Abstract

In this Master Thesis memory will be described a full end-to-end data science project performed in CleverData, a successful start-up specialized in machine learning techniques and analytics tools. Over all its capacities, it offers a huge variety of solutions to nowadays business needs from different domains.

This project was performed for one of its client, an important retail company from Spain. It consist of analysing the market basket of customers. Thus, the main goal is to find which items are purchased together in their stores.

Through the memory, the reader will see how, step by step, the project grows. Since the first step of defining objectives, until the last one of results delivery. Moreover, the reader will see one of the most promising tools used for machine learning as a service nowadays, BigML.

At the end of the project, the reader will have a general idea how data science projects are structured, and how machine learning can be used to solve real problems in today's companies.

1 Introduction

1.1 Introduction & Motivation

Retail has evolved through its life. Since the common corner stores from the 1900s, until the new e-commerce that has shaken the retail world to its core. This changing process has lead to new era possibilities for the commerce and the consumer.

Consumers nowadays has a wide range options, independently the commerce domain. In the past, when the consumer had to buy something, he only could choose a product from the catalog of the store. However, with the new era of information and globalization, the list of options has increased exponentially. Now consumers can choose between a huge variety of products. Limitations as geography, season are not more an issue. Products that years ago were considered as luxury goods are considered as common. All of this, caused that companies have a limitless possibilities nowadays. However, this limitless of possibilities caused a huge amount of new competitors as well. Companies have being forced to think new strategies in order to attract new customers or keep its current customers.

This concept is the one that caused this project. Our client is an important retail company from Spain. He posses a supermarket chain with a wide list of daily consumers. To increase the experience of the customer and increase its incomes as well, the client decided to invest analysing customer's behaviour and its purchases using knowledge discovery and data mining process [1], and specifically, the items associations rules of its stores [2]. This field in retail domain is known as market basket analysis.

Market basket analysis [3] encompasses a broad set of analytics techniques aimed at uncovering the associations and connections between specific objects, discovering customers behaviours and relations between items. In retail, is used based in the following idea, if a customer buy a certain group of items, is more (or less) likely to buy another group of items. For example, it is known that when a customer buy beer, in most of cases, buys chips as well. These behaviours produced in the purchases is what the client was interested on. The client was interested in analysing which items are purchased together in order to create new strategies that improved the benefits of the company and customers experience.

1.2 Definition of the problem & Objectives

For any client we have in CleverData, when a data science project is done we end-up with a set of results. This results tells to the client what is happening on its business. However, the results by itself are just a part of the entire project, to obtain a really benefit, some actions or

strategies has to be taken to extract value of it. There are three domains where market basket analysis is used for.
The first domain is the creation of personalized recommendations [4]. This methodology is well known nowadays. During the explosion of the e-commerce, personalized recommendations has appeared as a part of the marketing process. Basically, the idea consist in suggesting items to customer based on its preferences. There are two basic ways to do it. The first one, is suggesting items similars the ones the customer has purchased in the past. The second one, is looking for similar customers and recommending items that had purchased the others. Both strategies are often used for companies in order to realize cross-selling and upselling strategies.

The second domain where market basket analysis is used is in the analysis of spatial distribution in chain stores [5]. Due the increasing number of products that nowadays exist, physical space in stores has started to be a problem. More and more, stores invest money and time trying to find which distribution of items can lead them to obtain more sells. Due that, knowing in advance which items are commonly purchased together, the distribution of the store can be changed in order to sell more products.

The last domain is in the creation of discounts and promotions. Based in customers behaviour, special sales can be performed. For example, if the client knows which items are often purchased together, he can create new offers based in order to increase the sells of those items.

As it can be seen, market basket analysis can be used to help retail business in many fields. That's why the client contacted with CleverData, to help him to discover, which associations rules were in its stores.

1.3 Our Market Basket Analysis strategy

The first issue we had at the moment the client decided to realize this project with us was the decision of how to approach the project. The objective was clear, find which items are purchased together. However, the way we had to focus the project was not so clear. When the client contacted us in the first meeting, he had the idea to create associations rules for each store. However, when we started to consider the project and think which results could be useful for the client, we started to consider an issue, if it had sense discover associations rules for each store.

One can simply think that find the associations rules using all the company's tickets would end the project, however, that's not the case. For instance, imagine we want to create a new offer based on the rules we have discovered. Then, we choose a random store where we want to apply the new offer. However, at the moment to create the offer we see that this store doesn't sell the items from the rule, so it can not be created. This obviously can be solved just looking for another rule, however this made us to realise that stores have different behaviours, and maybe, just discovering associations rules for each store was not enough to obtain truly valuable results. We needed something else.

That's how we ended-up with the idea of creating a store clustering [6]. With that, we could capture the behaviours of each store and create rules that were more valuable. The process is the following, once created the different clusters, for each of them, we selected the store with less distance to the centroid, then we found the association rules on that shop and extrapolated the results to all the stores of that cluster. In addition, the client was interested with this clustering due the way they classify their stores is based in other metrics not related to the behaviour of the stores.

Another issue we had to consider is in which level we had to perform the associations rules. Each item belongs to a set of levels. For instance, the item *"patatas lays clasicas 170 grs"* belongs to family *"patatas fritas y fritos"*, the section *"alimentación seca"* and the sector *"alimentación y bebidas"*. This is used by the client to classify its itemset and logistics processes [7]. Is a common practise that any retail company do. In our case, the client wasn't interested in finding rules of items, he was interested in rules based at family level. The client wanted to know which families are purchased together in order to change its distribution on the stores. In addition, the client was just interested in the items from the sectors: "alimentación y bebidas", "productos frescos", "drogueria y perfumeria" and "bazar". Due that, all the project was done with the items of these sectors.

To conclude, we defined that clusters and associations will have to be retrained periodically. Over time, people behaviours change, appears new products or new stores. Due that, machine learning models have to be retrained in order to capture new behaviours [8]. This periodically task has to be done in any data science project.

Once we knew the client needs, we defined how would be the project and the objectives we wanted to achieve. The project was defined in two steps, the first one was the creation of the store clustering, and the second one was the association discovery for each cluster. We agreed with the client that the project will be considered finalized with the delivery of the list of clusters and the associations rules for each of them.

In the next chapter will be described the procedure we usually use in order to perform a data science project.

2 State of the art

2.1 A Data Science project

To realize a data science project there are some steps that always have to be done. This steps are the skeleton of any project. In CleverData, we always follow this procedure (Figure 2.1), and consequently, it was the one we performed in this project. Each step form part of the total project procedure, with its own characteristics and objectives.

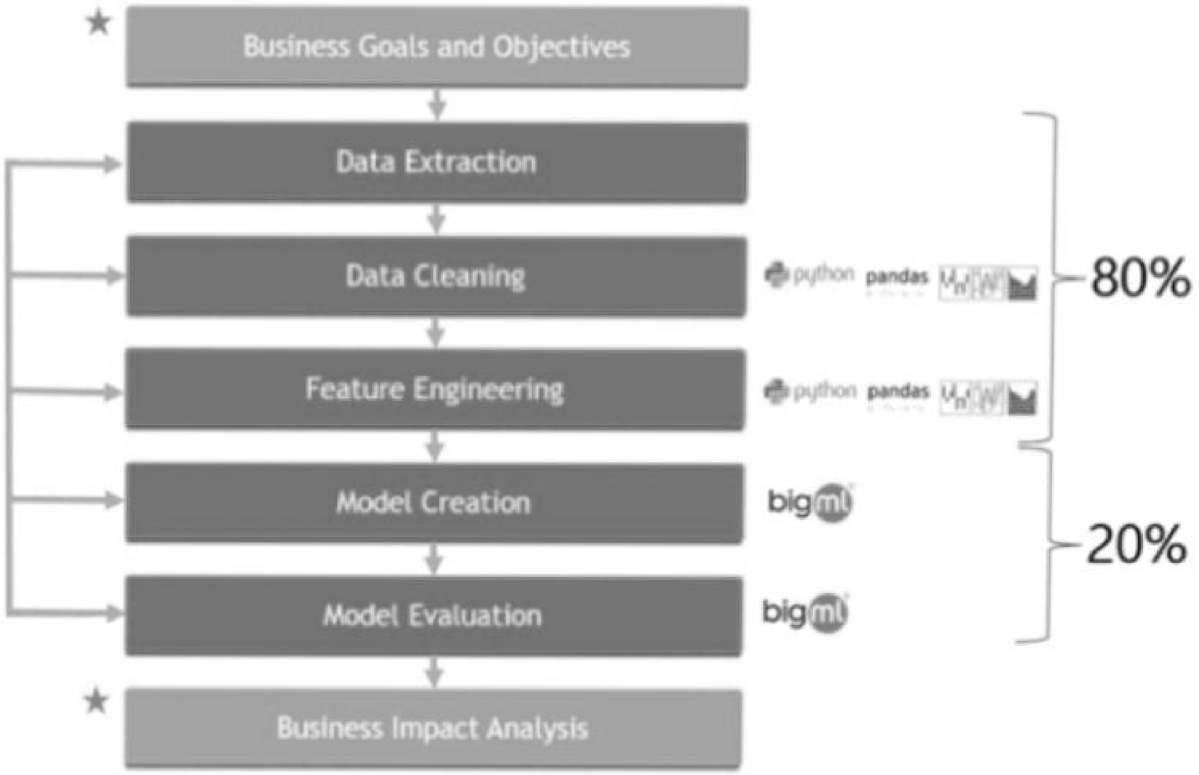


Figure 2.1: data science project skeleton.

Each rectangle represent a step in the project. On the one hand, the steps from *Data Extraction* until *Model Evaluation* are related to a common data science project. Those steps are iterative between them. On the other hand, the steps with a star represent the ones that vary depending the business and its needs.

The images on the right of each rectangle are the tools used on that step. As the reader can see, the project itself didn't need a lot of tools to be performed, that's because the tools used are really powerful. Those tools used in project are Python [9] and the Pandas library for the steps related to data mining, and BigML [11] for the model algorithms steps.

Big numbers are time cost approximations over total time of the project. One could have a priori idea that most time spended in a project is the model creation and evaluation. However, is absolutely opposite. Data transformation process consume most time of the project.

Rows are the flux between steps. This is one of the most remarkable characteristics of a data science project. Flux on traditionals projects are sequentials, there is just one iteration, however, in this type of projects, one has to work on iterations. A finished process or step can be repeated due some new condition or result.

2.1.1 Business Goals and Objectives

The base of any project are the goals and objectives that have to be achieved. This first step is really important. Decisions and strategies decided here will affect all the project itself and the direction where it will be developed .

In this step the client introduce what he wants to achieve using machine learning techniques. Then, our task is perform an analysis of those objectives in order to understand them and decide if they can be achieved using machine learning algorithms. If they can be achieved, we define how the project will be performed and which results are the one wants we want to achieve.

Sometimes, companies contact with us thinking that the issue they have can be solved using machine learning algorithms when are not. Most of the time, is caused due a general confusion about what really is machine learning and its capacities.

2.1.2 Data Extraction

Data extraction is the process of collecting all the historical data of a company. This data is considered raw due it hasn't received any treat previously.

Data extraction is the first step that can be considered part of the data transformation process. The collection of data sometimes can be a hard work due each client has it own way to store the data. Oftenly, data is distributed among different resources and have different formats. Other times, data is poorly structured or even unstructured. All these aspects makes data extraction a hard task.

There are many tools nowadays that suits in this type of problems. Each of them has its own characteristics and methodologies, however, even with this help, the process to collect data can imply a huge work due use this tools and process are not a simple job.

2.1.3 Data Cleaning

Data cleaning is the process of detecting and removing corrupt or inaccurate records from historical data [12]. One of the most different things about a data science project done at the university and the real world is this cleaning process. Datasets used to learn and practise are most of times already clean, they don't need to be treated.

The problem is that in real world, that's not the case. Datasets have a lot of errors. This errors are caused from different causes and the detection of them is vital for the project. Invalid records will imply deterioration of the future model adding noise or false information.

Cleaning data is used in the process of removing data that is not relevant or needed as well. Part of the work is to know which information is relevant or can apport value to the algorithm and treat it for each specific case. Another common situation is data duplicated. Due databases are from big companies and comes from different sources sometimes the information is repeated. This provoke an overlap of information absolutely useless.

2.1.4 Feature Engineering

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work [13]. This process is fundamental in a data science project, however, is difficult and expensive. Due this high cost, most of the time of the project is spent in this task.

The task consist of finding features that add value and which ones don't. The process encompassed creation, transformation and deletion of features, and try over all the cases the quality of the model created with those features.

Features used to train a machine learning model affect the performance of it. As better are the features, better will be the performance. The quality and quantity of features have a huge impact in the project. More than the hyperparameter configuration of the algorithm, the features are the ones that add value to the model. It is worth investing time creating new features, analysing them and transforming data than trying different algorithms. Over 80% time of the project was this step.

A complete process of generating a predicting model could be the same as the one used in a cooking recipe. Ingredients would be the data and the algorithm the recipe. If the ingredients are in poor state doesn't matter that the recipe is the best one of the world, the food resulting will be bad. In the same way, if data has no quality, even with the best algorithm, the results are bad.

2.1.5 Model Creation

Once created the features, the machine learning model is trained. Models are feed using the data provided. Algorithm can be supervised or unsupervised and depending the objective, the project will be a classification or regression task.

In order to capt the changing behaviour of the data, those machine learning models have to be retrained periodically. This time has to be defined with the client and according to the needs of the problem.

2.1.6 Model Evaluation

The last step in the standard procedure is the model evaluation. The performance of it is the result of all the work done along the process. Depending the type of the problem there are some metrics to evaluate the performance of a model.

There are two types of evaluation, offline and online. The first one, analyse the performance of a model a priori before put it in production. The basic ways to do it is with a 80/20 split of the dataset or performing a cross-validation. The second one, test the model in current data and analyse its performance. One famous tactic used is the A/B testing. This consist of selecting a subset of instances from the total set and evaluating the results of the model to that subset.

