MDSAA

Master Degree Program in

Data Science and Advanced Analytics

City Clustering Tool at iFood

Data-driven approach to design online experiments

Rodrigo Rodrigues Simões Matias

Internship Report

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2022

Title:

Subtitle:

Student full name

MDSAA

[this page should not be included in the digital version. Its purpose is only for the printed version]

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

CITY CLUSTERING TOOL AT IFOOD

by

Rodrigo Matias

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Business Analytics

Supervisor: Roberto Henriques

September 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Rodrigo Matias

[Azeitão, September 2022]

DEDICATION

This internship report is dedicated to all my family, friends and colleagues at iFood.

ACKNOWLEDGEMENTS

I would like to salute the people who assisted me and contributed to this internship.

Specially all the colleagues who I have had the pleasure to work with in the profitability/finance data department at iFood for being first-class advisers and colleagues.

ABSTRACT

One of the many ways innovation occurs in big tech companies is due to A/B testing in order to achieve reliable results the design of these online experiments needs to be well thought. There are some business constraints that might hinder some key requirements of the design such as the fact that some tests can't be done under the granularity of users and most be done under the granularity of cities which might happen due to ethical and judicial constraints. In those cases in order to make sure that the chosen sample is a good representation of the population it's proposed a cientific approach of city clustering so that the test cities all together represent a bigger portion of the county plus a best matching city function in order to choose the control cities.

With the assumption that the introduction of a city clustering tool would improve the city A/B testing design consistency within the profitability department. The present document reports the descriptive details of the research, discovery, development and validation phase. Results show that new experiments done using said tool are more reliable than the ones done prior. Although results are positive, future steps are proposed, which includes a better UI/UX in order to facilitate stakeholder's interaction with the tool.

KEYWORDS

Food Delivery Platforms; A/B test; Clustering

INDEX

Content

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AOV	Average order value
CHS	Lack of demand Coefficient
PPH	Orders per habitant
RPH	Restaurants per habitant
GMV	Gross monetary value
Dau	Number of devices

1. INTRODUCTION

The present document is a scientific approach to the problem found during the internship that took place between the months of September and February of the year's 2021-2022 at iFood, which is a result of a partnership between Nova Information Management School and said entity. This internship report was made within the scope of the Master Program in Data Science and Advanced Analytics and crucial to my development as a professional on the Data field.

Even though iFood is a very data driven company and there is a big and robust data department there are still some decisions that are not made based on data. One of the first tasks I was asked to do was to analyze the results of an A/B Test regarding the increase of a fee on one city having another city has "control", my main concern was that those cities couldn´t be comparable and I couldn't find reason behind the decision on which city the test was going to be conducted.

Therefore, an accurate and reliable city clustering tool to support decision making regarding the choice of cities to conduct new tests would increase the performance and reliability of A/B testing using cities. The reliability of these tests is of extreme importance because they are rolled out to all of our customers and any mistake can be super costly due to iFood's dimensions. Also, the effects of elevated geopolitical risks in 2022 make companies that are dependent of Venture capital money injection more fragile which made iFood seek not only growth but also profitability.

Literature on city clustering and city A/B testing are quite limited, despite that there are several publications about user A/B testing and the usage of clustering procedures in the designing phase. The knowledge extracted from these past experiences is useful to determine guidelines for the following scientific approach.

In order to perform the clustering, it was used a K-Means algorithm where several features were used such as Demographic data, iFood Performance or iFood Behavior of those cities. Besides the clustering tool it was created a function making use of the Nearest Neighbors algorithm the user inputs the city where he wants to test a different feature and the output is the best city to compare it with.

After the implementation of the first version of the data product stakeholders found this tool helpful, the reliability and consistency of the experiments conducted increased.

2. COMPANY AND ROLE DESCRIPTION

This chapter will cover a brief history of the company, its mission, values and vision, and a more detailed view of the overall enterprise structure.

2.1. COMPANY PRESENTATION

iFood emerged in 1997 under the name of Disk Cook, working has a telephone exchange intermediating the process of receiving the orders and organizing the deliveries with restaurants.

iFood is a privately held Brazilian technology company, headquartered in São Paulo, Brazil. It was founded in 2011 by Patrick Sigrist, Eduardo Baer, Guilherme Bonifácio and Felipe Fioravante, and it was it this moment that the company migrated to the digital platforms where it started working as an app. In 2018 after the Series G founding round iFood became a unicorn surpassing the one-billion-dollar evaluation mark. The company´s core pillars are consumers, couriers, restaurants, and groceries stores. iFood´s mission is to feed the future of the world and the company values are Entrepreneurship, Results, Innovation and All Together.

iFood started in 2011 outside the technology world, as an impressed menu guide with a call central where customers would order, the company decided to deliver a better experience to the customer and launched an app and a website.

At the moment, iFood is a leader in online food delivery in Latin America with operations both in Brazil and Colombia. The company is the market leader in Brazil with +60 million orders per month, more than 270 thousand restaurants, and more than 40 million active customers per month.

iFood operates in Brazil and Colombia and has a work force of approximately 5000 employees just in Brazil. iFood´s structure has been through a constant evolution due to the high growth of the company.

2.2. TEAM AND ROLE PRESENTATION

The main objective of the Data Analysis department is to lead the business area in to making better decisions out of data insights in order to fulfill strategic iFood goals.

The goals of all the data analysis teams are aligned with their respective areas, this way the focus is to have a high impact in the organization.

An advanced data analyst identifies or understands a business problem, prepares the data in order to perform an analysis which will lead in a presentation of the actionable and relevant insights in order to communicate them to the business area. Those analysis evolve statistical and machine learning knowledge when necessary.

iFood takes pride on its high agility structure and I felt that on first-hand. I´ve been through the Finance data department and right now I am in the Profitability department which has been gaining a major importance in our company due to the fact that now more than ever it’s super important to not only measure growth but also profitability due to the investment decline that we are facing in 2022 which is a result of the global supply chain disruptions caused by lockdowns, the Russian war against Ukraine, a meltdown in technology stocks and a spike in inflation and subsequent interest rates. This investment decline affects mostly companies dependent in Venture Capital money injection.

3. LITERATURE REVIEW

The purpose of this chapter is to cover the research about topics related to the internship covered in this paper. In due course, the knowledge acquired along this literature review will be employed in the developed work during my internship.

