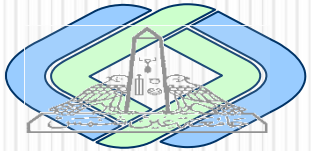
Faculty of Computer
and Information Sciences
Ain Shams University



كلية الحاسبات والمعلومات
جامعة عين شمس

Data Science

Code:

Instructor

Prof.Dr. Abeer M. Mahmoud

Professor of Computer Science -faculty of Computer and Information Sciences- Ain Shams University

Data Science and Big Data Analytics v2

Topics : Data Science and Big Data Analytics Course

Introduction to Big Data Analytics + Data Analytics Lifecycle	Review of Basic Data Analytic Methods Using R	Advanced Analytics – Theory and Methods	Advanced Analytics - Technology and Tools	The Endgame, or Putting it All Together + Final Lab on Big Data Analytics
Big Data Overview State of the Practice in Analytics The Data Scientist Big Data Analytics in Industry Verticals Data Analytics Lifecycle	Using R to Look at Data - Introduction to R Analyzing and Exploring the Data Statistics for Model Building and Evaluation	K-means Clustering Association Rules Linear Regression Logistic Regression Naive Bayesian Classifier Decision Trees Time Series Analysis Text Analysis	Analytics for Unstructured Data (MapReduce and Hadoop) The Hadoop Ecosystem In-database Analytics – SQL Essentials Advanced SQL and MADlib for In-database Analytics	Operationalizing an Analytics Project Creating the Final Deliverables Data Visualization Techniques + Final Lab – Application of the Data Analytics Lifecycle to a Big Data Analytics Challenge

Advanced analytics—theory and methods

Advanced analytics—theory and methods

Upon completing this module, you should be able to:

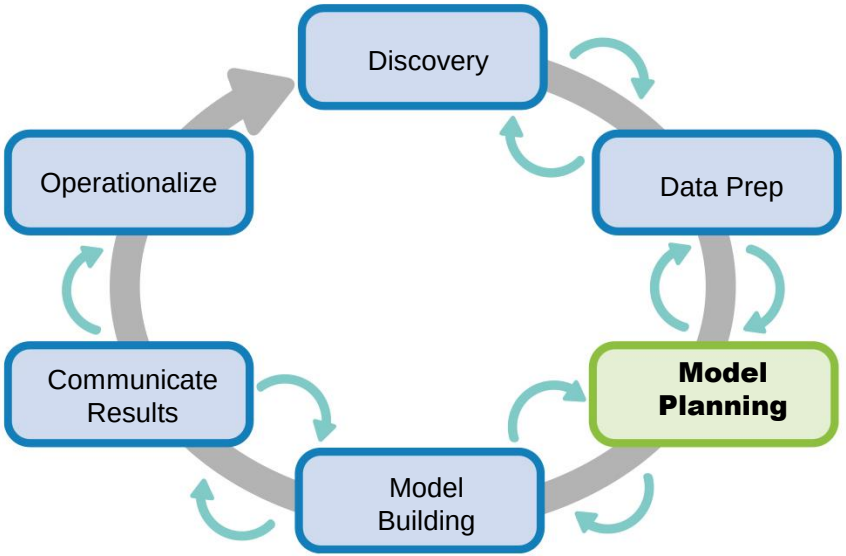
- ✓ Select an appropriate analytic technique based on the business problem reframed as an analytic challenge and based on the data's structure.
- ✓ Explain the technical foundations of commonly used analytic methods.
- ✓ Use R to fit, validate, and evaluate analytic models.

Lesson: Introduction to advanced analytics—theory and methods

Model planning—key activities

How do people generally solve this problem with the kind of data and resources I have?

- What are related or analogous problems? How are they solved? Can I do that? How can I improve on previous approaches?
- What are the model assumptions?
- Do I need extra data prep and transformations?



What kind of problem do I want to solve? How do I solve it?

Problem to solve	Category of techniques	Covered in this course
I want to group items by similarity. I want to find structure—commonalities in the data.	Clustering	K-means clustering
I want to discover relationships between actions or items.	Association rules	Apriori
I want to determine the relationship between the outcome and the input variables.	Regression	Linear regression Logistic regression
I want to analyze my text data.	Text analysis	Regular expressions, document representation—Bag of Words, TF-IDF
I want to assign known labels to objects.	Classification	Naïve Bayes Decision trees
I want to find the structure in a temporal process. I want to forecast the behavior of a temporal process.	Time series analysis	ARIMA

Why these analytic techniques?