2.1.7 Business Impact Analysis

The last step in a data science project is the impact analysis that had the solution. Companies usually trend to perform projects in order to obtain a monetary benefit. It can be directly or indirectly. On the one hand, an example of a model used to obtain a direct monetary benefit is one used for churn prediction. It give a direct income to the company due it prevents to lose clients that would churn. On the other hand, an example of a model used to obtain indirect benefits could be one that group customers based on its behaviours for a posteriori marketing strategy. This model doesn't feedback with a direct income, however, the knowledge of the patterns of that customers can lead to future incomes.

In CleverData we can help to our client to understand what is data telling about a business, however, at the end, the client is the one who has to apply the corresponding actions.

2.2 BigML

One of the most discussed topics in the BigData and Machine Learning fields are the methods and tools used. Searching on the internet, reading articles or speaking with other companies give a huge variety of options to choose. Each method or tool has its own properties and advantages, however, at the end, everyone has a different opinion. Each person will defend its own way to work.

This advantage of different resources can be a double-edged sword. At the moment to start a project, you can get lost over this huge world of options and get confused. My recommendation is to think twice which tool use and think about which resources you have. Spending time thinking which tool use can imply to reduce effort and time on the future. A bad selection of tools can consequence into future problems.

At CleverData our main tool used to develop data science projects is BigML. BigML is a pioneer system of machine learning as a service. Is a highly scalable, cloud based machine learning service that is easy to use, seamless to integrate and instantly actionable.

What makes BigML special is that any person, independently of the background, can use it. As Francisco J.Martin, Co-Founder and CEO of BigML said in an interview "*Es una herramienta que sirve para aprender de los datos de forma muy fácil. No hay que saber nada de data science para usar BigML. Es una cosa bastante mágica, el sistema encuentra patrones de forma automática. Nuestro objetivo es automatizar las tareas del machine learning y democratizarlo. BigML está a disposición de todo el mundo, basta con arrastrar un fichero de datos. De forma automática el sistema analiza los datos y crea un data set estructurado, después es capaz de encontrar los patrones y generar un formulario para jugar con las variables*" [14].

The service offers a wide range of different supervised and unsupervised algorithms. Moreover, it has resources that allows the user to create workflows in a easy way. The three main modes to use the service are:

- **Web interface**: This is the most common way to use it. Is a web user interface that is very intuitive. This is their main strong point. Allow the user to realize all the flux of steps in a very easy way.
- **Command Line Interface**: A command line tool call bigmler. Permit more flexibility than the web. I've never used it, I worked directly with the API.
- **API**: A RESTful API provided in many programing languages: Python, Java, Node.js, Clojure, Swift, Objective-C, C#, PHP.

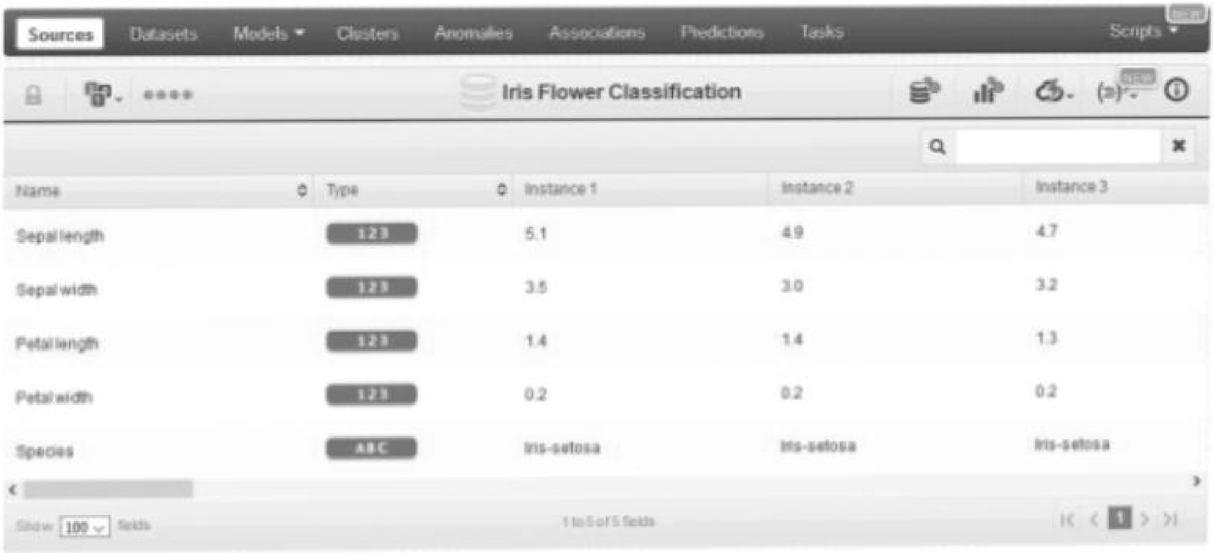
The service can be used in development mode or production mode. The first one is free, however, the drawback is the limitation of size tasks. The second one is a paid mode, there are different plans, each one of them with its own characteristics.

2.2.1 Supervised Learning

BigML offers a huge variety of resources very useful for the user. Following, will be summarized its main supervised learning resources. Remark that each resource has its own options and parameters that converts it in a powerful tool. However, due it will be endless for the reader, won't be described in detail each of its options. To go deep in details I recommend read the BigML documentation that is on the web.

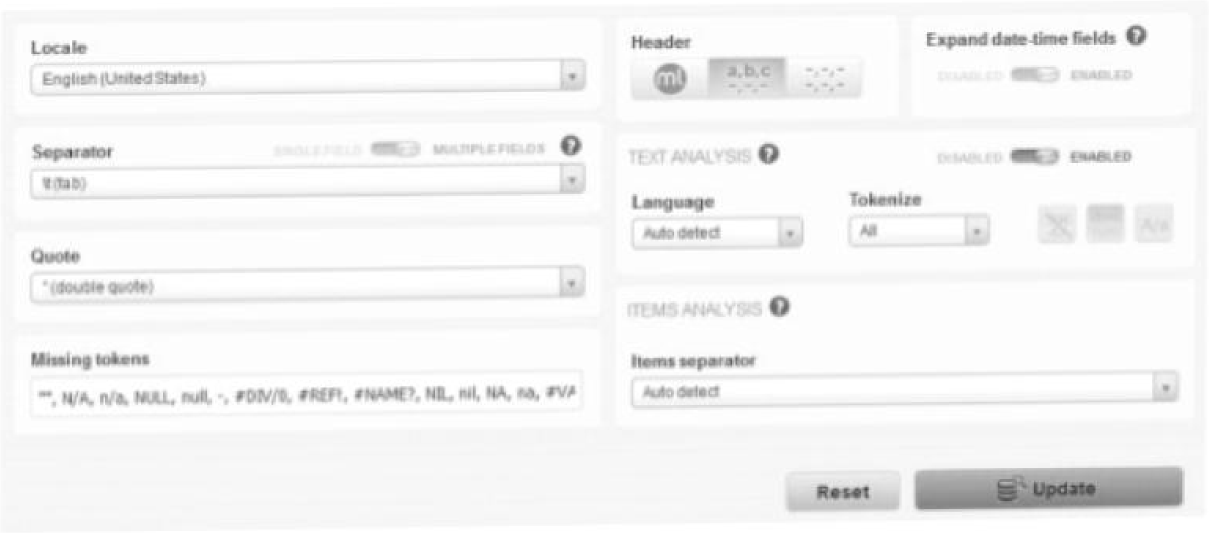
2.2.1.1 Sources

Sources are the raw data for the problem under study. BigML accept different formats file, however, the most common used is a CSV. BigML also accept as source remotes files by a specific URL or files from specific servers. Once the source is upload, there is a range of possibilities to configure it. Select the type of the features, the language of the source, missing values, how has to be treated text or item features are some of the options it has (Figures 2.2 and 2.3).



Name	Type	Instance 1	Instance 2	Instance 3
Sepal length	123	5.1	4.9	4.7
Sepal width	123	3.5	3.0	3.2
Petal length	123	1.4	1.4	1.3
Petal width	123	0.2	0.2	0.2
Species	ABC	Iris-setosa	Iris-setosa	Iris-setosa

Figure 2.2: Source data.



Locale: English (United States)

Header

Expand date-time fields: DISABLED ENABLED

Separator: \t (tab)

Quote: " (double quote)

Missing tokens: "", N/A, n/a, NULL, null, -, #DIV/0, #REF!, #NAME?, NIL, nil, NA, na, #VA

TEXT ANALYSIS: DISABLED ENABLED

Language: Auto detect

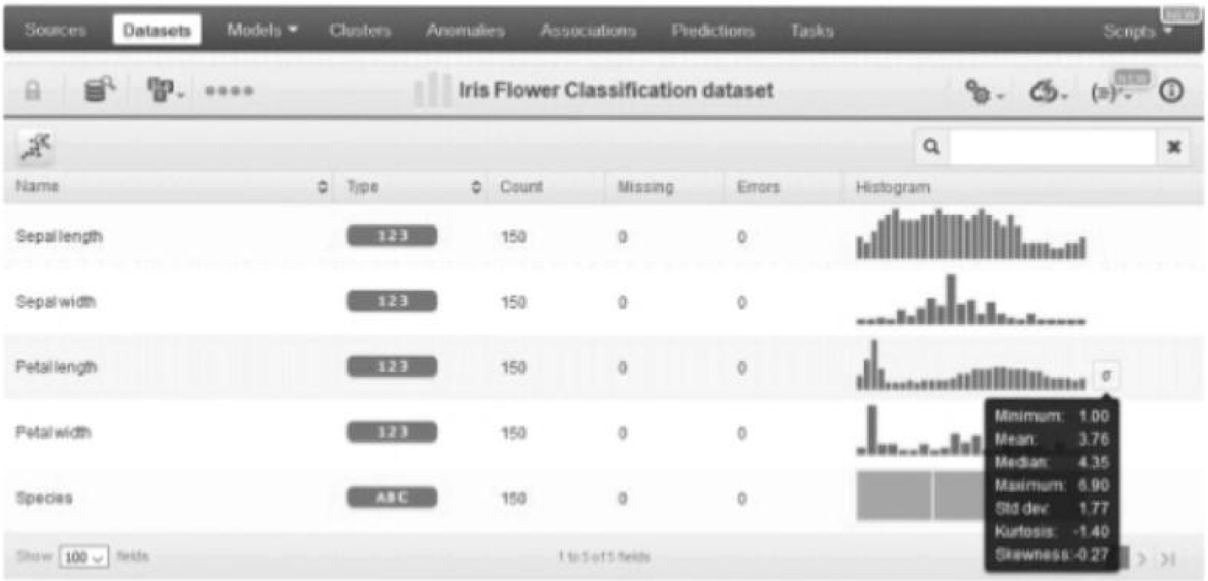
Tokenize: All

Items separator: Auto detect

Figure 2.3: Source configuration.

2.2.1.2 Datasets

Datasets are views of the data source that the user can use as the basis for building models. Datasets specify the target attribute (class in classification or output in regression). Each feature is summarized with a bar graph that permits its visualization (Figure 2.4). In addition, user can see some variables as mean, median, standard deviation over others that permits a first analysis of the features' distribution.

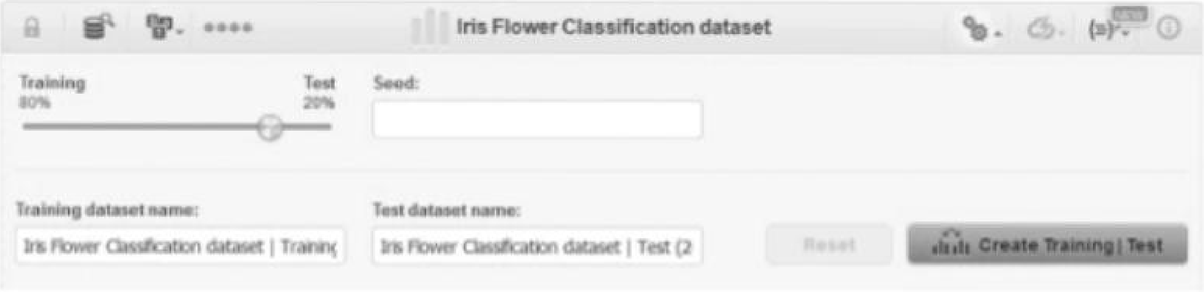


Name	Type	Count	Missing	Errors	Histogram
Sepal length	123	150	0	0	
Sepal width	123	150	0	0	
Petal length	123	150	0	0	
Petal width	123	150	0	0	
Species	ABC	150	0	0	

Minimum: 1.00
Mean: 3.76
Median: 4.35
Maximum: 6.90
Std dev: 1.77
Kurtosis: -1.40
Skewness: -0.27

Figure 2.4: Features distribution.

There is also a very common used training and test set split resource that separate an original dataset into a training and test dataset for a controlled evaluation of a models performance later. User can choose the proportion data of each set, however, the most common, is the typical 80-20 percent split (Figure 2.5)



Training 80%
Test 20%
Seed:
Training dataset name: Iris Flower Classification dataset | Trainin
Test dataset name: Iris Flower Classification dataset | Test (2

Figure 2.5: Split dataset.

2.2.1.3 Models

A Model is a decision trees created from a dataset. One of the best characteristics of BigML is the interactive interface it provides. User can see the confidence and support in the training data reflected in the model at each node and how the rules are build up, which is a clever and clean presentation of the model (Figure 2.6). BigML offers a sunburst view representation as well (Figure 2.7).

As in the different resources, models has its own properties that makes them flexible to the user demands. Som options as the balanced objective, or the number of leafs are examples of the variety parameters configuration.