Literature on city clustering tools for A/B testing design improvement and A/B testing on a city granularity is quite non-existing. Despite this, there are some publications that outline the key necessary elements of a well-designed A/B test on a user granularity and the guidelines extract are useful for the interpretation of the results.

3.1. UNSUPERVISED VS SUPERVISED LEARNING

Within machine learning there are two approaches: supervised learning and unsupervised learning. The main differences are that supervised learning require labelled data, while unsupervised learning does not. However, there are more distinctions between both.

Supervised learning is an approach where labeled datasets are created to train algorithms that will classify or predict outcomes. Supervised learning itself can be split into two types of problems: classification and regression. Some common algorithms of classification are logistic regression, decision tree, random forest, support vector machine, k nearest neighbor and naive bayes. (R. Sathya & Annamma Abraham, 2013)

Classification problems use an algorithm to classify test data and separate them into specific categories, one real world example is classifying customers that might buy your product. Regression is the other type of supervised method where an algorithm recognizes the connection between dependent and independent variables. Those algorithms aim to predict numerical values such as projections of revenue. (Love, 2002)

Unsupervised learning is an approach where machine learning algorithms discover patterns in data without any human intervention, the main tasks are clustering, association, and dimensionality reduction.

Clustering is a technique that groups data based on their similarities, Association is a method that finds relationships between variables in a given dataset and those methods are frequently used for market basket analysis, dimensionality reduction is used when a given dataset has to many features, so they are reduced to a manageable size while preserving data consistency.

There are algorithms that provide use for unsupervised and supervised learning which is the case of nearest neighbors. The concept behind nearest neighbor is to find a predetermined number of elements closest into the desired point. The distance is in general calculated using the standard Euclidean distance. (Chomboon et al., 2015)

3.2. CLUSTERING

Clustering is a simple conceptual activity, a process that is used in many sciences and a robust machine learning method. It evaluates data and tries to find patterns with no supervision with the objective of forming groups of similar elements (Shah & Koltun, 2017). There are four stages that define most of cluster analysis studies. The definition of variables, similarity criterion, algorithm and lastly is to conduct an analysis of the clustering solution (Jain et al., 1999).

3.2.1. Feature selection

Feature selection is the process of identifying the features that when combined minimize the intra-cluster distance and maximize the inter-cluster distances. In unsupervised learning there is no target variable to guide these decisions, but they can be guided by the domain knowledge and the reduction of highly correlated variables to guide the selection (Hancer et al., 2020).

This step is important because a simpler model is easier to interpret, has a shorter training time which leads to a reduce of costs, reduces overfitting, and reduces variable redundancy. There are three methods for feature selection: filter methods, wrapper methods and embedded methods.

Filter methods make use of statistical techniques to assess the relationship between the input variables and the target variable. There are several statistical tests that can be applied such as Pearson's correlation coefficient, Spearman's rank coefficient, ANOVA correlation coefficient (linear), Kendall's rank coefficient (nonlinear), Chi-Squared test (contingency tables) and Mutual Information.

Wrapper methods make different combinations that are evaluated and compared between them, a predictive model uses those features and scores them based on accuracy. Most common algorithm that uses the wrapper method is: RFE.

Embedded methods understand the contribution of each feature to the accuracy of the model while it's being produced, examples of algorithms are LASSO, Elastic Net and Ridge Regression. (Henriques, 2021)

3.2.2. Similarity criterion

Clustering methods often look at objects as points in a multidimensional space, where the similarities between those objects correspond to distances between the respective points. The most famous ones are:

- Euclidian distance: Distance between two points is the squared root of the sum of the squares of the differences between I and j values for all variables:

$$d_{ij} = \sqrt{\sum_{v=1}^{p} \left(X_{iv} - X_{jv} \right)^2} \, euclidean\, also\, known\, as\ L_2$$

$$d_{ij} = \sum_{v=1}^{p} \left| X_{iv} - X_{jv} \right| City\, Block\, or\, L_1$$

Fig. 1 – Formula of Euclidian distance. (Rossi & Testa, 2018)

- Minkowski distance: Distance between two points is defined from the absolute distance and it coincides with Euclidean distance when r = two and with Manhattan distance when r = one.

$$d_{ij} = \left(\sum_{v=1}^{p} \left| X_{iv} - X_{jv} \right|^{r} \right)^{1/r}$$

Fig. 2 – Formula of Minkowski distance. (Rossi & Testa, 2018)

3.2.3. Algorithms

Regarding the choice of a clustering technique there are 6 famous types of traditional clustering algorithms to choose from: Hierarchical methods, Partition methods, Density-based clustering, Fuzzy clustering, and Self-organizing maps.

Hierarchy-based algorithms make each data point its own cluster and then every two close clusters merge in order to assemble a new one and this process is repeated until only one cluster remains (Sasirekha & Baby, 2013)This hierarchy of clusters is represented as a dendrogram where the root is the unique cluster gathering all the data points.

The implementations of hierarchical algorithms often fall into two types: Agglomerative: each observation starts in its own clusters and clusters get paired as we move up the hierarchy and Divisive: Which is a "top down" method where all units start in one big cluster, and they are split recursively as we move down the hierarchy.

Partition based algorithms divide data into non-hierarchical clusters, the centers of the dataset are the centers of the clusters in the partitioning method it's up to the user to specify k number of partitions that will represent the number of clusters, the most popular algorithms that come under portioning method are K-means, K-Mediods and CLARA algorithm. (Velmurugan & Santhanam, 2011)

K-means algorithm starts with a random group of centroids, which is the starting points for every cluster, which then performs repetitive calculations in order to optimize the positions of the centroids.

In Density-based clustering, algorithms form different clusters regarding the data region's density at a certain area. This methodology is able of creating arbitrarily shaped clusters, where clusters are defined as dense regions separated by low-density regions. Two well-known density-based algorithms are DBSCAN and OPTICS (Ram & Kumar, 2010).

In fuzzy clustering algorithms clusters have assigned to them data units based on their level of belonging. That relationship is described by a continuous interval [zero,one] where 0 equals belong and one equal not belonging. In fuzzy clustering one element may belong to more than one cluster. Examples of such algorithms are FCS, FCM and MM (Akash et al., 2011)

Self-organizing map is an unsupervised neural network model, extensively used for clustering. SOM reduces data dimensions and displays similarities among data, this algorithm is commonly found used in conjunction alongside with other algorithms such as hierarchical-based algorithms in order to use a dendrogram that enables the user to find the correct number of clusters. (Vesanto & Alhoniemi, 2000).