- Most popular, frequently used:
  - These techniques provide the foundation of data science skills on which to build
- Relatively easy for new data scientists to understand and comprehend
- Applicable to a broad range of problems in several verticals

Lesson: K-means clustering

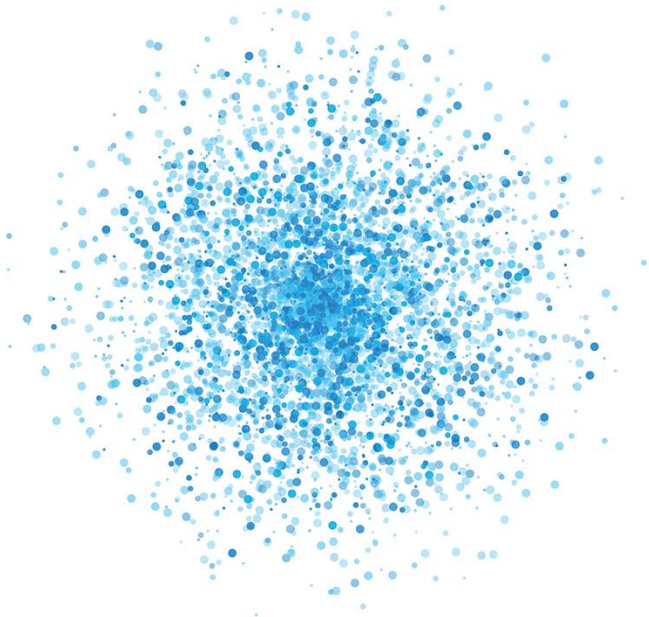
K-means clustering

During this lesson, the following topics are covered:

- Clustering—unsupervised learning method
- K-means clustering:
  - Use cases
  - The algorithm
  - Determining the optimum value for K
  - Diagnostics to evaluate the effectiveness of the method
  - Reasons to choose (+) and cautions (-) of the method

Clustering

- How do I group these documents by topic?
- How do I group my customers by purchase patterns?
- Sort items into groups by similarity:
  - Items in a cluster are more similar to each other than they are to items in other clusters.
  - Detail the properties that characterize similarity.
    - Or, detail the properties of distance, the "inverse" of similarity.
- Not a predictive method; finds similarities, relationships
- **Example: K-means clustering**



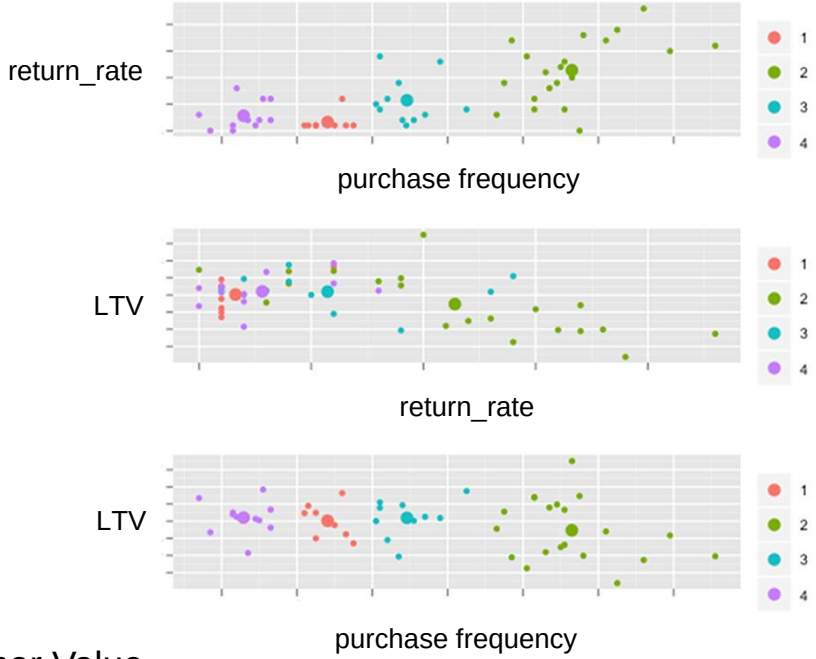
K-means clustering—what is it?