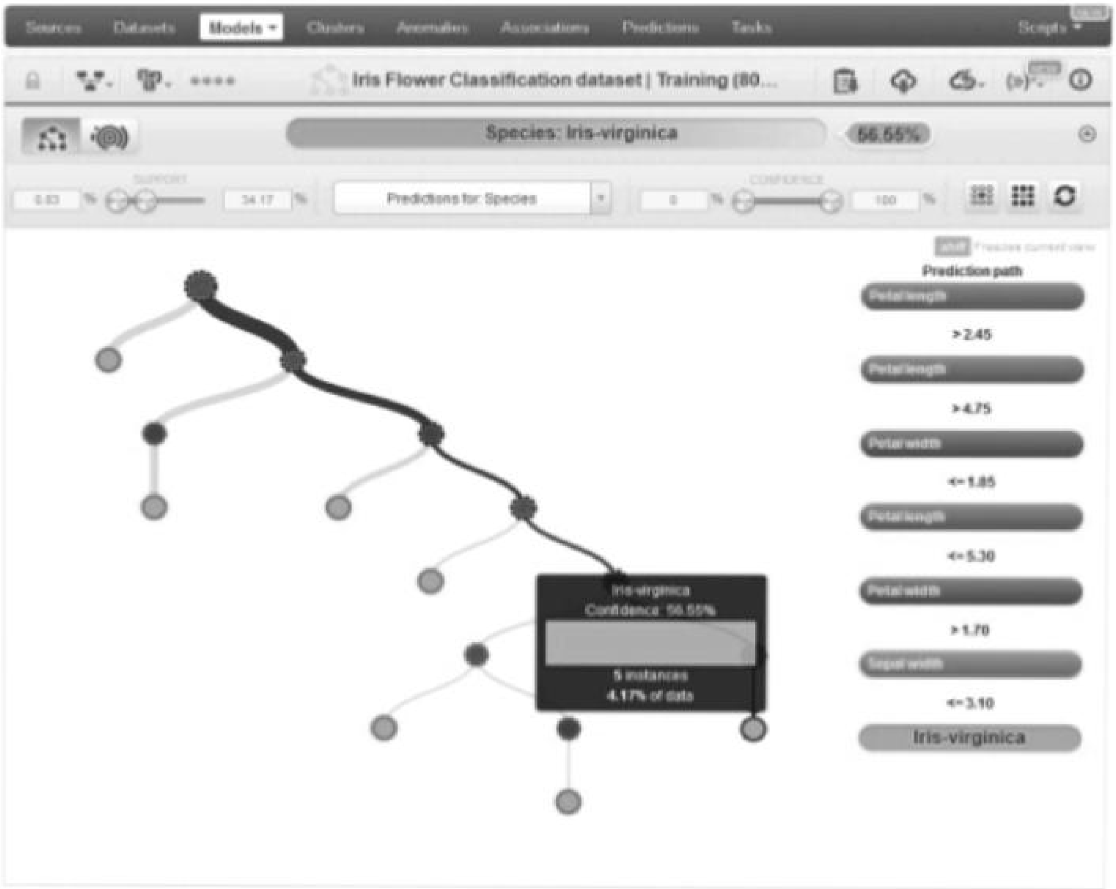
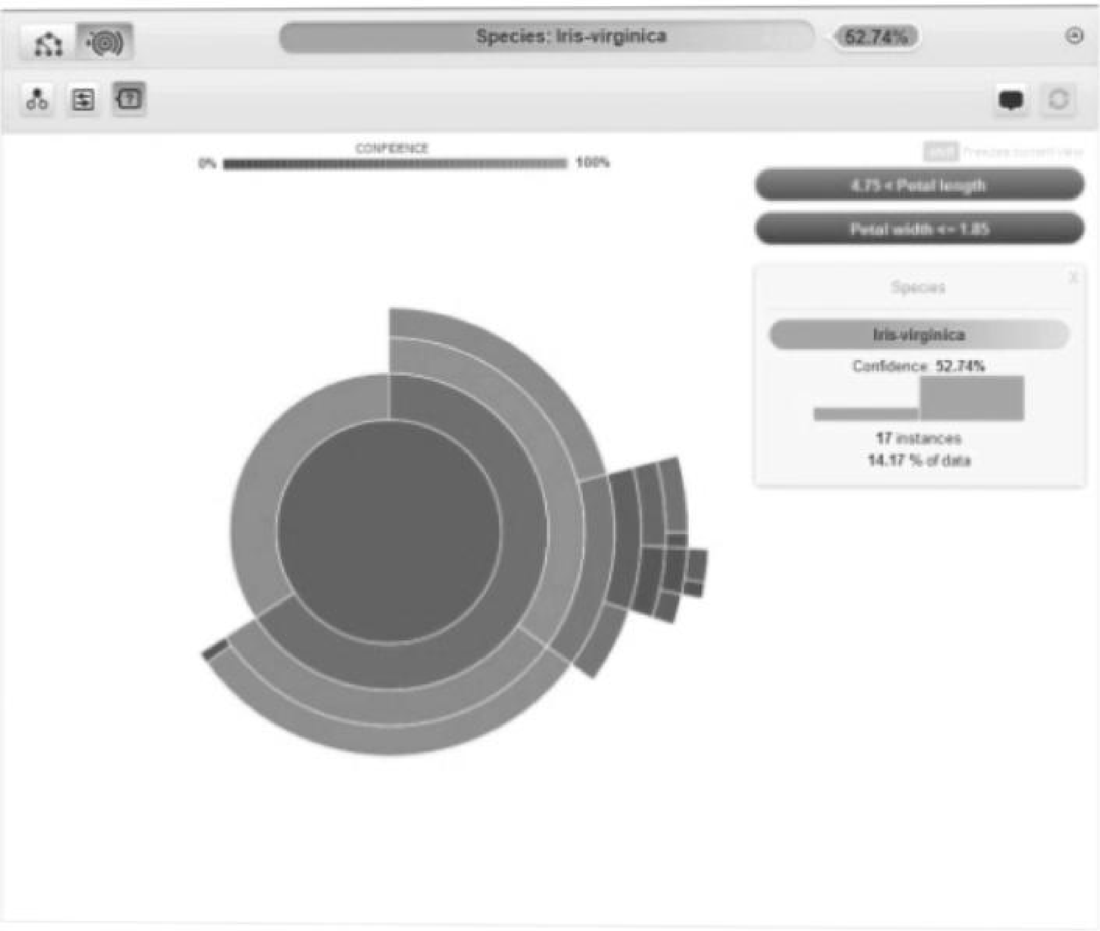


Figure 2.6: Model.



Species: Iris-virginica 52.74%

4.75 < Petal length

Petal width <= 1.85

Species

Iris-virginica

Confidence 52.74%

17 instances
14.17 % of data

Figure 2.7: Sunburst view.

2.2.1.4 Ensembles

An ensemble is a collection of models which work together to create a stronger model with better predictive (Figure 2.8). BigML currently provides two types of ensembles:

- **Bagging** (a.k.a. Bootstrap Aggregating): builds each model from a random subset of dataset. By default the samples are taken using a rate of 100% with replacement. While this is a simple strategy, it often outperforms more complex strategies.

- **Random Decision Forests**: similar to Bagging however, also chooses from among a random feature subset at each split.

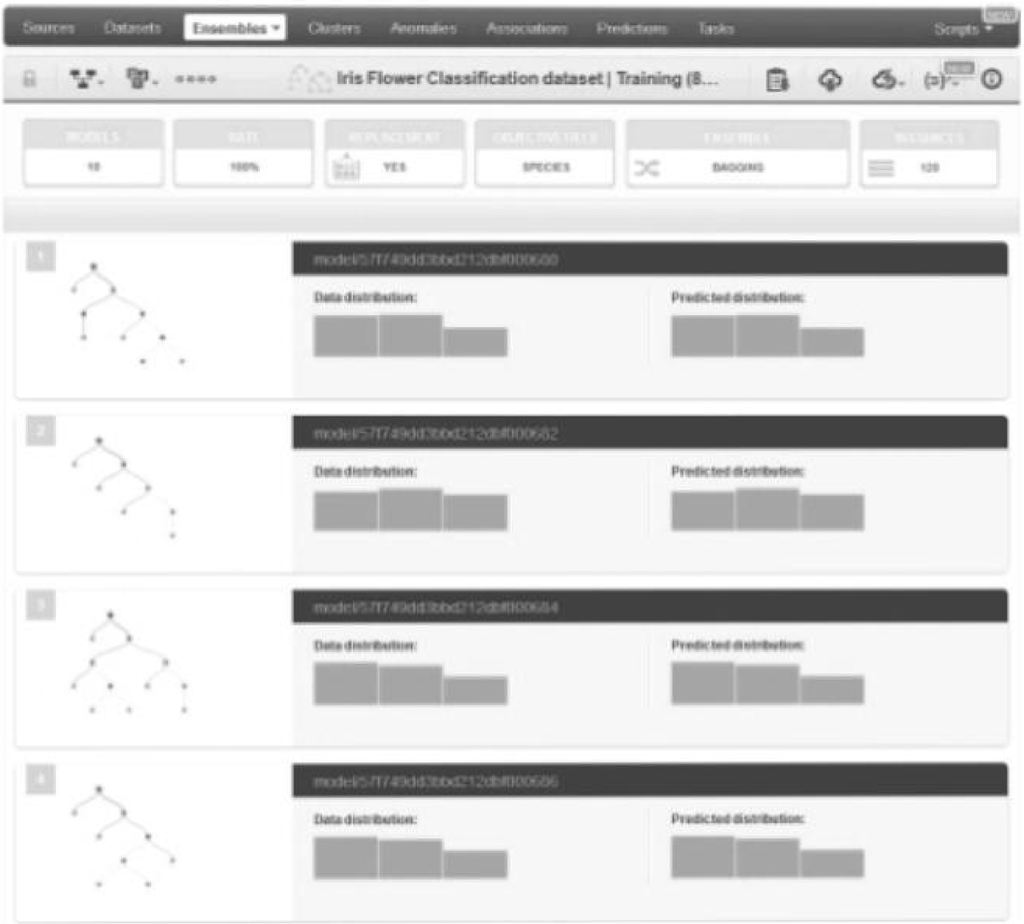


Figure 2.8: Ensemble.

Ensembles and models, once trained, have the option to visualize an ordered list with the field importance (Figure 2.9). This simple characteristic, that permits the user a first analysis, is even more useful for clients. As mentioned before, as important to understand the results, is the transmition of them to clients. Due most of time, clients wants brief and simple answers, this visualization is perfect.

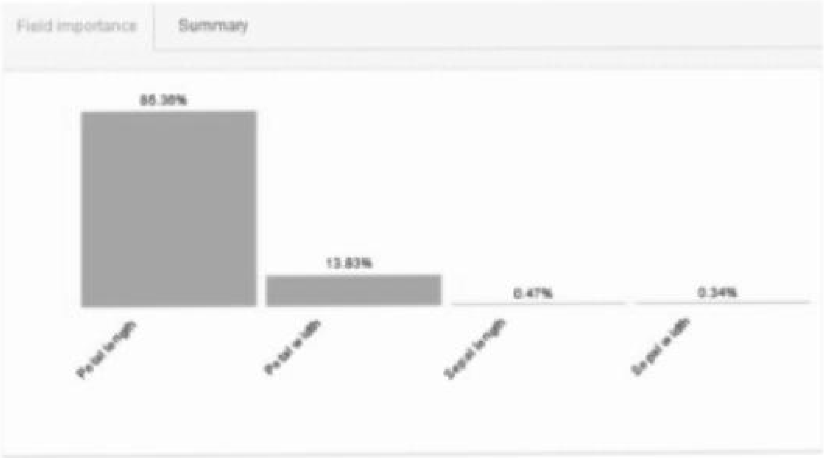
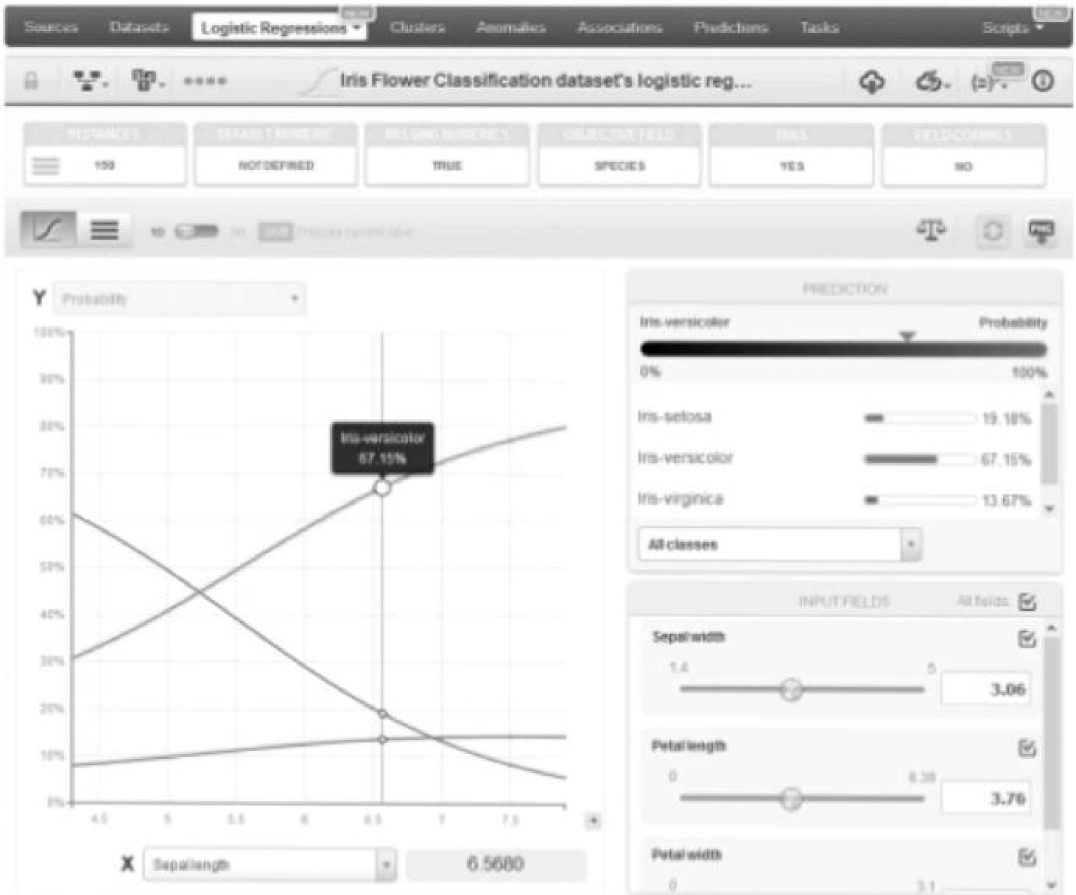


Figure 2.9: Field Importance.