3.2.1. Performance Evaluation

The evaluation of the performance of a classification problem which is a supervised learning approach where the target variable is categorical can be calculated using a wide range of different metrics such as classification metrics, confusion matrix, precision, and recall, F1 score, sensitivity and specificity, ROC curve and AUC.

When it comes to the evaluation of the performance of an unsupervised learning problem such as clustering is not as trivial as the calculation of a precision and recall or counting the number of missed categorization for a classification problem.

Clustering algorithms performance are assessed with a similarity or dissimilarity measure such as the distance between cluster points. Basically, if an algorithm was able to separate dissimilar observations apart and similar observations together, it performed well.

One method is the Silhouette Score which measures the separation distance between clusters. The Silhouette Plot shows how similar each observation in a cluster is to observations in the neighboring clusters. It ranges between [-1, 1] and is a notable tool to visually inspect the similarities within clusters and differences across clusters.

The calculation of this measure uses the mean intra-cluster distance (i) and the mean nearest-cluster distance (n) for each sample. The Silhouette Coefficient for a sample is (n - i) / max(i, n). n is the distance between each sample and the nearest cluster that the sample is not a part of while i is the mean distance within each cluster.

A common Silhouette Plot on the y-axis has the representation of the cluster label, while the actual Silhouette Score appears on the x-axis. The dimensions of the silhouettes are also equivalent to the number of observations inside that specific cluster.

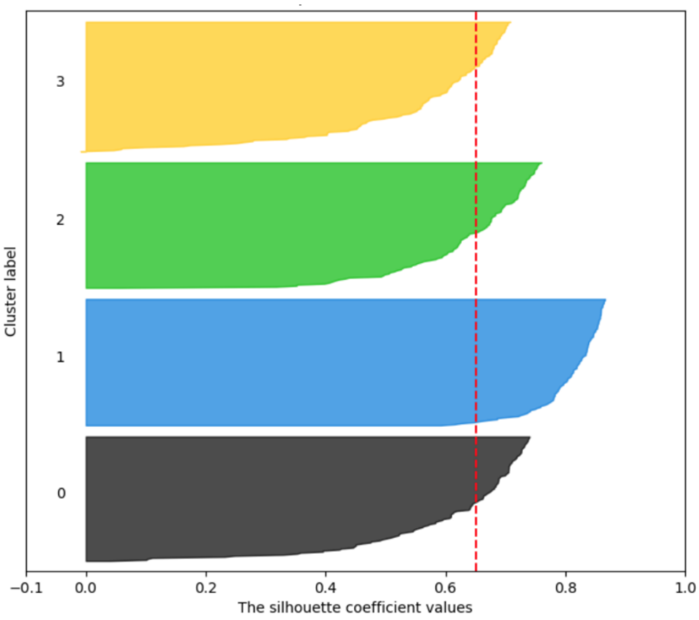


Fig. 3 – Example of a Silhouette Plot (Dinh et al., 2019)

As said before the Silhouette coefficients ranges between [-1,1] and the closer the Silhouette coefficient gets to +one, the remoter the cluster’s observations are from the neighboring clusters observations. A value of 0 is an indication that the observation is on the decision boundary between two neighboring clusters, on the other hand negative values, are an indication that those observations might have been allocated to the wrong cluster. Averaging those coefficients, we obtain the global Silhouette Score which is used to describe the entire population’s performance with a single number. (Dinh et al., 2019)

Another metric used to evaluate clustering algorithms performance is the Rand Index, the Rand Index is a similarity measure between two clusters by considering all pairs of observations and counting pairs that are allocated in the same or different clusters in the predicted and true clustering’s. The formula is:

$$RI = \frac{\text{Number of Agreeing Pairs}}{\text{Number of Pairs}}$$

Fig. 4 – Formula of Rand Index. (Santos & Embrechts, 2009)

RI ranges between [0-1], one being a perfect match. There is one drawback with this measure which is that has the assumption that we can find the true cluster labels and use it to compare the performance of the model.

There is also the Adjusted Rand Index which “adjusts for chance” the raw RI score using the following formula:

$$ARI = \frac{\text{RI - Expected RI}}{\text{Max(RI) - Expected RI}}$$

Fig. 5 – Formula of the Adjusted Rand Index. (Santos & Embrechts, 2009)

The Adjusted Rand Index also ranges from zero to one, having the same meaning has the raw Rand Index where zero is a random allocation and one is when the clusters are identical.

The Mutual Information is another metric frequently used in the evaluation of the performance of clustering algorithms. Measures the similarity between two labels of the same dataset. |Ui| is the count of observations in cluster Ui and |Vj| is the count of the observations in cluster Vj, the Mutual Information between clusters U and V is calculated using this formula:

$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

Fig.6 – Formula of the Mutual Information metric. (Haussler & Opper, 1997)

Just like the Rand Index there is one major drawback which is requiring to know the true labels a priori. Which makes this metric almost never being used in real-life scenarios for clustering.

Calinski-Harabasz Index also known as the Variance Ratio Criterion is explained as the ratio between the within-cluster dispersion and the between-cluster dispersion. The higher the Index the better was the clustering algorithm performance, this algorithm does not require information about the truth labels which makes it being a good metric for clustering performance evaluation.

The formula for this metric is the following:

$$s = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{n_E - k}{k - 1}$$

Fig. 7 – Formula of the Calinski-Harabasz Index. (Wang & Xu, 2019)

tr(Bk) being the trace of the between group dispersion matrix and tr(Wk) is the trace of the within-cluster dispersion matrix defined by:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

Fig. 8 – Formula of the within cluster dispersion matrix. (Wang & Xu, 2019)

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$$

Fig. 9 – Formula of the between group dispersion matrix. (Wang & Xu, 2019)

The Davies-Bouldin Index is the average similarity measure of each cluster with its most similar cluster. Similarity is the ratio of within-cluster distances to between-cluster distances. Clusters afar from each other and less spread lead to better scores.