- Is a type of unsupervised learning used when you have unlabeled data
- Aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean
- **Input:** numerical. There must be a distance metric defined over the variable space.
  - Euclidian distance
- **Output:** the centers of each discovered cluster, and the assignment of each input datum to a cluster
  - Centroid

K means clustering—use cases

- Often an exploratory technique
  - Discover structure in the data
  - Summarize the properties of each cluster
- Sometimes a prelude to classification
  - Discovering the classes
- Examples
  - The height, weight, and average lifespan of animals
  - Household income, yearly purchase amount in dollars, number of household members of customer households
  - Patient record with measures of BMI, HBA1C, HDL
  - Cluster regions across a country based on sales, sensitivity, risk

Use-case example—online retailer

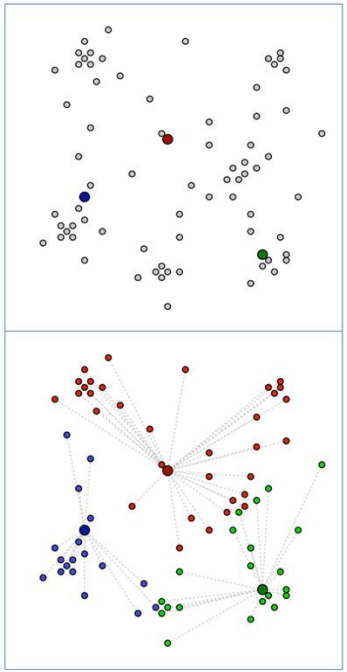


LTV – Lifetime Customer Value

Algorithm

Step 1: Choose K; then, select K random "centroids." In this example, K equals 3.

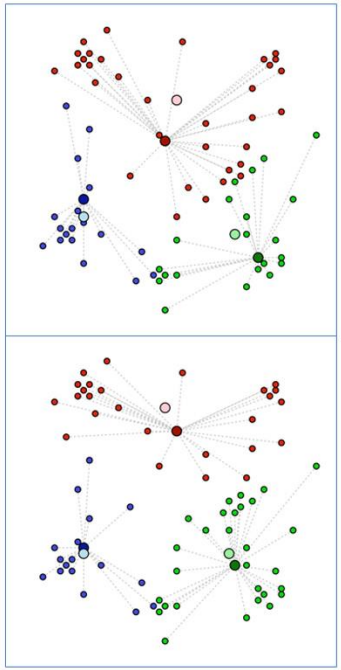
Step 2: Assign records to the cluster with the closest centroid.



Algorithm, cont.

Step 3: Recalculate the resulting centroids.
Centroid: the mean value of all the records in the cluster.

Step 4: Repeat steps 2 and 3 until record assignments no longer change.



Picking K

Heuristic: find the elbow of the Within Sum of Squares (WSS) plot as a function of K.

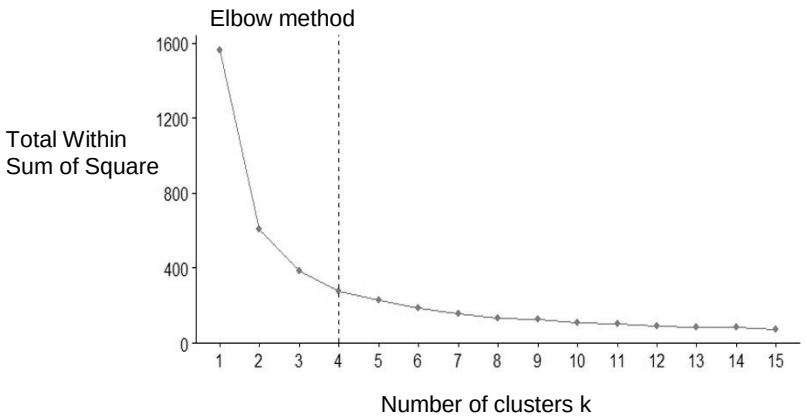
$$\text{WSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |x_{ij} - c_i|^2$$

K: # of clusters
n_i: # points in i th cluster
c_i: centroid of i th cluster
x_{ij}: j th point of i th cluster
$|x_{ij} - c_i|$ Denotes distance between the centroid and point x

Elbow in the plot is at K = 4

Scaling enables each variable to have equal impact in clustering.