2.2.1.5 Logistic Regressions

A logistic regression is a supervised Machine Learning method to solve classification problems. For each class of the objective field, the logistic regression computes a probability modeled as a logistic function value, whose argument is a linear combination of the field values (Figure 2.10).



INSTANCES	DEFAULT NUMERIC	MISSING NUMERICS	OBJECTIVE FIELD	BIAS	FIELD CODINGS
150	NOT DEFINED	TRUE	SPECIES	YES	NO

X: Sepal length 6.5680

Iris-setosa	19.18%
Iris-versicolor	67.15%
Iris-virginica	13.67%

Sepal width 3.06

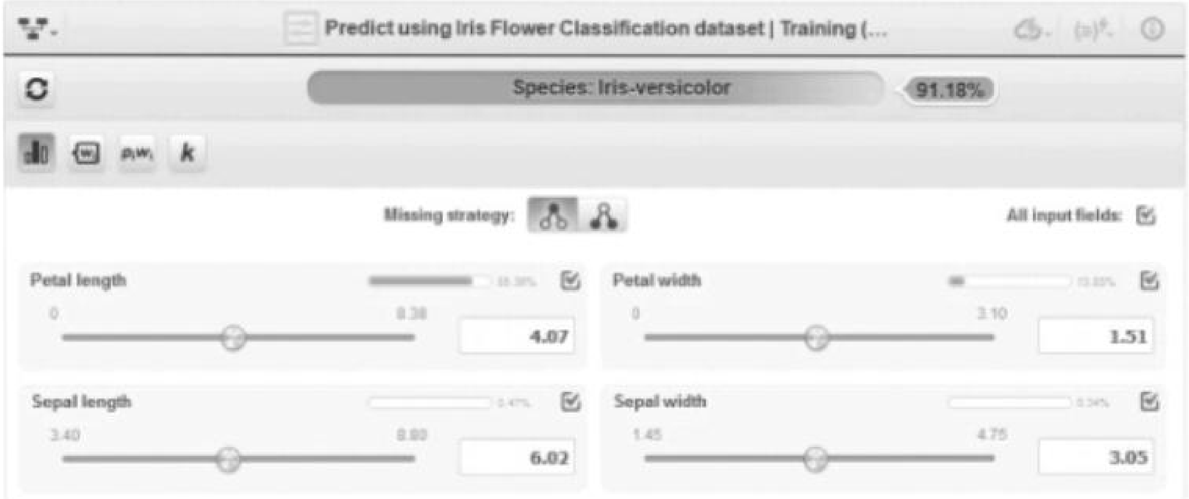
Petal length 3.76

Petal width

Figure 2.10: Logistic Regression.

2.2.1.6 Predictions

BigML permits predictions for single instances or for many instances in a batch (Figure 2.11). Each prediction has a categorical or numerical output depending if it is a classification or regression problem respectively. In addition, for each prediction there is its confidence or expected error respectively.



Predict using Iris Flower Classification dataset | Training (...

Species: Iris-versicolor 91.18%

Missing strategy:

All input fields: ☑

Petal length ☑ 0 8.38 4.07

Petal width ☑ 0 3.10 1.51

Sepal length ☑ 3.40 8.80 6.02

Sepal width ☑ 1.45 4.75 3.05

Figure 2.11: Single Prediction.

2.2.1.7 Evaluations

BigML provide an easy way to measure and compare the performance of classification and regression models. The main purpose of evaluations is twofold:

- First, obtaining an estimation of the model's performance in production (i.e., making predictions for new instances the model has never seen before).

- Second, providing a framework to compare models built using different configurations or different algorithms to help identify the models with best predictive performance.

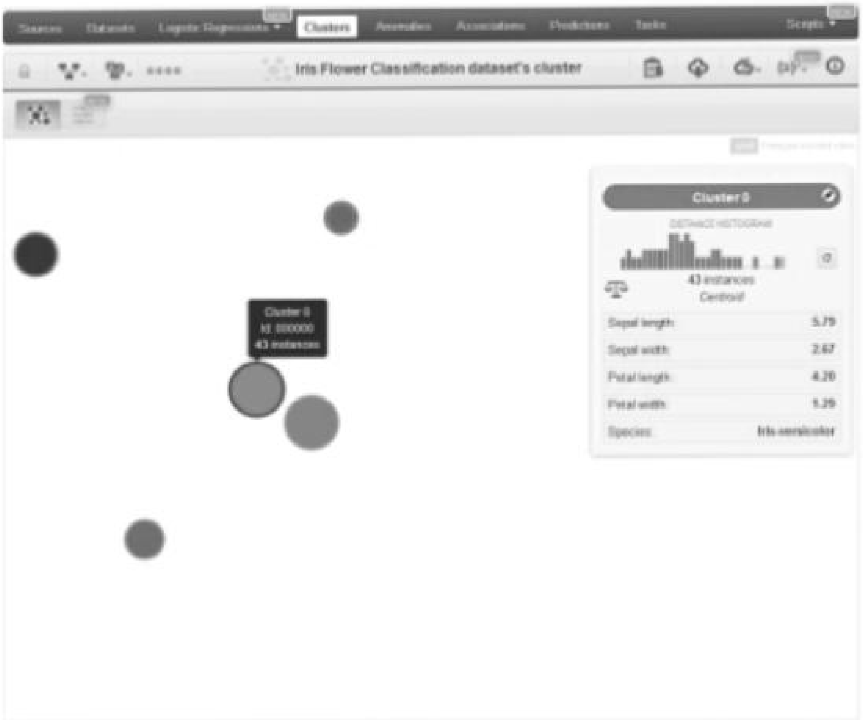
The basic idea behind evaluations is to take some test data different from the one used to train a model and create a prediction for every instance. Then compare the actual objective field values of the instances in the test data against the predictions and compute several performance measures based on the correct results as well as the errors made by the model.

2.2.2 Unsupervised Learning

BigML offers a variety of unsupervised learning resources as well. Due the project was developed with unsupervised learning resources, they will be more detailed than the supervised learning resources. However this resources will be deeply explained in the next section, here are just introduced.

2.2.2.1 Clusters

BigML Clusters provide powerful visualizations of the results of clustering data instances, which gives insight into their internal structure. In addition their visual representations, clusters also provide a textual summary view of the most essential information about them (Figure 2.12). Clusters uses proprietary unsupervised learning algorithms to group together the instances that are closer together according to a distance measure, computed using the values of the fields as input.



Sepal length	5.79
Sepal width	2.67
Petal length	4.20
Petal width	1.29
Species	Iris-versicolor

Figure 2.12: Clusters.

BigML Clusters can be built using two different unsupervised learning algorithms:

- **K-means**: the number of centroids need to be specified in advance.

- **G-means**: learns the number of different clusters by iteratively taking existing cluster groups and testing whether the cluster's neighborhood appears Gaussian in its distribution.

Both algorithms support a number of configuration options, such as scales and weights, over others.

2.2.2.2 Anomalies

Identify instances within a dataset that do not conform to a regular pattern (Figure 2.13). BigML's anomaly detector is an optimized implementation of the Isolation Forest algorithm, a highly scalable method that can efficiently deal with high-dimensional datasets.

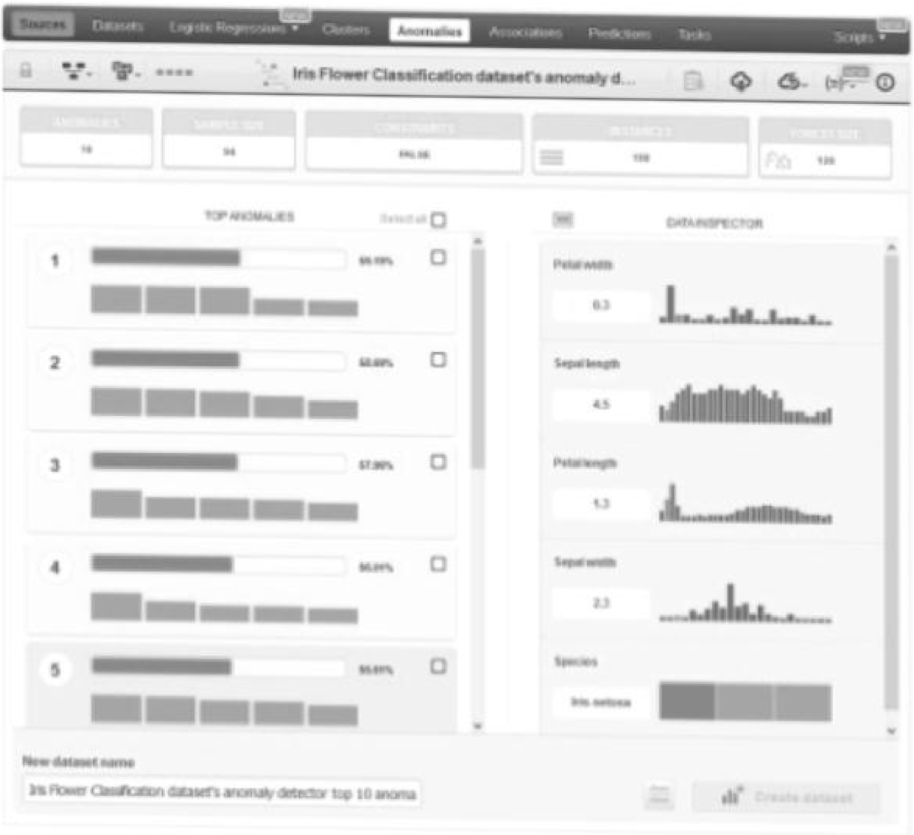
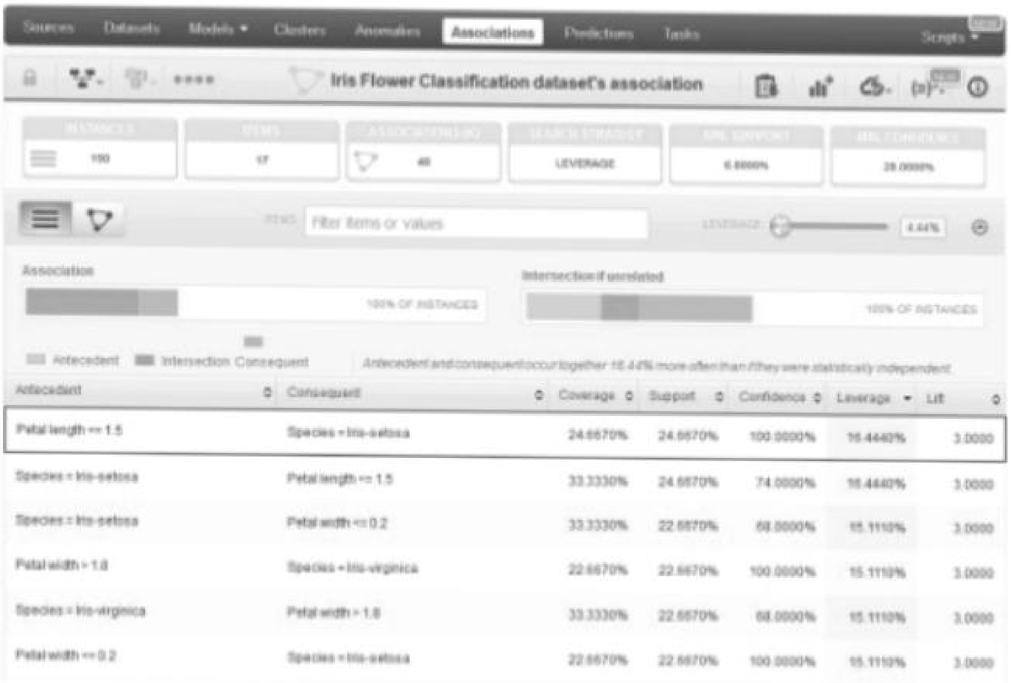


Figure 2.13: Anomaly Detection.

2.2.2.3 Associations

Find meaningful relationships among fields and their values in high-dimensional datasets (Figure 2.14).



Antecedent and consequent occur together 16.44% more often than if they were statistically independent

Antecedent	Consequent	Coverage	Support	Confidence	Leverage	Lift
Petal length <= 1.5	Species = Iris-setosa	24.6670%	24.6670%	100.0000%	16.4440%	3.0000
Species = Iris-setosa	Petal length <= 1.5	33.3330%	24.6670%	74.0000%	16.4440%	3.0000
Species = Iris-setosa	Petal width <= 0.2	33.3330%	22.6670%	68.0000%	15.1110%	3.0000
Petal width > 1.8	Species = Iris-virginica	22.6670%	22.6670%	100.0000%	15.1110%	3.0000
Species = Iris-virginica	Petal width > 1.8	33.3330%	22.6670%	68.0000%	15.1110%	3.0000
Petal width <= 0.2	Species = Iris-setosa	22.6670%	22.6670%	100.0000%	15.1110%	3.0000

Figure 2.14: Associations rules.

3 Design and Application of Market Basket Analysis Methodology

3.1 Project Methodology

Through previous chapters, most aspect of the project were already described. However, a brief summary of the project process will be described following to have a general view of the procedure. In addition, there are some details that has to be mentioned about the design of the project that affected its procedure.

Our objective in his project was the analysis of customers purchases and its behaviour. To do it, the project was divided in two steps. The first one, was a store clustering. Group stores based on its behaviours. The second one, was the analysis of associations rules of its items for each cluster.

The first step we realized was the problem definition. Which thing the client wanted to achieve with this project. All this step was described in the "*Introduction*" chapter.

The second step was the obtention of data. Companies usually have its data in data warehouse [21] or databases [22] and the extraction of it is a difficult task that requires a huge work. In order to design a good data science project, automatic workflows of that extraction has to be done. That's because, as it was mentioned previously, machine learning models have to be retrained periodically. However, for the first model of associations rules, the client was not interested in this automatic workflow. Due that, all the data used in this project was given in a *csv* format.

Once we obtained the data we had to clean it. Throughout the project we removed data that was incorrect or invalid. Is common that data have mistakes, is impossible to have everything in order in a company, that's why a data cleaning task is performed. However, there are cases where some instances have to be removed although are correct because they are considered anomalies. Anomaly detection [23] is the identification of observations which do not conform to an expected pattern or other items in a dataset. This concept has not to be confused with data cleaning, due the cleaning data process search for invalid records. Anomalies detection just looks for records that not form part of a pattern. Depending the objective of the project, this anomalies can be noise or be exactly what you are looking for. For instance, in fraud detection problems, those instance that don't conform a regular pattern are possible fraudulent transactions [24].

