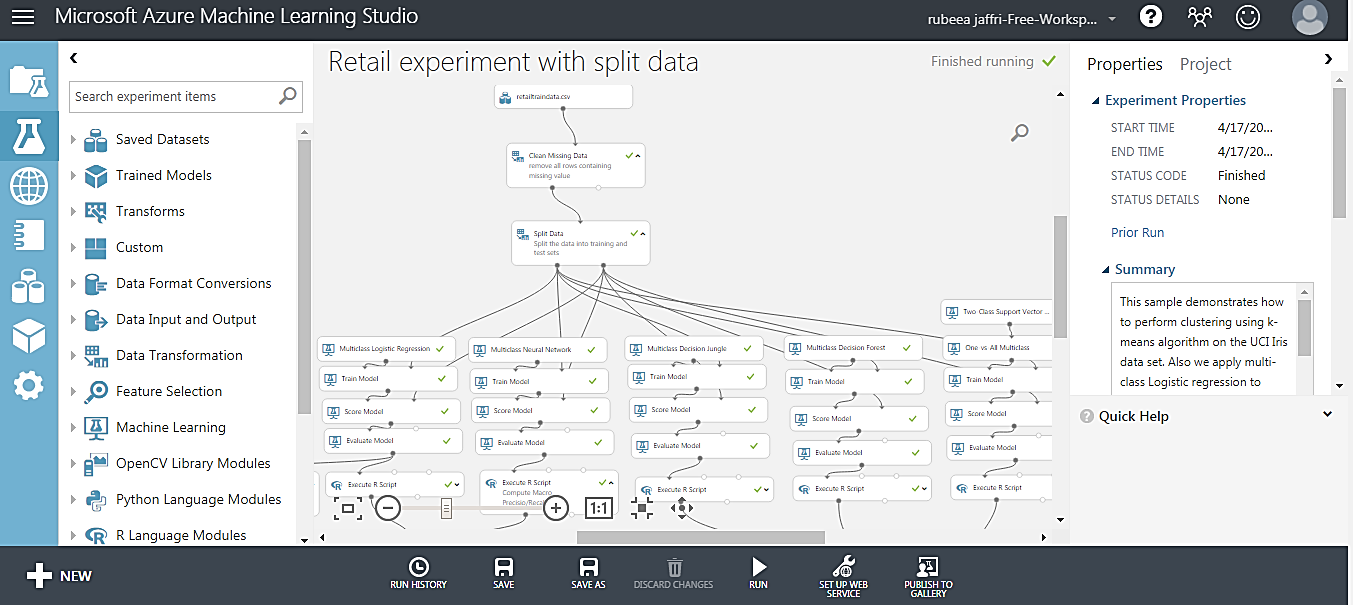
A dataset is data that has been uploaded to Machine Learning Studio so that it can be used in the machine learning process. A number of sample datasets are included with Machine Learning Studio for you to experiment with, and you can upload more datasets as you need them. When you build an experiment you can choose from the list of datasets available to the left of the canvas.

**Modules:**

A module is an algorithm that you can perform on your data. Machine Learning Studio has a number of modules ranging from data ingress functions to training, scoring (prediction), and validation processes. When you build an experiment you can choose from the list of modules available to the left of the canvas.

Modules and datasets to drop onto the canvas on right

Experiment canvas to drag and drop items from left



Run button to execute the experiment when done.

**Association Rule Learning:** Association rule learning is a method for discovering interesting relations between variables in large datasets. It is intended to identify strong rules discovered in datasets using some measures of interestingness.

For example, the rule

**{onions, potatoes}=> {burgers}**

{\displaystyle \{\mathrm {onions,potatoes} \}\Rightarrow \{\mathrm {burger} \}}

found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection, continuous production, and bioinformatics.

**Note:** In contrast with sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions hence the itemsets **{onions, potatoes}** or **{potatoes,onions}** are both equivalent to each other.