In this metric the minimum score is zero and the lower the scores are the better performance the clustering algorithm had. Just like the Silhouette Score D-B Index doesn’t need the a-priori true labels.

Data abstraction is a step that might be next, it is a process of extracting meaningful information of the dataset it can be machine oriented if this information is used to enhance the performance of the process or human oriented in order to be easily comprehended and appealing. A typical data abstraction is a simple description of each cluster, such as the descriptive statistics of the cluster itself or of the centroid. (Jain et al., 1999)

3.3. BIG DATA ANALYTICS

Big data is a denomination for colossal data sets that have a huge and complex structure where storing, analyzing and visualizing data comes with more difficulties. Big organizations produce massive amounts of data and it's why revealing patterns and providing deeper insights it's key on getting an edge over their competition.(Elgendy & Elragal, 2014)

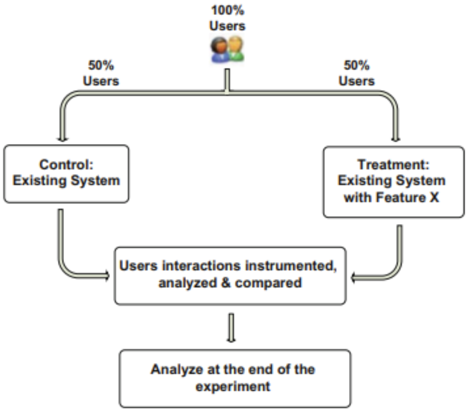
There are several big data frameworks with a common aim which is persistent data storage, maintenance of data availability and data consistency insurance the most common being Hadoop and Apache Spark.

Hadoop is an open-source batch processing framework that relies on modules designed under the assumption that the hardware will eventually fail. It operates by dividing files into blocks of data and spreads them over the nodes present in a cluster.(HGST, 2016)

Apache Spark is not only a batch processing framework but has also the capacity of stream processing. Making it a hybrid framework that is particularly easier to use, user-friendly to write apps in various programming languages such as Java, Scala, Python and R. (Veith & Assunção, 2018)

3.4. A/B TESTING

In a data-driven company there is no room for gut feelings, A/B testing is a methodology to test new features. The principle of an A/B test is to split your elements into two groups where you show the existing feature to the control group and the feature to the experiment group. Eventually evaluating the response of the users to the two different groups. Examples of tests are changes in the user interface, such as button colors, headline messages, page layout or even page load time. (Kohavi & Longbotham, 2017)



High-level structure of an user online experiment. (Kohavi & Longbotham, 2017)

It was only found literature about user A/B testing instead of city A/B testing as the ones that the clustering tool would be supporting but the fundamentals of their design should be the same:

The process of A/B testing begins with a hypothesis formulation. The null hypothesis assumes that both groups performed equally, and any difference is due to chance. The alternative hypothesis assumes the null hypothesis is wrong and the outcomes are different.

In order to analyze the results one can, use statistical hypothesis tests where if the p-value is lower or equal than the chosen significance level (alpha) the result is significant, and the null hypothesis is rejected and in the other hand if p-value is bigger than alpha the result is not significant, and the null hypothesis is not rejected.

The definition of the sample size is also very important in an A/B test, randomness and representativeness is extremely important in order to reach unbiased results and capture all types of users. (Kaufmann et al., 2014)

3.4.1. Cluster-Adaptive Network A/B Testing

One of the assumptions of A/B testing is the Stable Unit Treatment Value Assumption, which states that each user's response is affected by their own treatment only, which might not be valid and the test leads to a wrong conclusion. In (Zhou et al., 2020) research a cluster-adaptive network A/B testing procedure was proposed and numerical studies suggest that it achieves consistent estimation with higher efficiency.

This is a much more viable procedure to follow in a city A/B test due to the fact that cities have more dissimilarities between them than humans and it's logistically easier to choose N cities from each cluster than to follow a randomized procedure.

I will add a step forward by providing a tool for comparison between cities when said tests are difficult to make in a lot of cities, and needs to be a small number of test cities compared to a small number of control cities.

4. METHODOLOGY

The methodology that was used to solve this problem was CRISP-DM which is an abbreviation to Cross Industry Standard Process for Data Mining. It´s a model with six sequential phases that will be described on the next steps of this thesis:

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

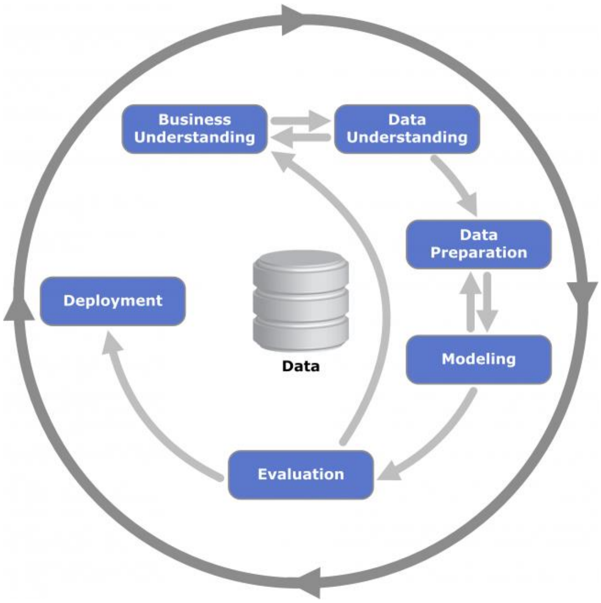


Fig. 10 – CRISP-DM Process image (Meeting & Chapman, n.d.)

Business understanding focal point is the comprehension of the goals and demands of the project. It's the foundation of the project and there are several steps like the determination of the business success criteria the determination of the requirements and risk assessment. In addition to the business success criteria, one should also specify technical data mining success criteria, and lastly select which tools and technologies will be used in each phase of the project.

Data understanding is an add on to the foundation of business understanding by identifying, collecting, and analyzing the data sets available that could be useful in order to achieve the project goals. Firstly, one should acquire and load the necessary data and scrutinize it in order to understand the data format, the number of rows and even the variables. Then, a deeper look on the data in order to explore it and understand and identify correlations among the data. Lastly, is the data quality verification.