UNIDADES	Numerical
CANTIDAD	Numerical
IMPORTE_TICKET	Numerical

Figure 3.2: Ticket dataset features.

3.3.2 Articulos dataset

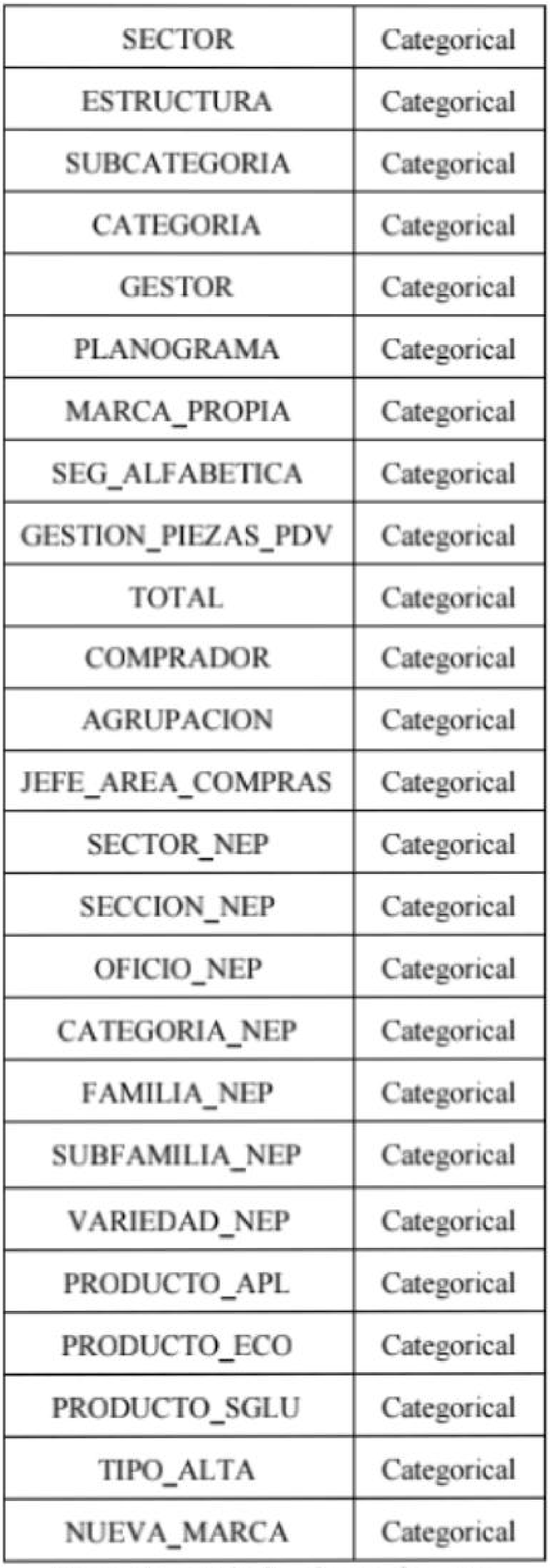
- Number of Instances: 60,587
- Number of Attributes: 74
- Missing Values? Yes
- Size of Dataset: 65.5 MB

The second dataset used in the project was the stock of items our client has. This dataset contains information of each item like the family group it belongs, if it is ecologic, if it has gluten and so on.

We used this dataset basically for all the feature engineering and analysing of items per level. As we commented in previously, the client is interested in associations rules at family level. Due that, the features that will be created during the clustering step will try to capt as better as possible the behaviour of the shops related to that.

To avoid overextend the summary table (Figure 3.3) are merged some features into one. For instance, most of the features in the database are repeated twice, one has the code and the other the description. Other features are repeated in catalan and spanish, that for us is not needed. Due this overlap of information, here are just listed a list of features that represent the concept of the totally features.

Feature	Type
ARTICULO	Categorical
DEPARTAMENTO	Categorical
SECCION_VENTA	Categorical
VARIEDAD	Categorical
SUBFAMILIA	Categorical
FAMILIA	Categorical
SECCION	Categorical



SECTOR	Categorical
ESTRUCTURA	Categorical
SUBCATEGORIA	Categorical
CATEGORIA	Categorical
GESTOR	Categorical
PLANOGRAMA	Categorical
MARCA_PROPIA	Categorical
SEG_ALFABETICA	Categorical
GESTION_PIEZAS_PDV	Categorical
TOTAL	Categorical
COMPRADOR	Categorical
AGRUPACION	Categorical
JEFE_AREA_COMPRAS	Categorical
SECTOR_NEP	Categorical
SECCION_NEP	Categorical
OFICIO_NEP	Categorical
CATEGORIA_NEP	Categorical
FAMILIA_NEP	Categorical
SUBFAMILIA_NEP	Categorical
VARIEDAD_NEP	Categorical
PRODUCTO_APL	Categorical
PRODUCTO_ECO	Categorical
PRODUCTO_SGLU	Categorical
TIPO_ALTA	Categorical
NUEVA_MARCA	Categorical

Figure 3.3: Articulos dataset features.

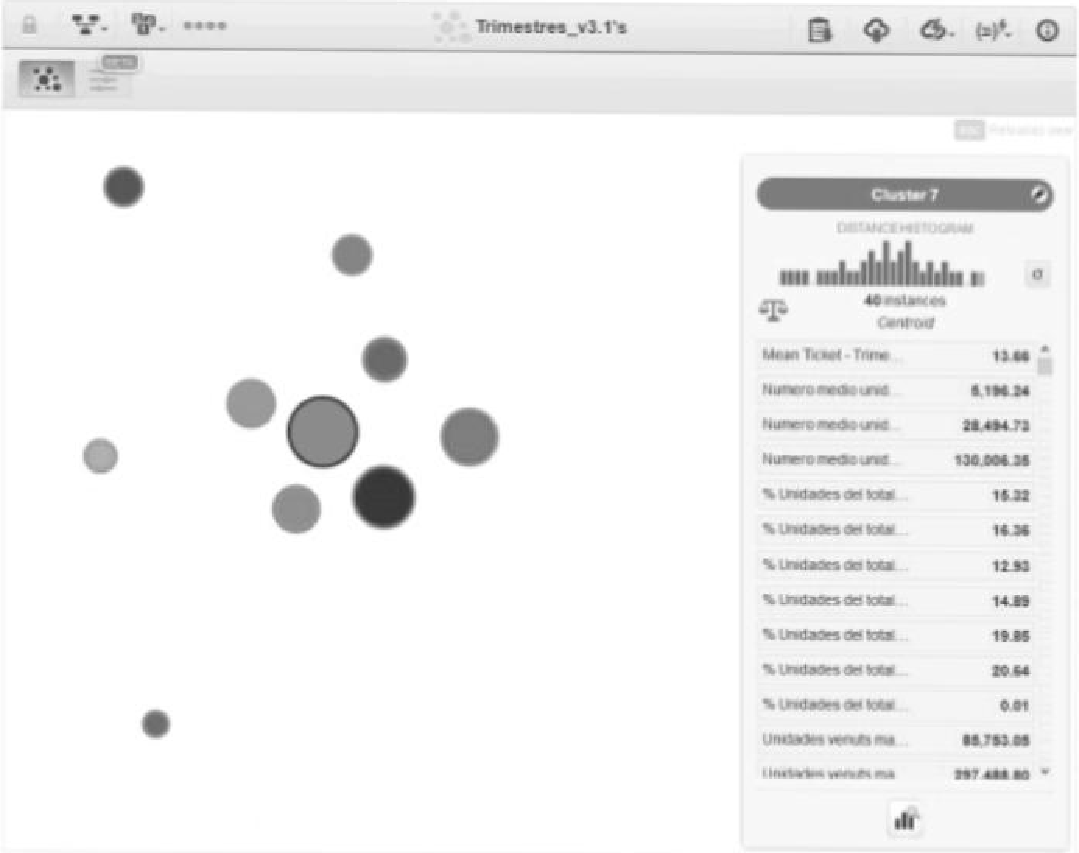
3.4.1.3 Version 3

In this version were added two new concepts in the dataset, the revenue of each section and the market penetration rate. Both concepts were important in order to obtain the final cluster and represented an important point of view from the client side.

The first concept is similar to the ones created in the previous version, when we started to add information about the revenue of the shop. This features are the revenue obtained for each section along a specific trimester. With this, we wanted to find similars shops based on the revenue of each section.

The second concept added in this version was the penetration rate. Penetration rate [28] and market share rate go hand in hand as metrics descriptions in retail. This features represents which presence has each section on the shop's tickets. For instance, if a shop has 4 different tickets and on 3 of them there are at least one item of an specific section, the penetration rate of that section will be 75%.

The total number of features used is 319. The next figure (Figure 3.5) was the resulting clusters.



Mean Ticket - Trime...	13.66
Numero medio unid...	5,196.24
Numero medio unid...	28,494.73
Numero medio unid...	130,006.35
% Unidades del total...	15.22
% Unidades del total...	16.36
% Unidades del total...	12.93
% Unidades del total...	14.89
% Unidades del total...	19.85
% Unidades del total...	20.64
% Unidades del total...	0.01
Unidades venuts ma...	85,753.05
Unidades venuts ma...	297,488.80

Figure 3.5: Clusters version 3.

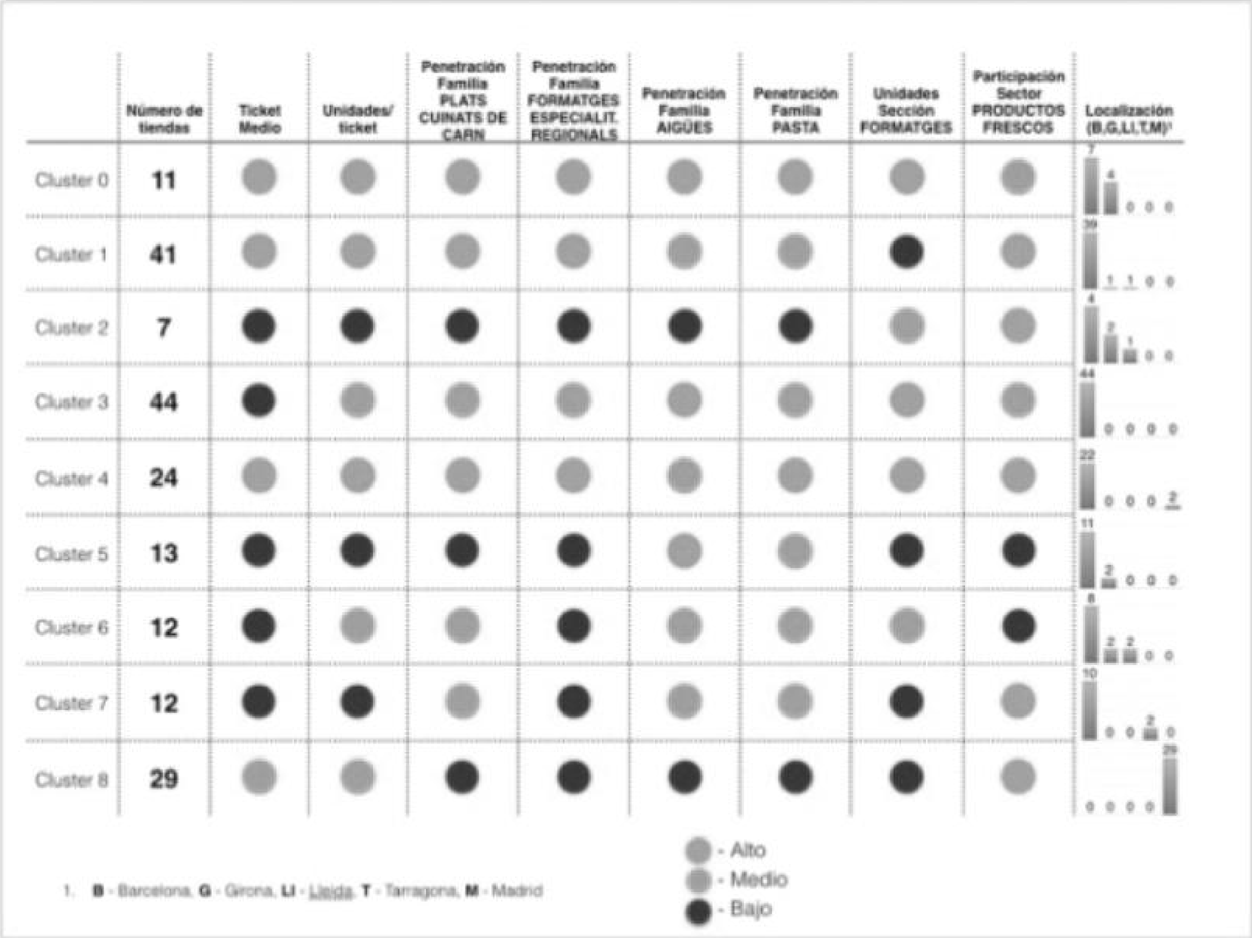
Figure 3.11: Associations rules with Lift strategy.

3.4.3 Results Delivery

With all the differents steps of the project done, our last work was deliver those results to the client. An important moment in all projects we perform in CleverData is when results have to be transmitted. It is true that the result by itself are important, however, the transmission of them is even more important. This step can be seen as trivial but, in real world projects, is absolutely the opposite. Most of times, when a project is done, the results has to be explained to people that is not from the same domain or don't have the same background. Due this, the explanation of results is a hard work and needs time. After all, the client is the one that has to apply the results.

This concept is the one that leads us to create two brief tables summary that described the clusters and the similitude between them. Before start with the analysis of associations rules, when the results of clustering were delivered to the client, this asked if it was possible to create something that makes easy distinguish which characteristics had each cluster. Basically they wanted something that tells them what had each cluster in particular and why the shops were putted together.

For people from machine learning world, is easy to explain that what defines a cluster is the sum of the different features, however, for person from marketing that's another story, they need something that tells them what has a cluster in particular. So, in order to have this little help that described what had each cluster in particular we created the brief tables summary (Figures 3.12 and 3.13).