**Calculating Association Rules:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Example dataset with 5 transactions and 5 items** | | | | | |
| transaction ID | milk | bread | butter | beer | diapers |
| 1 | 1 | 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 0 | 0 | 1 | 1 |
| 4 | 1 | 1 | 1 | 0 | 0 |
| 5 | 0 | 1 | 0 | 0 | 0 |

Association rules for the above dataset can be derived using the following measures:

**Support:** Support is an indication of how frequently the itemset appears in the dataset. It is calculated as follows:

The support of itemset **X** with respect to dataset **T** is defined as the proportion of transactions **t** in the dataset which contains the itemset **X.**

**Sup(X)= (X is found in t)/|T|**

For example, the itemset **{bread,butter}** has a support of **1/5=0.2** since it occurs in 20% of all transactions (1 out of 5 transactions).

**Confidence:** Confidence indicates the number of times the if/then statements (called association rules) have been found to be true. It is calculated as:

**Conf(X=>Y)= sup(XUY)/supp(X)**

The confidence value of a rule, **X=>Y**, with respect to a dataset of transactions **T**, is the proportion of the transactions that contains **X** which also contains **Y**, where **X** is called the head or **antecedent** of the rule while **Y** is called the tail or **consequent** of the rule.

For example, the rule **{butter,bread} =>{milk}** has a confidence of **0.2/0.2=1.0** in the dataset, which means that for 100% of the transactions containing butter and bread the rule is correct (100% of the times a customer buys butter and bread, milk is bought as well).

**Lift:** The lift of a rule is defined as:

**lift (X=> Y)= supp (XUY)/ (supp(X) x supp(Y))**

or the ratio of the observed support to that expected if X and Y were independent.

For example, the rule **{milk,bread} =>{butter}** has a lift of **{0.2}{0.4\times 0.4}}=1.25.**

* If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no association rule can be drawn involving those two events.
* If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.
* If the lift is < 1, that lets us know the items are substitute to each other. This means that presence of one item has negative effect on presence of other item and vice versa.

**Creating Data Mining Experiment in Microsoft Azure ML Studio:**

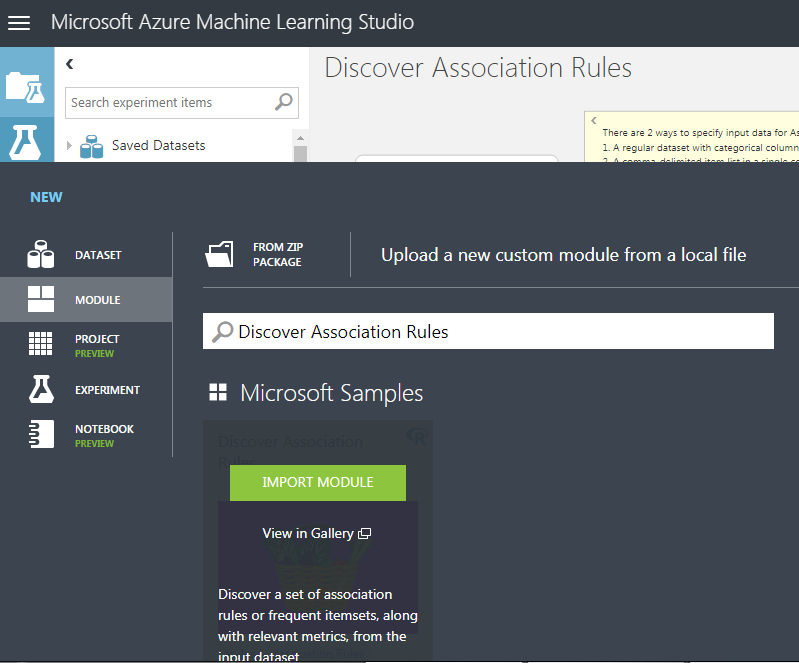
This lab would be using Microsoft Azure ML Studio to find important Association rules and frequent item sets from our data. The steps to create the experiment are as follows:

**Step 1: Import Association Rules Module:**

By default, Microsoft Azure ML Studio doesn’t have its own built-in Association rules module to discover rules from any dataset so we’ll import a custom module that will perform this task for us in our ML Studio workspace. To import a custom module (built by other users and the published online), follow the steps below:

Here's how to get any custom module into your experiment.

1. Import a custom module by clicking **+NEW** at the bottom of the Machine Learning Studio window, select **MODULE** from left pane, type in the name of the custom module as **Discover Association Rules** and then click on **Import Module**.



1. The module is given a default name that you can see while importing. You can rename it because this needs to be unique.