Data Preparation is the phase where occurs the final preparation of the data prior to the modeling. Normally this phase takes 80% of the time in the whole data science project. It starts by selecting the data that will be selected, the values are then corrected, imputed, or removed, new features are created in order to provide more value and new information, different data sets may be joined during this step and the datasets are formatted as required.

Modeling is regarded as the most thrilling phase of a data science project and it starts with the determination of which models to try which is followed by splitting the data into training, test, and validation data sets which is dependent on the modelling approach or problem the data scientist is tackling, for example there is no data split on a clustering problem. After building the models itself they are compared and interpreted.

Evaluation is a comparison between the business and data mining success criteria and of what was achieved so far during the project, the work is reviewed, and next steps are discussed, in order to understand if the deployment is to proceed, if there will be more iterations of the model, or even if this project will end.

Deployment complexity can vary widely but the overall concept is that a customer/business team/stakeholder needs to access the model results. It starts with the planning on how the model should be deployed it can be as simple as a report or a model into production on an app used by 200 million persons a month. It’s where the development of a maintenance plan takes place in order to avoid issues once the project is in its operation phase. Also, the team needs to document and summarize the project and it might include a presentation of the results. If the model is going to production the work is not ending here, it’s imperative that the models are maintain and monitored so they can be occasionally tuned, in order to achieve better results and keep up to the new emergent technologies.

There are at least two different styles of CRISP-DM implementations: Agile and Waterfall. A Waterfall style implementation follows CRISP-DM precisely, defines detailed plans for each step of the project and documents heavily every iteration. On the other hand, if CRISP-DM is followed in a more flexible way, quickly iterated, and layered in with an agile process you will end up following an agile approach. Agile approach, when possible, provides, sooner value to the stakeholders, since they can provide feedback also sooner, it’s possible to evaluate the model performance earlier and the plan is adjustable based on feedback.

4.1. BUSINESS UNDERSTANDING

In big tech companies testing is key to surpass competition, iFood is no different, tests like user interface changes are common and easy to implement but tests where costumers in the same environment receive different fees are quite tricky since they can lead to unfairness. In order to overcome this problem, the profitability department who is in charge to maximize our revenue without damaging our orders numbers, designs A/B tests within cities, for example: All the clients who make an order below 50 R$ have to pay an additional fee of 5 R$ and we want to test in Rio de Janeiro if our orders drop if we move the threshold to 60R$.

As a data analyst in such a fast-paced environment as iFood’s I came across many quasi-A/B tests regarding new fees that had to be done under the granularity of cities, most of the times it was the business department choosing the city where the test would be done without even choosing a control city. Brazil is an extremely heterogeneous country how can we have the confidence that a test done in city A will have the same result as in the whole country? Which city is better to compare city A?

From a business perspective it´s important to improve the confidence in iFood´s A/B testing and make these decisions using data.

Regarding a technical data mining perspective, the goals are to provide a descriptive analysis of the cities clusters and a function that determines which city is the most comparable with the test city.

4.2. Data Understanding

iFood´s datalake is very extensive and there is plenty of data to work with, the biggest challenge is always to understand which dataset is required and the consistency of said dataset.

All the data produced by the app such as every time a client opens it, makes a purchase, opens a restaurant catalog and other services such as Zendesk (client support) every client support chat is stored or salesforce (merchant information) all the legal information of our clients (merchants) are stored and sent to the transient zone where the data is not in a usable state the data format at this stage can vary, csv, doc, xml, xls, txt. It then gets through a process that changes the data into tabular format where sensible client's information is cryptographed which is the raw zone.

Then there are two different usable zones curated and sandbox, curated zone are raw archives that were treated, validated, documented and have been through a quality process. Sandbox is a 60-day experimentation zone just to analyze fast and not perpetual datasets.

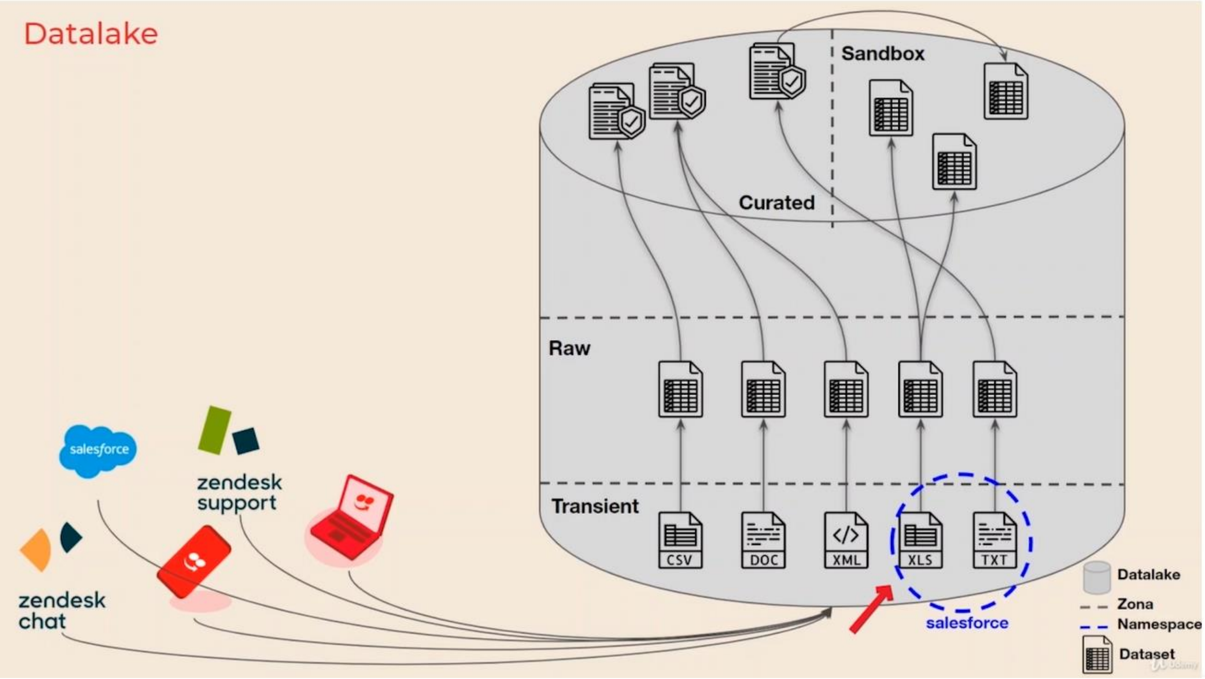


Fig. 11 – Representation of iFood's datalake

Collected data from datasets in the curated zone that had information regarding: Orders, sessions, demographics, and merchant information.