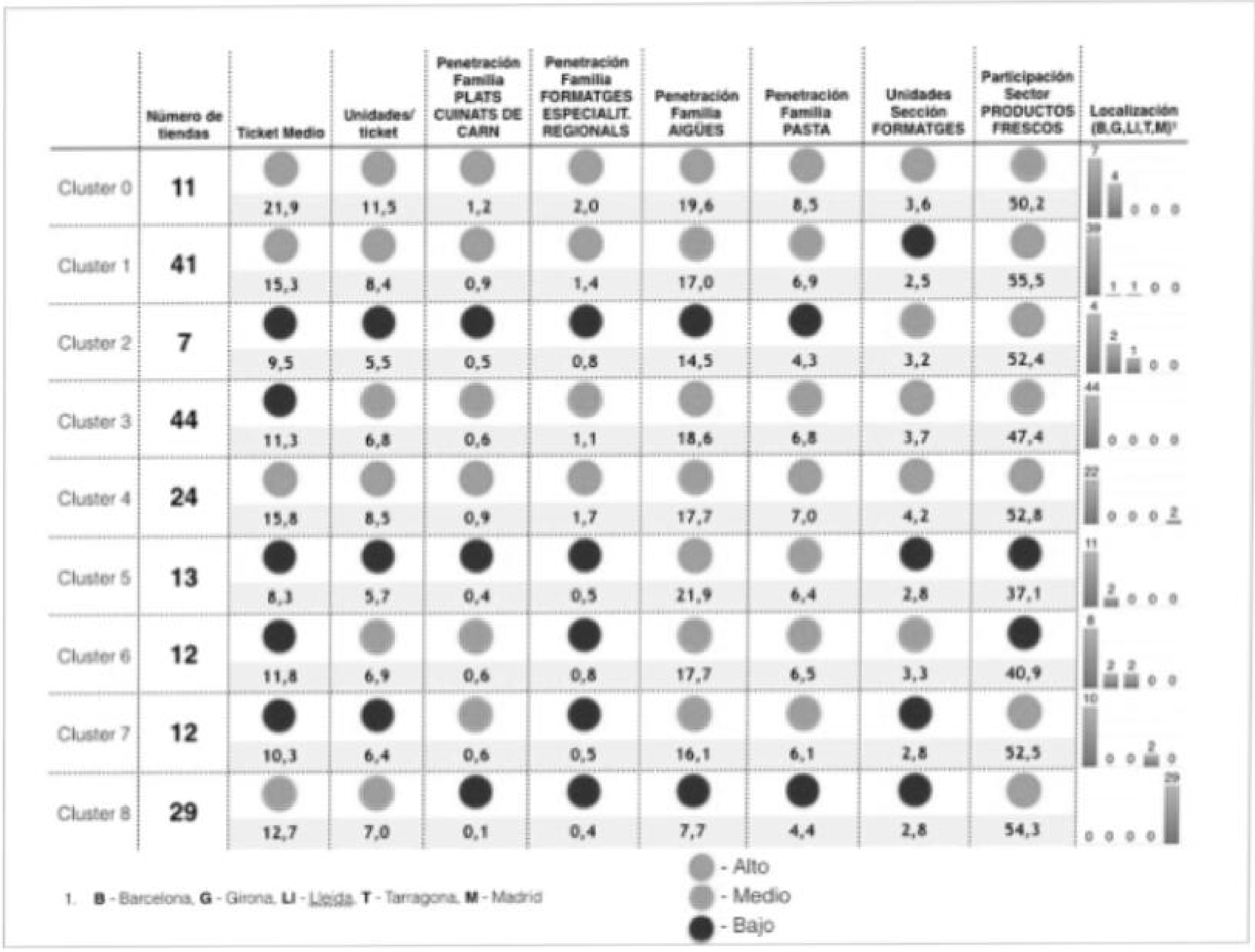


	Número de tiendas	Ticket Medio	Unidades/ ticket	Penetración Familia PLATS CUINATS DE CARN	Penetración Familia FORMATGES ESPECIALIT. REGIONALS	Penetración Familia AIGÜES	Penetración Familia PASTA	Unidades Sección FORMATGES	Participación Sector PRODUCTOS FRESCOS	Localización (B,G,Ll,T,M)[1]
Cluster 0	11	○	○	○	○	○	○	○	○	7 4 0 0 0
Cluster 1	41	○	○	○	○	○	○	●	○	39 1 1 0 0
Cluster 2	7	●	●	●	●	●	●	○	○	4 2 1 0 0
Cluster 3	44	●	○	○	○	○	○	○	○	44 0 0 0 0
Cluster 4	24	○	○	○	○	○	○	○	○	22 0 0 0 2
Cluster 5	13	●	●	●	●	○	○	●	●	11 2 0 0 0
Cluster 6	12	●	○	○	●	○	○	○	●	8 2 2 0 0
Cluster 7	12	●	●	○	●	○	○	●	○	10 0 0 2 0
Cluster 8	29	○	○	●	●	●	●	●	○	0 0 0 0 29

○ - Alto
○ - Medio
● - Bajo

[1] **B** - Barcelona, **G** - Girona, **Ll** - Lleida, **T** - Tarragona, **M** - Madrid

Figure 3.12: Table summary.



	Número de tiendas	Ticket Medio	Unidades/ ticket	Penetración Familia PLATS CUINATS DE CARN	Penetración Familia FORMATGES ESPECIALIT. REGIONALS	Penetración Familia AIGÜES	Penetración Familia PASTA	Unidades Sección FORMATGES	Participación Sector PRODUCTOS FRESCOS	Localización (B,G,LL,T,M)[1]
Cluster 0	11	21,9	11,5	1,2	2,0	19,6	8,5	3,6	50,2	7 4 0 0 0
Cluster 1	41	15,3	8,4	0,9	1,4	17,0	6,9	2,5	55,5	39 1 1 0 0
Cluster 2	7	9,5	5,5	0,5	0,8	14,5	4,3	3,2	52,4	4 2 1 0 0
Cluster 3	44	11,3	6,8	0,6	1,1	18,6	6,8	3,7	47,4	44 0 0 0 0
Cluster 4	24	15,8	8,5	0,9	1,7	17,7	7,0	4,2	52,8	22 0 0 0 2
Cluster 5	13	8,3	5,7	0,4	0,5	21,9	6,4	2,8	37,1	11 2 0 0 0
Cluster 6	12	11,8	6,9	0,6	0,8	17,7	6,5	3,3	40,9	8 2 2 0 0
Cluster 7	12	10,3	6,4	0,6	0,5	16,1	6,1	2,8	52,5	10 0 0 2 0
Cluster 8	29	12,7	7,0	0,1	0,4	7,7	4,4	2,8	54,3	0 0 0 0 29

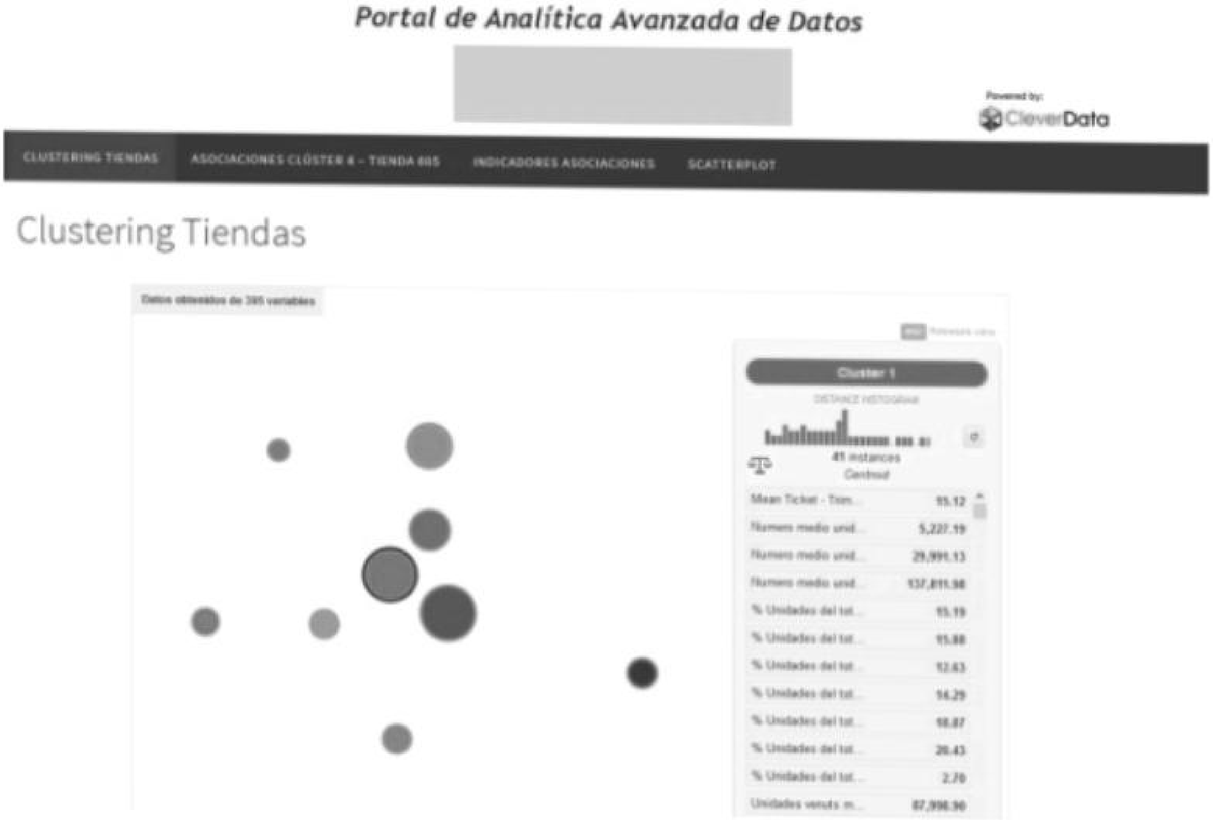
● - Alto
● - Medio
● - Bajo

1. **B** - Barcelona, **G** - Girona, **Ll** - Lleida, **T** - Tarragona, **M** - Madrid

Figure 3.13: Table summary with values.

This tables were created using some chosen variables. Moreover we added a color to each variable to distinguish if the value was high, medium or low. Both tables are basically the same, the unique difference is that in the second one the value of the corresponding characteristic is added.

After this summary tables, we created a web (Figure 3.14), where we upload the results of the project. With that, the client was able to analyze the results always he needed. This web was divided in 4 pages, each of them described a part of the project. Moreover, we delivered 3 different files. The first one was an excel with information of the different clusters with the shops' id it was conformed. The second and third one, was the list of associations rules using the lift and leverage strategies.



Portal de Analítica Avanzada de Datos

[redacted]

Clustering Tiendas

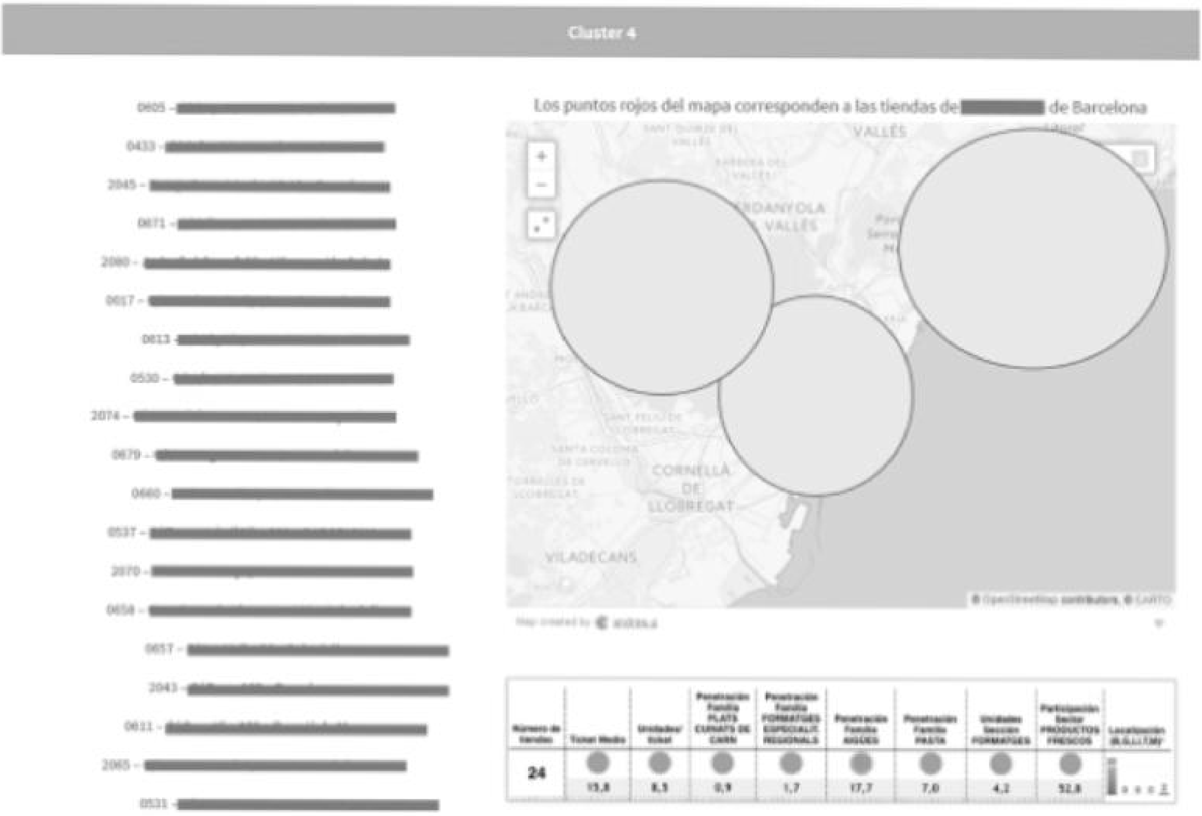
Datos obtenidos de 385 variables

41 instances
Centroid

Mean Ticket - Trim...	55.12
Numero medio unid...	5,227.19
Numero medio unid...	29,991.13
Numero medio unid...	137,811.98
% Unidades del tot...	55.19
% Unidades del tot...	55.88
% Unidades del tot...	52.63
% Unidades del tot...	54.29
% Unidades del tot...	58.87
% Unidades del tot...	20.43
% Unidades del tot...	2.78
Unidades venuts m...	87,998.90

Figure 3.14: CleverData web to visualize the results.