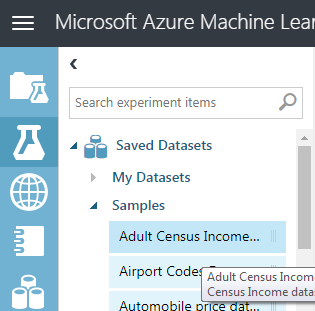
**Step 2: Get data:** There are 2 ways to specify input data for Association Rules:

1. A regular dataset with categorical columns.

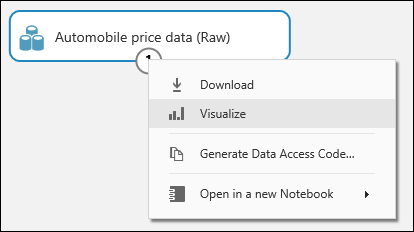
2. A comma-delimited item list in a single column.

We’ll use both of the ways. For the first way we’ll use an already available **“Adult census income”** dataset and for the second way we’ll provide our own dataset separated by commas.

1. Create a new experiment by clicking **+NEW** at the bottom of the Machine Learning Studio window, select **EXPERIMENT**, and then select **Blank Experiment**.
2. The experiment is given a default name that you can see at the top of the canvas. Select this text and rename it to something meaningful, for example, **Discover Association Rules**. The name doesn't need to be unique.
3. To the left of the experiment canvas is a palette of datasets and modules. Type **adult census income** in the Search box at the top of this palette to find the dataset labeled **Adult Census Income Binary Classification data (Raw)**. Drag this dataset to the experiment canvas.

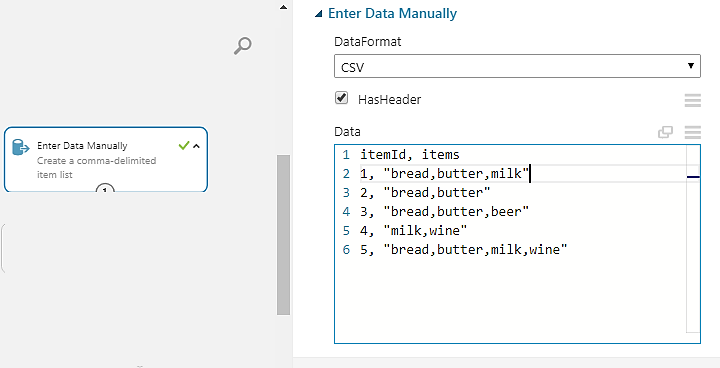


To see what any dataset looks like, click the output port (small circle) at the bottom of the automobile dataset, and then select **Visualize**. You can close the visualization window by clicking the "**x**" in the upper-right corner.



In this sample dataset, each instance of census data appears as a row, and the variables associated with each census appear as columns.

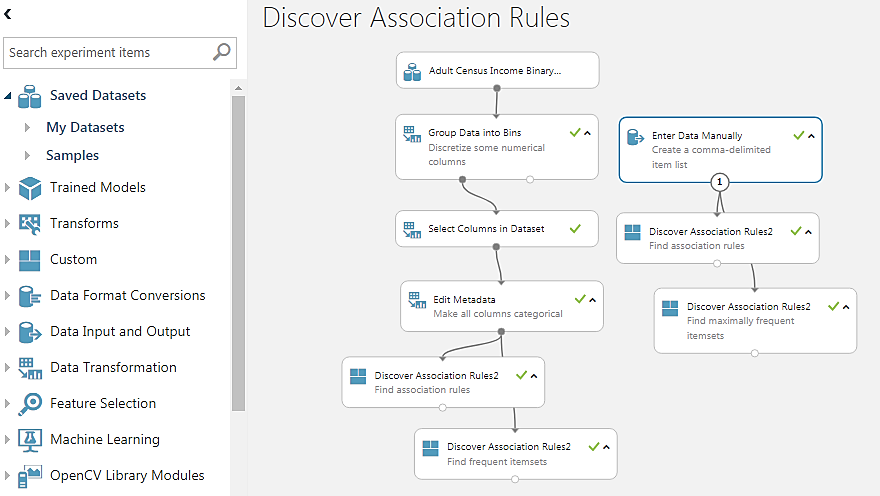
1. For the second data option, from the left pane, drag and drop the **Enter Data Manually** module onto your canvas. Click on it to specify the dataset in CSV in the right properties pane window.