Brazil has 5570 municipalities and iFood operates in around 1500 cities. After all the data was collected the dataset had: 1507 rows and 27 columns described below:

Feature	Description
City	Name of the city
State	State abbreviature
Population_2021	Estimated population
Population_2011	Last Census Numbers
Mean_salary	Mean salary of the active population
CHS	Coefficient of lack of demand
PPH	Orders per habitant
RPH	Restaurants per habitant
% orders_unsubsidized	%Organic orders
City classification	Tier of the city
Total orders	Sum of orders in the city
Active restaurants	Number of restaurants with orders
Orders_dawn	Orders during dawn shift
Orders_breakfast	Orders during breakfast shift
Orders_lunch	Orders during lunch shift
Orders_snack	Orders during snack shift
Orders_dinner	Orders during dinner shift
Users	Total users of the city
GMV	Gross monthly value
AOV	Average order value
Dau	Total Devices
Total Sessions	Sum of sessions in the city
Merchant Online Time	Average merchant online time of the city

Table 1 – Variables

4.3. Data Preparation

Most of the work of a data analyst is spent on this step and working on this project wasn’t different. Datasets were collected and in order to gather more value I did some feature engineering:

Feature	Description
Slope	Average gain in orders in the last 5 months
% orders breakfast	(orders dawn / total orders) * 100
% orders lunch	(orders lunch / total orders) * 100
% orders snack	(orders snack / total orders) * 100
% orders dinner	(orders dinner / total orders) * 100
Total_rent_population	Mean salary * Population
Conversion	Total orders / Total sessions

Table 4 – Feature engineering table

Besides the feature engineering the datasets contained some noisy data like null values and outliers. In order to handle it there are a few ways, you may delete rows with missing values, or you may impute those values. Two of the columns (Dau (total devices) and merchant online time) had more than 80% of null values so they were dropped, lastly the column Total sessions had a median imputation of the null values.

Regarding outliers, 99.78% of the data was kept using a manual outlier removal method. After the outliers were removed then I analyzed the metric variables correlations using the Pearson method.

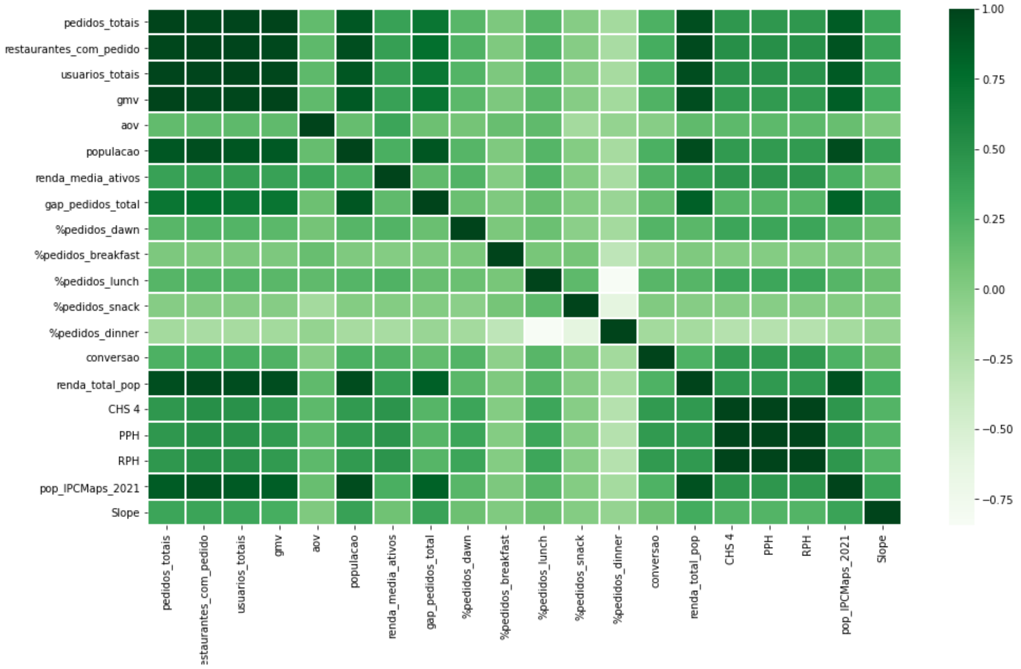


Fig. 12 – Variables correlation matrix

Analyzing the variable correlation matrix and conducting some tests in order to compare the results the final variable choice was conducted, and it resulted in 10 variables:

Feature	Description
City classification	Tier of the city
Slope	Average gain in orders in the last 5 months
PPH	Orders per habitant
RPH	Restaurants per habitant
Aov	Average order value
Conversion	Orders / Sessions
Total orders	Sum of orders in the city
Mean salary	Average salary in the city
Population_2021	Estimated population in 2021
CHS	Coefficient of lack of demand

Table 3 – Final Feature Selection

In order to get the numeric columns to a common scale two scalers were tested: Minmax and Standard. Minmax scaler transforms the variables by scaling each one to a given range where the default is zero and one. This scaling is given by:

X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))

X_scaled = X_std * (max - min) + min

Where min,max = feature range

I normalized the values using the standard scaler because it performed better when comparing silhouette scores and the interpretability. The standard scaler removes the mean and scales to unit variance, the standard score of a sample x is calculated as: z = (x-u) / s, where u is the mean of the training samples and s is the standard deviation of the training samples.

4.4. MODELING

The chosen model was K-means, K-means has a high performance when it come to large datasets, it's easy to implement and provides an intuitive result interpretation which is key in this project given the fact that it will be presented to stakeholder with no data science knowledge

As previously explained in the literature review K-means first step is the definition of the number of clusters for that I runed k-means with an initialization K-means ++ which selects the clusters centroids in a smarter way (not random) which speeds up the convergence of the clusters, K-means ++ initializes the centroids distant from each other, which leads to probably better results. In order to define the number 11 times, starting with one cluster ending with 11 clusters the results were the following:

The K-Means algorithms clusters data trying to segment groups of data (samples) in k groups of equal variances and minimizing inertia, which is a measure of how internally coherent clusters are.