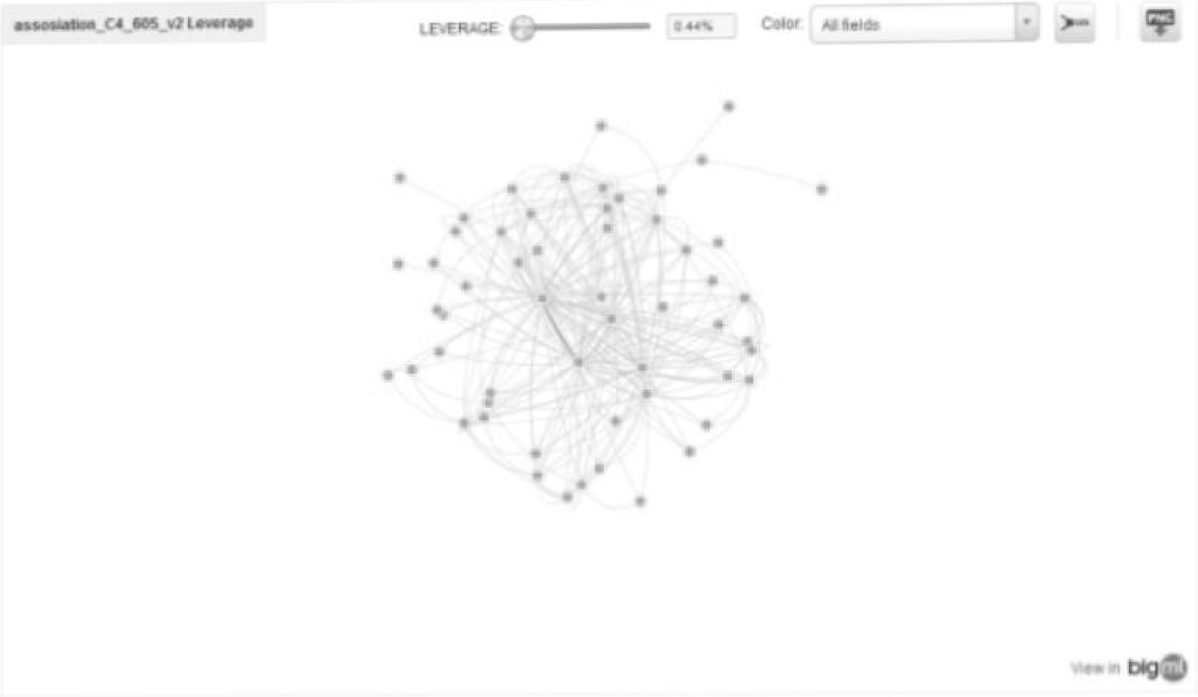
The first page of the web is related to the shops that conform a cluster and its characteristics. In addition, for each cluster, we created a map where the position of each shop was plotted. In addition, were plotted the shops from its main competitor to analyze the distance between shops. This maps were created using the Carto [29] tool. The figure 3.15 is an example of this structure corresponding to the the cluster 4.



Número de tiendas	Ticket Medio	Unidades/ ticket	Penetración Familia PLATS CUINATS DE CARN	Penetración Familia FORMATGES ESPECIALIT. REGIONALS	Penetración Familia ANISES	Penetración Familia PASTA	Unidades Sección FORMATGES	Participación Sector PRODUCTOS FRESCOS	Localización (B,G,L,I,T,M)
24	● 15,8	● 8,5	● 0,9	● 1,7	● 17,7	● 7,0	● 4,2	● 52,8	

Figure 3.15: Cluster 4 Web.

On the second page of the web, the client could saw the results according the associations rules. Both associations rules based on lift and leverage strategies were plotted. Using the BigML API, we were able to create a widget that communicated with our BigML account and plotted the corresponding associations, with that, the client was able to analyze the results in a dynamic way. Both associations are the ones of the figures in the previous section of this memory. In addition, we added two relationship diagrams for the leverage and lift rules (Figures 3.16 and 3.17).

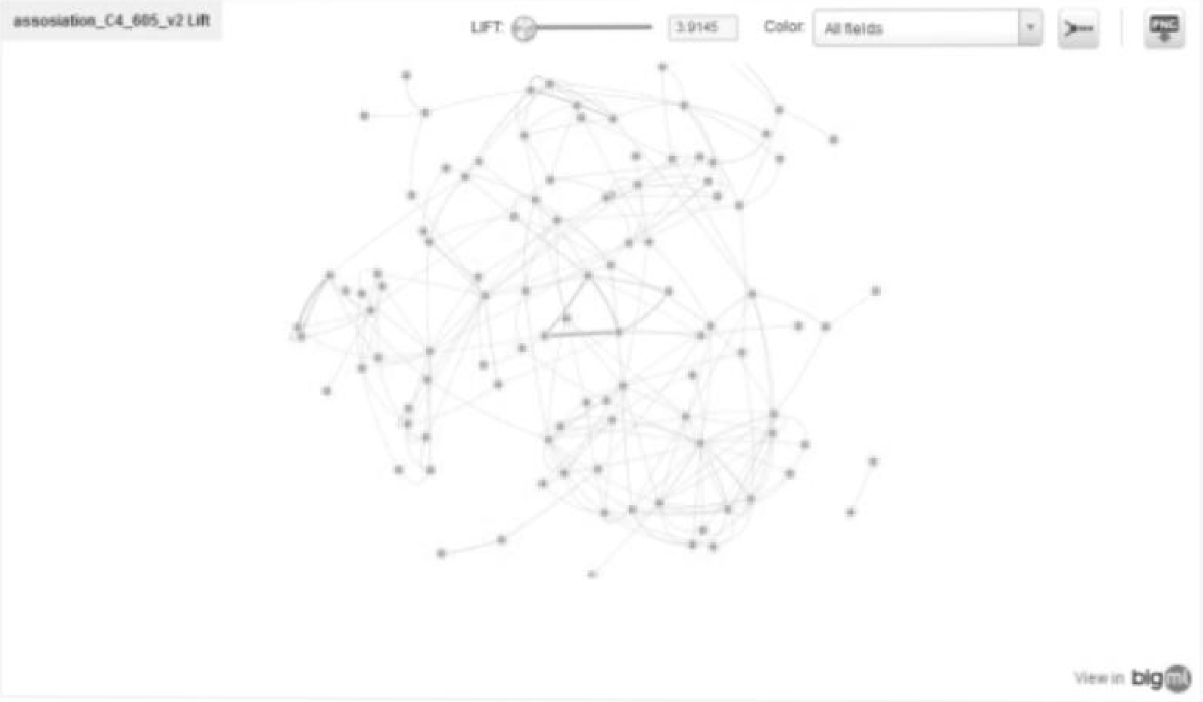


assosiation_C4_605_v2 Leverage

LEVERAGE: 0.44%

Color: All fields

Figure 3.16: Leverage Diagram.



assosiation_C4_605_v2 Lift

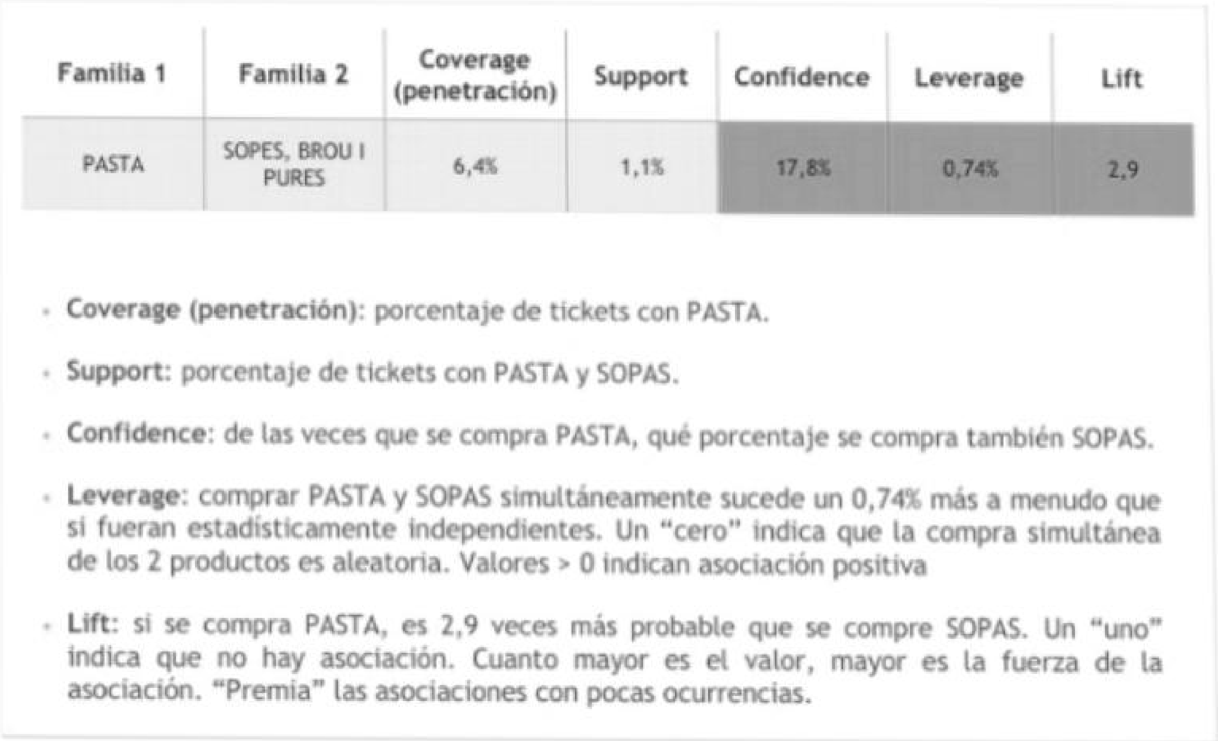
LIFT: 3.9145

Color: All fields

Figure 3.17: Lift Diagram.

To decide which associations were interesting we let the client choose between them. One interesting point with the projects we realize, is that once the results are obtained and analyzed, the corresponding action has to be taken using the knowledge of the client. At the end, the one who knows better the company is the client itself, and he has to be the one who decide what to do. Our results help supporting information, however, don't tells which is the action to be taken.

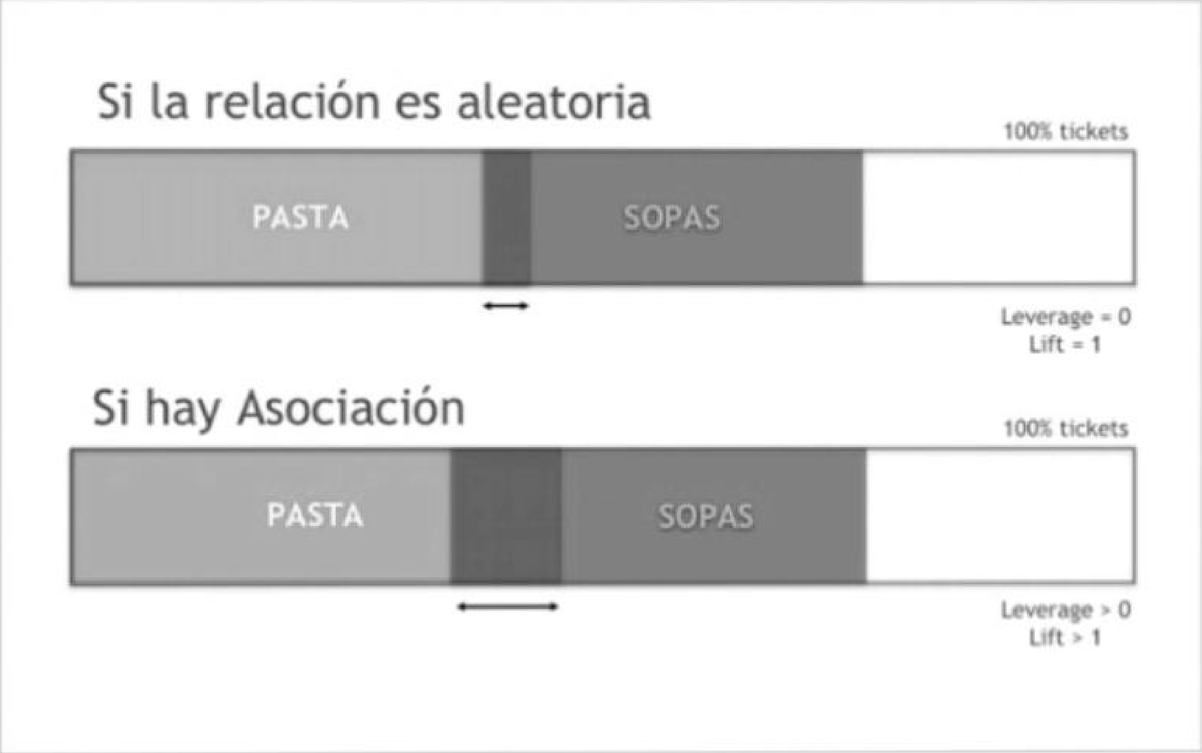
The third page was just a reminder of which metrics have the association rules and how can they be interpreted (Figures 3.18 and 3.19).



Familia 1	Familia 2	Coverage (penetración)	Support	Confidence	Leverage	Lift
PASTA	SOPES, BROU I PURES	6,4%	1,1%	17,8%	0,74%	2,9

- **Coverage (penetración):** porcentaje de tickets con PASTA.
- **Support:** porcentaje de tickets con PASTA y SOPAS.
- **Confidence:** de las veces que se compra PASTA, qué porcentaje se compra también SOPAS.
- **Leverage:** comprar PASTA y SOPAS simultáneamente sucede un 0,74% más a menudo que si fueran estadísticamente independientes. Un "cero" indica que la compra simultánea de los 2 productos es aleatoria. Valores > 0 indican asociación positiva
- **Lift:** si se compra PASTA, es 2,9 veces más probable que se compre SOPAS. Un "uno" indica que no hay asociación. Cuanto mayor es el valor, mayor es la fuerza de la asociación. "Premia" las asociaciones con pocas ocurrencias.