Optimal number of clusters



Diagnostics—evaluating model

- Do the clusters look separated in at least some of the plots when you do pair-wise plots of the clusters?
  - Pair-wise plots can be used when there are not many variables.
- Do you have any clusters with few data points?
  - Try decreasing the value of K.
- Are there splits on variables that you would expect but do not see?
  - Try increasing the value K.
- Do any of the centroids seem too close to each other?
  - Try decreasing the value of K.
- Do the means of variables used for clustering vary across the clusters?
  - Try decreasing the value of K.

K-means clustering—reasons to choose (+) and cautions (-)

Reasons to choose (+)	Cautions (-)
Easy to implement	Does not handle categorical variables
Easy to assign new data to existing clusters Which is the nearest cluster center?	Sensitive to initialization—first guess
Concise output Coordinates the K cluster centers	Variables should all be measured on similar or compatible scales Not scale-invariant!
	K, the number of clusters, must be known or decided in a way based on theoretical deduction Wrong guess: possibly poor results
	Tends to produce "round," equal-sized clusters Not always desirable

Test yourself

1. Why is K-means clustering considered an unsupervised machine learning algorithm?
2. Detail the four steps in the K-means clustering algorithm.
3. How do you use WSS to pick the value of K?
4. What is the most a common measure of distance used with K-means clustering algorithms?
5. The attributes of a dataset are purchase decision (Yes/No), gender (M/F), income group (<10K, 10-50K, >50K). Can you use K-means to cluster this dataset?

K-means clustering—summary

During this lesson, the following topics were covered:

- Clustering—unsupervised learning method
- K-means clustering
- Use cases with K-means clustering
- The K-means clustering algorithm
- Determining the optimum value for K
- Diagnostics to evaluate the effectiveness of K-means clustering
- Reasons to choose (+) and cautions (-) of K-means clustering

Lesson: Association rules

Association rules

During this lesson, the following topics are covered:

- Association rules mining
- Apriori algorithm
- Prominent use cases of association rules
- Support and confidence parameters
- Lift and leverage
- Diagnostics to evaluate the effectiveness of rules generated
- Reasons to choose (+) and cautions (-) of the Apriori algorithm

Grocery store—scenario

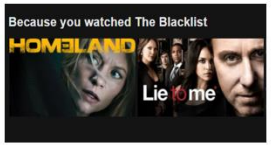
- In order to increase the volume of sales, grocery store managers may want to perform an analysis for understanding the products that shoppers purchase together.
- For example, If the managers find that there is a higher probability of a customer purchasing bread and milk together, they may want those items in aisles that are close to each other. This way, there is more possibility of customers buying milk easily when they come in to buy bread, which increases the volume of sales.
- At this point, association rules come into play and help identify those patterns.

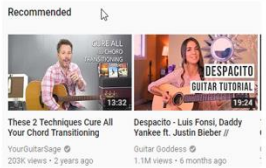
Association rules : do u consider it a predictive method??

- Help identify interesting patterns and connections among sets of items
  - Rules take the form of "If X is observed, then Y is also observed"
  - The definition of "interesting" varies with the algorithm used for discovery.
- Use case: Understand customer buying habits by finding associations between the different items that customers place in their "shopping basket"
  - Known as market basket analysis
- Not a predictive method

Association rules—examples

Examples where association rules could be applied





Amazon: Notice the "Customers who bought this item also bought" section in the amazon website.

Netflix: For every movie watched, there is a recommendation for movie Y "Because you watched movie X."

YouTube: Based on your viewing pattern, YouTube has a "Recommended" section that finds the relationships.

Algorithm for association rules—Apriori

- Apriori—an algorithm for mining frequent itemsets for the Boolean association rules
  - Uses a bottom-up approach where frequent subsets are extended one item at a time
  - Designed to operate on datasets containing transactions
- Used over itemsets—sets of discrete variables that are linked
- Possible transactional datasets
  - Retail items that are purchased together
  - A set of tasks completed in one day
  - A set of links one user clicks in a single session
- Four common ways to measure association
  - Support
  - Confidence
  - Lift
  - Leverage

Apriori algorithm

- Earliest of the association rule algorithms
- Frequent itemset: a set of items L that appear together often enough:
  - Formally: meets a **minimum support** criterion --? Explain
  - **Support:** the percentage of transactions that contain the itemset, which shows how popular an itemset is; this percentage is measured by the proportion of transactions that contain it
- Apriori property: Any subset of a frequent itemset is also frequent
  - It has at least the support of its superset.

Apriori algorithm—support example

Consider the grocery store example with 1000 transactions and the following items in the basket. Minimum support is 50 percent.

Items in the basket
Milk, Bread, Eggs, Cookie, Meat, Potatoes

Frequent itemset	Support
Milk	70%
Bread	81.3%
Milk, Bread	62.7%

Itemset {Milk, Bread} has minimum support

Possible rules are Milk → Bread, Bread → Milk

If you find that an itemset greater than a certain proportion tends to have significant impact on profits, then that proportion can be used as threshold for support. You can identify itemsets with support values greater than the threshold as significant itemsets.

Apriori algorithm—confidence

- Iteratively grow the frequent itemsets from size 1 to size K, or until you run out of support.
  - Apriori property tells you how to prune the search space.
- Frequent itemsets are used to find rules X → Y with a minimum **confidence**:
  - **Confidence:** The percentage of transactions that contain X that also contain Y.
- Output: The set of all rules X → Y with minimum support and confidence.

030. Apriori algorithm—confidence example

In the grocery store example, you have 1000 records with the following combinations:

	Meat	Bread	Eggs	Total
Cookies	44	186	70	**300**
Milk	34	627	39	**700**
	78	**813**	**109**	

Out of 813 shoppers who buy bread, 627 buy Milk, as well.

Milk → Bread = 627/700 = 89.57%

Bread → Milk = 627/813 = 77.12%

One drawback of the confidence measure is that it might misrepresent the importance of an association. It only accounts for how popular Milk is, but not Bread. If Bread is also popular, in general, there is a higher chance that a transaction containing Milk also contains Bread, inflating confidence measure.

To account for the base popularity of both constituent items, you use a third measure called lift.

Lift and leverage

Lift explains how likely item Y is purchased when item X is purchased, as if X and Y are statistically independent.

$$\text{Lift}\,(X \rightarrow Y) = \frac{\text{Support}\ (X \cup Y)}{\text{Support(X)} * \text{Support(Y)}}$$

Lift = 1 indicates no association between items. This relationship is coincidental.

Lift > 1 indicates that item Y is likely to be bought if item X is bought. Relationship is interesting.

Lift < 1 indicates that item Y is unlikely to be bought if item X is bought.

Leverage measures the difference in the probability of X and Y appearing together in the dataset compared to what would be expected as if X and Y were statistically independent.

$$\text{Leverage}\,(X \rightarrow Y) = \text{Support}\,(X \cup Y) - \text{Support(X)} * \text{Support(Y)}$$

$$\text{Lift}(X \to Y) = \frac{\text{Support}(X \wedge Y)}{\text{Support}(X) * \text{Support}(Y)}$$

$$\text{Leverage}(X \to Y) = \text{Support}(X \wedge Y) - \text{Support}(X) * \text{Support}(Y)$$

Apriori algorithm—lift and leverage example

In this grocery store example:

	Meat	Bread	Eggs	Total
Cookies	44	186	70	**300**
Milk	34	627	39	**700**
	78	**813**	**109**	

You calculated the confidence for the following rules.

Milk → Bread = 627/700 = 89.57%

Bread → Milk = 627/813 = 77.12%

Lift of these two rules = 0.627/0.700*0.813 = 1.10.

Since lift is > 1, the relationship is interesting and true.

Sketch of algorithm

- If L_k is the set of frequent k-itemsets:
  - Generate the candidate set C_{k+1} by joining L_k to itself.
  - Prune out the (k+1)-itemsets that do not have minimum support. Now, you have L_{k+1}.
- You know this algorithm catches all the frequent (k+1)-itemsets by the Apriori property.
  - A (k+1)-itemset cannot be frequent if any of its subsets are not frequent.
- Continue until you reach k_{max}, or run out of support.
- From the union of all the L_k, find all the rules with minimum confidence.

Step 1—1-itemsets (L1)

- Let min_support = 0.5
- 1000 transactions
- Scan the database
- Prune

Frequent Itemset	Count
Bread	813
Milk	700
Soda	631
Potatoes	550

Step 2—2-itemsets (L2)

- Join L1 to itself
- Scan the database to get the counts
- Prune

Frequent Itemset	Count
Milk, Soda	544
Milk, Bread	627

Step 3—3-itemsets

- You have run out of support.
- Candidate rules come from L2:
  - Milk → Soda
  - Soda → Milk
  - Milk → Bread
  - Bread → Milk

Finally—find confidence rules

Rule	Set	Cnt	Set	Cnt	Confidence
IF Milk THEN Soda	Milk	700	Milk AND Soda	544	544/700 =77%
IF Milk THEN Bread	Milk	700	Milk AND Bread	527	527/700 =75%
IF Soda THEN Milk	Soda	631	Soda AND Milk	544	544/631 =86%
IF Bread THEN Milk	Bread	813	Bread AND Milk	627	627/813 =77%

If you want confidence > 80%:

IF Soda THEN Milk

Diagnostics

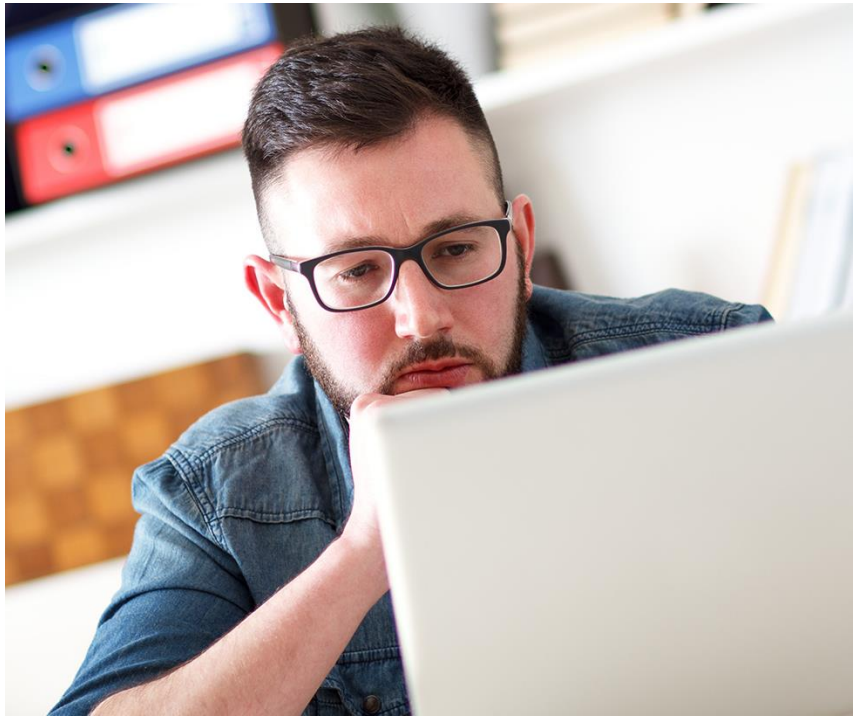
- Do the rules make sense?
  - What does the domain expert say?
- Make a "test set" from hold-out data:
  - Enter some market baskets with a few items missing, selected at random. Can the rules determine the missing items?
  - Remember, some of the test data may not cause a rule to fire.
- Evaluate the rules by lift or leverage.
  - Some associations may be coincidental, or obvious.

Apriori—reasons to choose (+) and cautions (-)

Reasons to choose (+)	Cautions (-)
Easy to implement	Requires many database scans
• Uses a clever observation to prune the search space – Apriori property	Exponential time complexity
Easy to parallelize	• Can mistakenly find spurious, or coincidental, relationships – Addressed with lift and leverage measures

Check your knowledge

1. What is the Apriori property, and how is it used in the Apriori algorithm?
2. List three popular use cases of the association rules mining algorithms.
3. What is the difference between lift and leverage? How is lift used in evaluating the quality of rules discovered?
4. Define support and confidence.
5. How do you use a hold-out dataset to evaluate the effectiveness of the rules generated?



Association rules—summary

During this lesson, the following topics were covered:

- Association rules mining
- Apriori algorithm
- Prominent use cases of association rules
- Support and confidence parameters
- Lift and leverage
- Diagnostics to evaluate the effectiveness of rules generated
- Reasons to choose (+) and cautions (-) of the Apriori algorithm