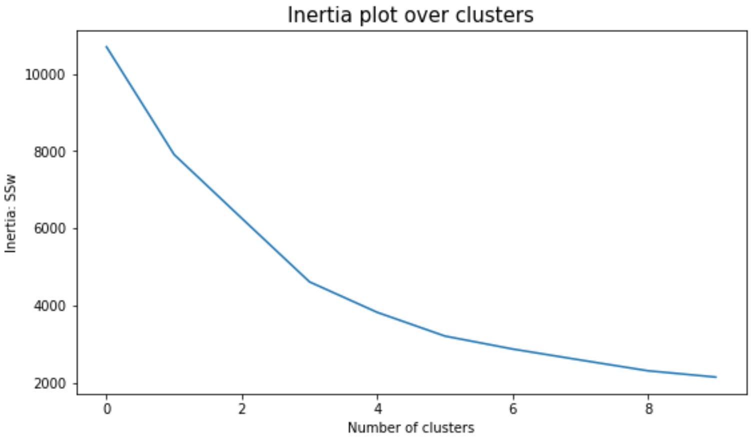


Fig. 13 – Inertia plot

After checking the inertia, I calculated the silhouette score for the various numbers of clusters:

Number of clusters	Silhouette Score
2	0.575
3	0.480
4	0.515
5	0.527
6	0.338
7	0.348
8	0.359
9	0.359
10	0.355

Table 4 – Silhouette score

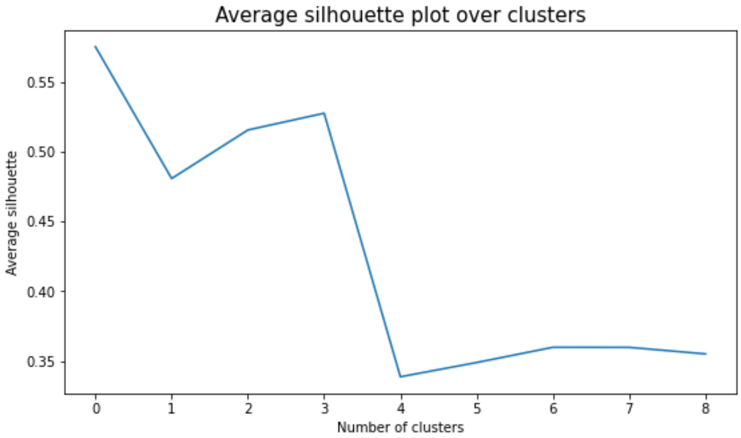


Fig. 14 – Average silhouette score

Based on the results and the domain knowledge the number of clusters chosen was six and the results were the following (average of the most important features for each cluster):

Clusters	Number of cities	Slope	PPH	Total Orders	Aov	Conversion	Average Salary
1	737	0.208296	0.224220	0.202334	0.010848	-0.073964	0.127528
2	258	0.737339	1.278421	0.700951	0.066003	0.141913	0.698419
3	36	0.177760	0.503224	0.156373	0.183707	0.345319	0.367595
4	703	0.209981	0.247778	0.200858	0.017942	-0.006197	0.139541
5	24	5.177626	1.691734	5.064442	0.043253	0.285432	1.350929
6	21	0.278851	0.492972	0.256938	0.196912	0.250171	0.270797

Table 5 – Average of the most important features within each cluster

After reading the results we could understand that regarding the slope of growth the cities in the cluster 5 are the ones with the higher value and they also lead regarding the orders per habitant, total orders, and average population salary. Average order value is led by cluster number 6, while conversion is led by cluster number 3.

Using the same variables of the segmentation I developed a function using Pyspark, which is an interface for Apache Spark in Python, that receives as input one city and indicates n (user defined) cities that would be a great match to compare with. The algorithm used in order to make this match was Nearest Neighbors which calculates the distance between cities using the standard Euclidean distance.

4.5. Evaluation

The metric used to evaluate our city segmentation was the silhouette score. The silhouette score is calculated using the following formula: (b-a)/max(a,b), where a= average intra-cluster distance and b= average inter-cluster distance.

The silhouette score for 6 clusters was 0.338, which leaves room for improvement. Choosing less clusters would have led to a better score, but we would lose the Brazilian heterogeneity which was a business request.

Using the city segmentation and the best match city tool our stakeholders in the profitability department were able to design diverse and consistent tests making the results more reliable.

4.6. Deployment

The segmentation itself was used to create a table which is being updated every other week using Apache Airflow where the granularity is cities and all the variables used in the clustering and the output itself are present.

This table is mainly used by business and data team's designing a test which needs to be done by city, so they try to choose a different city from each cluster in order to achieve a better sense of diversity. The table in order to be located in the curated zone of iFood's datalake had to go through a data quality process.

This project also resulted in the creation of an in-house function where the user inputs the name of the test city, and the output is the name of the best matching city which is then used to compare the results. The algorithm chosen to conduct this match was nearest neighbors.

The concept behind nearest neighbor is to find a preset number of training elements closest in distance to the new point. The used metric measure was standard Euclidean distance.

The way our team chose to cater this solution was resorting to a databricks notebook in order to validate the need of this data product which so far has been receiving great feedback and has been used several times.

Example: Stakeholders want to know the best control city for Florianópolis, the Output would be a comparison between both cities:

City	City Classification	Slope	AOV	Conversion	Total_orders	Average Salary
Florianópolis	Tier 1	5.46	2.59	0.41	5.93	3.58
Ribeirão Preto	Tier 1	5.98	2.86	0.44	6.31	3.88

Table 6 – Best matching city function output

5. RESULTS AND DISCUSSION

The purpose of this chapter is to go through the results of the modeling and deployment steps while discussing their impact.

Presentation of the model scores for the city clustering:

Algorithms	Silhouette score
K-means	0.338
DBSCAN	0.31
Agglomerative Clustering	0.25

Table 7 – Silhouette scores (City Clustering)

Regarding the city clustering the utilized metric in order to measure the quality of the utilized model was the silhouette score: 0.338, this result is nowhere near of a good result, but it’s a good first step given that there were some specific business requirements the project had to attend. The overall silhouette scores show a possible lack of the data quality and the next step is to improve the information about each city and making use of the data provided by the Brazilian census that will be available in 2023.