Figure 3.18: Metrics.



Si la relación es aleatoria

Si hay Asociación

Figure 3.19: Description of rules.

The last page consisted in a dynamic scatterplot. As we did with the associations, this widget communicated with our BigML account, in concrete, the scatterplot tool that has BigML. With this, the client was able to visualize different variables and how they were correlated. In the scatterplot each point is a shop and each color a cluster. Some examples of scatterplots that the client could visualize are the following: ticket mean price (Figure 3.20), the region (Figure 3.21) and mean units per ticket (Figure 3.22)

With the files delivered and the web constructed, the project was considered concluded.

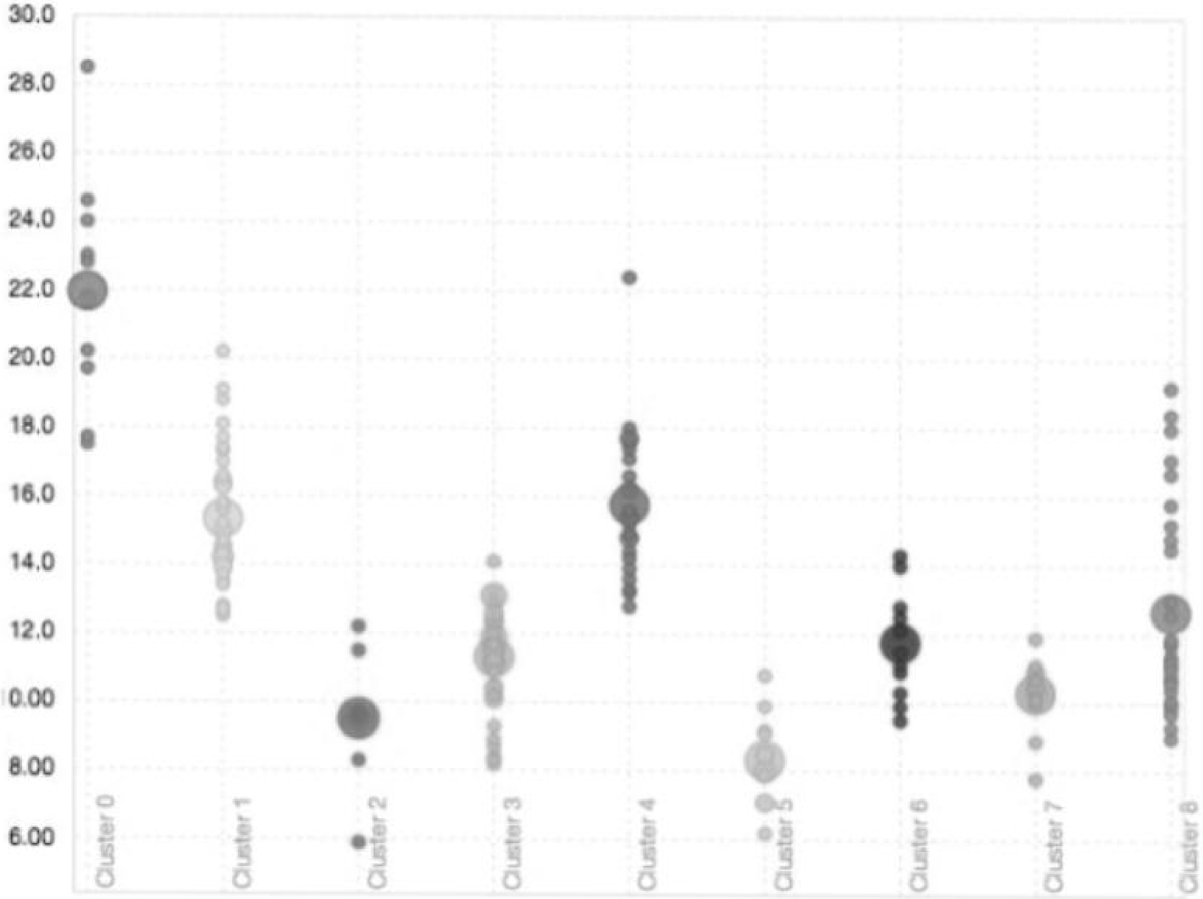


Figure 3.20: Mean price ticket scatterplot.

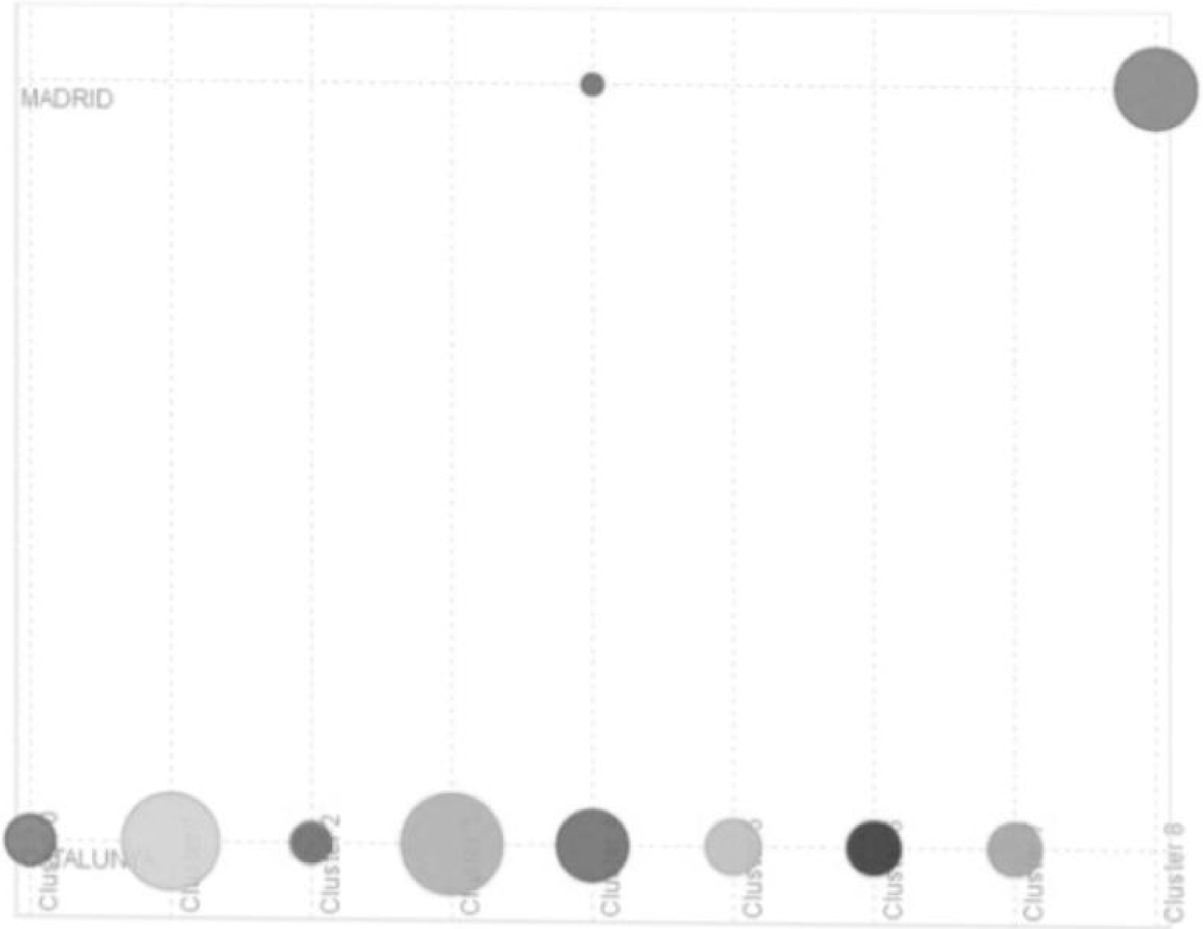


Figure 3.21: Region scatterplot.

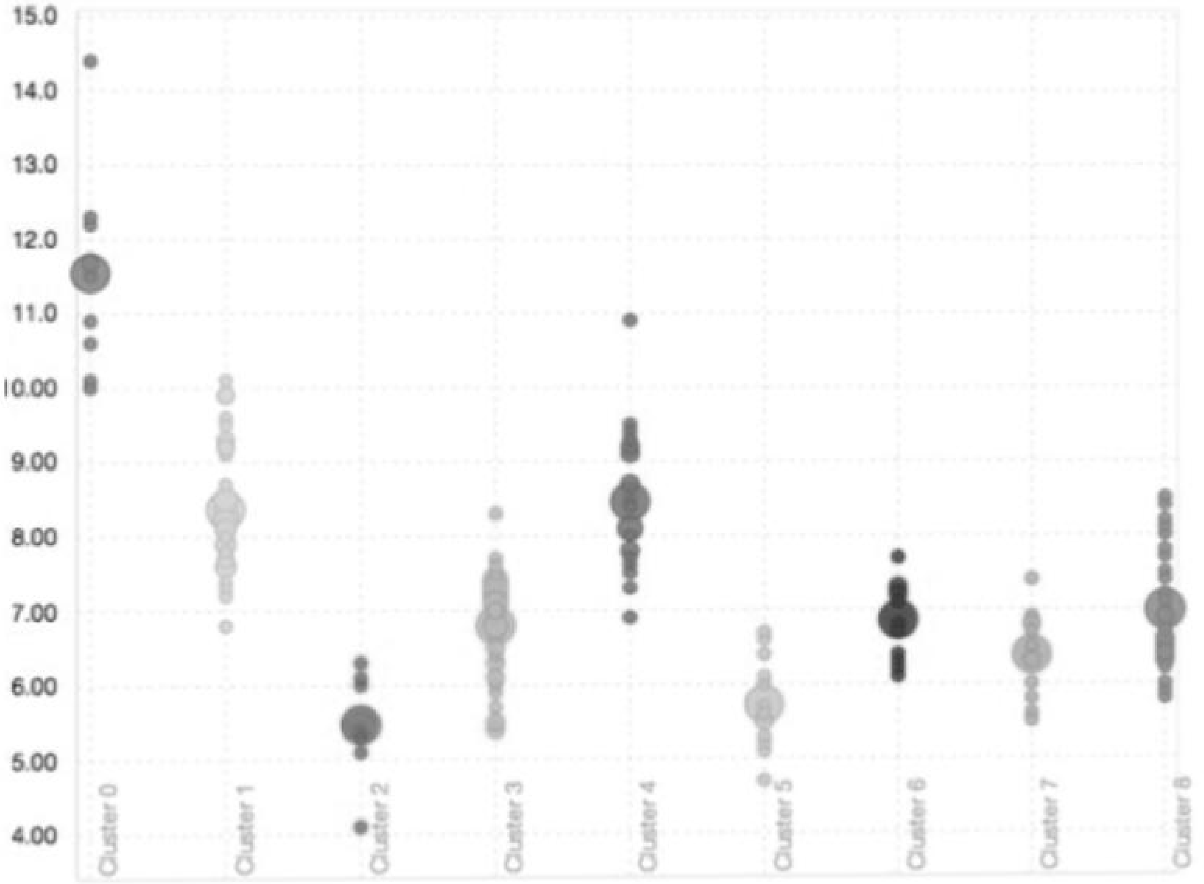


Figure 3.22: Mean units per ticket scatterplot.

5 Conclusions

The results obtained in this project were satisfactory. Objectives defined were achieved and the association rules discovered will provide a competitive advantage to our client over its competitors. Moreover, clustering provide a new vision of store's behaviour that can lead to future strategies. Unfortunately, we couldn't analyse the real impact of our solution to the business yet. The implantation of our solution to the business is difficult and has a huge impact, due that, the client need time to get organized.

Through the project, we detected the needs of the company and provided a solution to them. We proved that machine learning algorithms can be used to solve real world problems and how the use of them, can provide a quality advantage to companies over its competitors. In addition, we learnt a huge amount of valuable information about a retail business that helped us to approach different projects to other potential clients from retail domain.

We demonstrated the high potency of BigML. How machine learning as service breaks with the traditionals methods used nowadays to develop data science projects without losing performance or flexibility. Unfortunately, in this project we couldn't demonstrate the easy implantation BigML's models have in production.

Machine learning will change the world as we know it today. We understand machine learning as the key process in the business transformation, how companies think and make decisions. Those companies that don't adapt its methodologies and procedures to this new era are doomed to failure.

4 Evaluation of the project

Evaluation of a project is one of the most important task of any project. Analysis of results is vital. It tells if the project is going in the right direction or not. Basically, there are two ways to evaluate a project, offline and online. The first one, is performed during the project development, commonly done via evaluation of the model or meetings with the client. The second one, is performed once the project has been finished and is in production.

Through all the project, in order to evaluate the quality of the results we were obtaining, we were in continuous contact with the client. Periodically, we had meetings with the client where we presented the results and decided which was the next step to perform in the project. The priori idea was that those meetings where always at the same day every week, however, due it was difficult to coordinate client schedule with ours, those meetings were done when it was possible for both.

For the clustering part, those meetings were vital for the project, due as it was mentioned in a previous chapter, there is no metric that analyse a clustering model like in a supervised model. Thus, in order to analyse if the results we were obtaining had sense, at the end of each clustering version, we presented the clusters to the client. In this case, in all the versions we presented to the client, we received good news about how the algorithm split the stores. One thing that confirmed us that at the end of the clustering process we had already achieved a good clustering was that all the shops, except one, from Madrid, were automatically classified in the same cluster by the algorithm.

Once we finished the clustering, we discovered the association rules for the cluster number 4 as it was described in the previous chapter. Those associations rules were presented to our client in a last meeting with a brief summary of all the tasks that were done in the project.

A common practice realized in data science projects is the analysis offline of it. Results obtained during the project are important, however, the real test is the one performed in production. Those results tells if the project was an exit or a failure. In this project, the offline test proposed was the analysis of the customers purchased after the proper actions performed based on the association rules found. For instance, check if an item was purchased more often than before a special offer was created or the distribution of the store was changed. Unfortunately, we couldn't analyse the quality of the project yet. That's because the client is working in many other projects and the application of our results has a huge impact in the company.

However, even with no information about the impact that could have association rules in the stores, based on the experiences we had in the meetings with the client and the positive opinion he had, we considered that the project was an exit.