Both the datasets have now been placed in your experiment.

**Step 3: Follow the experiment map:**

Complete the experiment following the map shown below:



Drag and drop modules from the left pane into the experiment canvas according to the map shown above and then specify the module properties as per your needs.

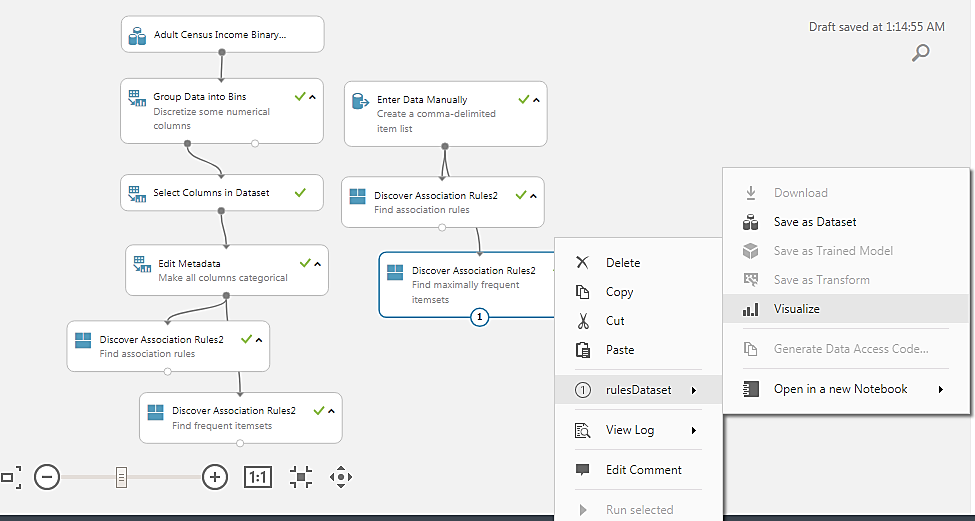
Module settings for ready-made dataset:

|  |  |
| --- | --- |
| **Module Name** | **Properties Setting** |
| Group Data into Bins | * **Binning mode:** quantiles * **Number of bins:** 04 * **Quantile normalization:** QuantileIndex * **Columns:** age,capital-gain,capital-loss,hours-per-week * **Output mode:** Inplace |
| Select Columns in Dataset | * **Columns:** workclass,education,marital-status,occupation,relationship,race,sex,native-country,income,age,capital-gain,capital-loss,hours-per-week |
| Edit Metadata | * **Columns:** workclass,education,marital-status,occupation,relationship,race,sex,native-country,income,age,capital-gain,capital-loss,hours-per-week * **Data type:** Unchanged * **Categorical:** Make categorical * **Fields:** Unchanged |
| Discover Association Rules (finding rules) | * **Input Data type:** Data Frame * **Columns:** All * **Minimal support:** 0.01 * **Minimal Confidence:** 0.5 * **Minimal number of items in a Rule:** 3 * **Maximal number of items in a Rule:** 5 * **Sort by:** support * **Left Hand Side:** (keep this empty) **Right Hand Side:** income=>50K, income=<=50K * **Return type:** Rules |
| Discover Association Rules (finding frequent itemsets) | * **Input Data type:** Data Frame * **Columns:** All * **Minimal support:** 0.1 * **Minimal Confidence:** 0.5 * **Minimal number of items in a Rule:** 2 * **Maximal number of items in a Rule:** 5 * **Sort by:** Confidence * **Left Hand Side:** (keep this empty) **Right Hand Side:** (keep this empty) * **Return type:** Frequent Itemsets |

The module setting for your manually entered data can be altered by yourself accordingly.

**Step 4: Run the experiment:** Finally run your experiment by clicking at the **Run** button located at the bottom of your experiment canvas.

**Step 5: Interpreting the results:** The output from both the datasets can be seen by clicking the output port (small circle) at the bottom of the **Discover Association Rules** module, and then selecting **Visualize.**



|  |
| --- |
| Lab Tasks |

* **Take any dataset in CSV format and derive association rules and frequent item sets from it using any one of the approaches described above.**