6. CONCLUSION

Testing in big tech companies is key to innovation and innovation is what keeps those companies thriving, the consistency and reliability of those tests need to be an uttermost priority. Ensuring business teams have that in mind was the purpose of my internship in iFood. Making A/B tests in iFood regarding a color of a button or a page layout is trivial but A/B testing in the profitability department where we charge the client different fees can be quiet delicate even in a judicial level, which is why it has to be done under all the restaurants in certain cities leaving more external factors come into play during the test.

Making these cities A/B tests as reliable as the ones directly targeting users with changes in the UI is almost impossible since there are always external factors such as different behaviors in the restaurants or external factors in one city. But with this segmentation and best matching city function a step forward was made regarding this issue.

6.1. WORK EVALUATION

Working at iFood this past year has been crucial to my life as an advanced data analyst. Big tech companies supply enormous amounts of learning supplies and iFood really supports its employees in their self-improvement journey. My skillset and experience expanded both on technical skills and soft skills.

Regarding technical skills I became fluent with Python and learned Pyspark, I became more acquainted with working with a data warehouse and big data. Concerning soft skills I became a better problem solver, improved my storytelling abilities, and learned how to conduct meeting with non-technical stakeholders

6.2. FUTURE WORK

I have been receiving good feedback about the work I've done by my managers, colleagues, and business team peers. I will continue to work under the profitability team as an Advanced Data Analyst now with more responsibility given the fact that I was made effective in the company.

Regarding the cities clustering it will be under our maintenance and improvement pipelines, so that it keeps being useful as it is to the A/B test designs. The datasets we use in order to receive data from each city have dependencies so it's extremely important to monitor problems that may arise from other tables upstream.

7. REFERENCES

Akash, B. A., Al-Kilany, A. Z., Al-Maaitah, A. A., Al-Nimr, A., Badran, A. A., Sawaqed, N. M., Al-Ghandoor, A., Abdul-Hafiz, A., Dina Institute Superior Tecnico, M., Badiru, P., Bassam, E., Rashid, A., Malaysia, K., Tzou, M., & Eng Hasan Al-Ba, B. B. (2011). THE INTERNATIONAL ADVISORY BOARD EDITORIAL BOARD SUPPORT TEAM 2 B Language Editor. In *Jordan Journal of Mechanical and Industrial Engineering*. JJMIE.

Love, B. (2002). comparing supervised and unsupervised. *Psychonomic Bulletin & Review*.

Chomboon, K., Chujai, P., Teerarassammee, P., Kerdprasop, K., & Kerdprasop, N. (2015). *An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm*. 280–285. https://doi.org/10.12792/iciae2015.051

Dinh, D.-T., Fujinami, T., & Huynh, V.-N. (2019). *Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient* (pp. 1–17). https://doi.org/10.1007/978-981-15-1209-4_1

Elgendy, N., & Elragal, A. (2014). Big data analytics: A literature review paper. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8557 LNAI*, 214–227. https://doi.org/10.1007/978-3-319-08976-8_16

Hancer, E., Xue, B., & Zhang, M. (2020). A survey on feature selection approaches for clustering. *Artificial Intelligence Review*, *53*(6). https://doi.org/10.1007/s10462-019-09800-w

Haussler, D., & Opper, M. (1997). MUTUAL INFORMATION, METRIC ENTROPY AND CUMULATIVE RELATIVE ENTROPY RISK. In *The Annals of Statistics* (Vol. 25, Issue 6).

Henriques, R. (2021). *Feature selection Master in Data Science and Advanced Analytics BA and DS majors*.

HGST. (2016). *Taming the Long Tail in Apache Hadoop ® : Storage Tiering for Big Data*. https://www.usenix.org/system/files/conference/lisa14/lisa14-paper-kambatla.pdf

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Kaufmann, E., Cappé, O., & Garivier, A. (2014). *On the Complexity of A/B Testing LTCI, Télécom ParisTech & CNRS LTCI, Télécom ParisTech & CNRS* (Vol. 35).

Kohavi, R., & Longbotham, R. (2017). Online Controlled Experiments and A/B Testing. In *Encyclopedia of Machine Learning and Data Mining* (pp. 922–929). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_891

Meeting, B. S., & Chapman, P. (1999). *The CRISP-DM User Guide*.

R. Sathya, & Annamma Abraham. (2013). *THE SCIENCE AND INFORMATION ORGANIZATION INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ARTIFICIAL INTELLIGENCE*. www.ijarai.thesai.org

Ram, A., & Kumar, M. (2010). A Density based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases Sunita Jalal. *International Journal of Computer Applications*, *3*(6), 975–8887. https://doi.org/10.13140/RG.2.1.4420.1448

Rossi, G. C., & Testa, M. (2018). Euclidean versus Minkowski short distance. *Physical Review D*, *98*(5). https://doi.org/10.1103/PhysRevD.98.054028

Santos, J. M., & Embrechts, M. (2009). *On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification*.

Sasirekha, K., & Baby, P. (2013). Agglomerative Hierarchical Clustering Algorithm-A Review. *International Journal of Scientific and Research Publications*, *3*(3). www.ijsrp.org

Shah, S. A., & Koltun, V. (2017). Robust continuous clustering. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(37). https://doi.org/10.1073/pnas.1700770114

Veith, A. da S., & Assunção, M. D. de. (2018). Apache Spark. In *Encyclopedia of Big Data Technologies* (pp. 1–5). Springer International Publishing. https://doi.org/10.1007/978-3-319-63962-8_37-1

Velmurugan, T., & Santhanam, T. (2011). A Survey of partition based clustering algorithms in data mining: An experimental approach. *Information Technology Journal*, *10*(3), 478–484. https://doi.org/10.3923/itj.2011.478.484

Vesanto, J., & Alhoniemi, E. (2000). Publication 7 Clustering of the Self−Organizing Map Clustering of the Self-Organizing Map. In *IEEE Transactions on Neural Networks* (Vol. 11, Issue 3).

Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering*, *569*(5). https://doi.org/10.1088/1757-899X/569/5/052024

Zhou, Y., Liu, Y., Li, P., & Hu, F. (2020). *Cluster-Adaptive Network A/B Testing: From Randomization to Estimation*. http://arxiv.org/abs/2008.08648

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa