Machine Learning in Business Intelligence 4.0: Cost Control in a Destination Hotel

Fulgencio Sánchez-Torres[1]*, Iván González[2], Cosmin C. Dobrescu[2]

[1] Higher Polytechnic School, Universidad de Alicante, Alicante (Spain)
[2] MAmI Research Lab at Castilla-La Mancha University (Spain)

Received 17 January 2021 | Accepted 4 February 2021 | Published 25 February 2022

Abstract

Cost control is a recurring problem in companies where studies have provided different solutions. The main objective of this research is to propose and validate an alternative to cost control using data science to support decision-making using the business intelligence 4.0 paradigm. The work uses Machine Learning (ML) to support decision-making in company cost-control management. Specifically, we used the ability of hierarchical agglomerative clustering (HAC) algorithms to generate clusters and suggest possible candidate products that could be substituted for other, more cost-effective ones. These candidate products were analyzed by a panel of company experts, facilitating decisions based on business costs. We needed to analyze and modify the company's ecosystem and its associated variables to obtain an adequate data warehouse during the study, which was developed over three years and validated HAC as a support to decision-making in cost control.

Keywords

Business Analysis With Expert Assessment, Business Intelligence 4.0, Candidate Product, ICT Ecosystem, Machine Learning.

DOI: 10.9781/ijimai.2022.02.008

I. Introduction

When we talk about industry 4.0, we associate the term with the fourth industrial revolution, with artificial intelligence as a differentiating element. This work studies the use of data science as a support for business intelligence to control company costs [1], [2], [3], [4].

Cost management in companies is almost always a problem solved from the financial point of view. This work proposes and implements an alternative to cost management based on the analysis of product consumption. To do this, we apply Machine Learning (ML), specifically, hierarchical agglomerative clustering (HAC) algorithms to support decision-making [5], [6], [7], [8], [9]. The research took place in a hotel on the southeastern coast of Spain.

To delimit the investigation and ensure its viability, we studied the company in its environment and context. We analyzed the company's ecosystem and its Information and Communications Technology (eICT) ecosystem, together with the possible Artificial Intelligence environments to be used. Later, we acted on business and technological processes.

Since this is applied research, we incorporated a panel of experts in the areas of: ICT management, purchasing, and food and beverages. These professionals provided deductive knowledge to identify the items to be considered and evaluated the results obtained. This knowledge was combined with the inductive knowledge generated by Machine Learning, increasing the differential value of the research [10], [11].

* Corresponding author.
E-mail address: fulgenciosancheztorres@gmail.com

To ensure the viability of the work and avoid additional problems in the company, the investigation was carried out in phases, where the completion of one allowed the completion of the next

II. Methods

A. Hypothesis and Objectives

To deal with profit ratio problems, the tourism industry focuses on selling as much as possible at as high a price as possible, without addressing organizational difficulties in depth. Food and Beverages (F&B) are an important part of a hotel's operating costs, where the consumption of raw materials is difficult to manage. We have found works based on food costs during a trial period [12] and those focused on room or energy costs associated with hotel management [13]. Most solutions avoid the problem of control by allocating a percentage of the sales produced. Thus, these questions arise: Is applying a percentage of sales the best way to control consumption? Is this the only way to control consumption? This work proposes and develops an alternative to cost management by controlling product consumption instead of controlling financial cost [14], [15] [16], [17], [18], [19].

B. Scope of Work

The tourism sector is a market in need of innovation. The 4.0 paradigm can help digitize its value chain [20]. To do this, it is necessary to analyze companies as a whole, taking into account their environment, context, users, and computer systems to get a complete vision. A company's environment focuses on its surroundings and the data associated with systems other than in-house customer service. Context focuses on the circumstances of the company as an entity and its interaction through business processes [21], [22]. The data generated by these processes must be usable, so it is necessary to ensure that they are correctly produced, complying with the technical

and legal requirements established by the mandatory data protection regulations (RGPD) [23], [24] for companies in the European Union.

We grouped computer systems by the business processes related to the clients of the hotel. To do this, we used two dimensions, one referring to the company and the other to the clients of the hotel. We studied the systems' functionality and interaction, their usefulness in the study, and compliance with the GDPR. We established three degrees: high, medium, and low, to quantify the degree of interaction among the available systems.

C. Business Ecosystem

Using the cost-control approach based on the company's consumption and its associated elements, we focused on the F&B area. Consumption control can be carried out in two ways depending on the warehouse: weekly inventories of the products for production warehouses, in this case, the hotel kitchen, and permanent inventories that record the data in real time for the regulatory warehouse, in this case, the general warehouse of the hotel. This regulatory warehouse control allows management to keep track of direct consumption.

It was necessary to carry out some tasks prior to the study. The relationships between business processes and the data they generate were analyzed [25], [26], [27]. Internal meetings were held with the people involved to avoid rejection. The systems designed to control product consumption were tested, and support measures were implemented regarding regulation warehouses, safety stock, and real-time inventories.

Counting on a panel of experts is a basic part of proposals based on data science to obtain results that are consistent with the problem they address. This has been treated in previous works [28] [29], [30], [31]. Knowing what and how to assess is best carried out by experts [32].

Contextualizing the data, their selection, and the degree to which they could be used by the expert panel defined the variables. We kept in mind the viewing preferences of reference in the sector [33] and when the categorization of the data and items that define them condition their analysis and use [34]. It was crucial to analyze the dimensions of the products based on the needs of the study and not on manufacturers' specifications. For this reason, we created and implemented a new concept for product management called the generic product, which grouped all the products with equivalent dimensions. Each product used had its generic code Fig.1 and a unit of measure that characterized it (liter, kilogram, etc.). This was essential to reconfigure and provide feedback to the system based on need and the analyses carried out over time.

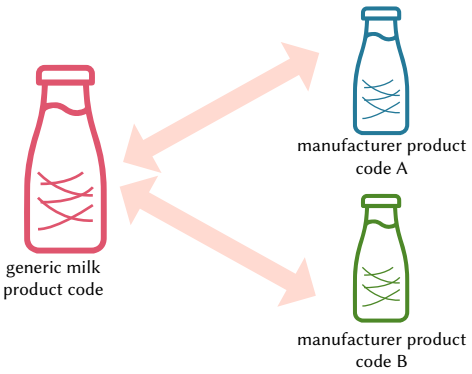


Fig. 1. Generic product.

D. The ICT Used

The technologies used were grouped by their functionalities for data collection and treatment, data analysis, and AI techniques.

1. Data Generation and Access

We started with a double handicap. The ICTs used in the tourism sector normally collect and store data in isolation. Each type of software generates its island of data. Moreover, software manufacturers do not facilitate access to the data directly. Therefore, it was first necessary to ensure valid, accurate, complete, and consistent data over time for the data warehouse. The selected systems were Warehouse Management System, Supply Chain Management, Property Management Systems, Point Of Sale, and Finance System.

We identified the data to use and its associated fields in the database. To obtain the data, we used MSQL Server Management Studio, SoapUI, Postman, and XML and JSON developments we carried out. During the research, it was necessary to add new software functionalities and make changes to the management systems. The modifications were made from the functional and technical points of view, for example, incorporating mobility and barcode systems.

2. AI Environment

The work took place where scientific and expert knowledge converged, making it difficult to distinguish where one began and the other ended. The domain experts performed deductive tasks, and the engineering experts, inductive ones. We needed a work environment that would allow us to evaluate the possibilities of ML, although soon after the work began, we decided to use HAC algorithms. These, by grouping the possible candidate products in an unsupervised way, facilitated business analysis with expert assessment (BAEA), and this knowledge became part of the company's know-how.

The chosen suite had to be used in the company and be especially oriented to data science [35]. We chose Python [36] with the Anaconda distribution as it is an open-source suite conceptually designed for data science and encompasses applications and libraries.

3. AI Techniques

Looking for an alternative to classical management, we avoided techniques such as support vector machines based on training, kNearest algorithms based on supervised learning with training sets, or tree and forest algorithms. Initially, clustering [37] and classification techniques seemed to be the most appropriate, although these can be considered similar to each other when identifying the two behavior patterns. The essential difference is that data classification uses predefined classes to perform the grouping, and clustering identifies similarities between the objects of the data sets by grouping them by their common characteristics.

In our study, input data was available without any type of labeling, from which information was obtained without conditioning the final result. This led us to unsupervised learning. We were looking for common patterns in the items that would give us candidate products to be substituted for more cost-effective products. The decision to substitute these products would be made in the [38] BAEA. This led us to agglomerative clustering algorithms about which there are different works [39], such as those that compare clustering algorithms [40].

E. Implementation

We decided to carry out the work in phases, with the completion of each allowing the following to begin. In this way, we were able to better delimit the study and cause the least possible disturbance to the hotel.

1. Phase 0. Preparing the Data Set

We wanted to work with complete and quality data sets to be able to test the AI environment, variables, and algorithms. The information needed to be consistent over time. Therefore, the data warehouse had to be separated from the business processes of the hotel. To extract the data, we used SQL, XML, and JSON.

Previous data. First, we analyzed needs from the point of view of business logic. For example, to obtain information about product consumption, it was necessary to consider all the warehouses involved in the product's life cycle and the business processes that affected it.

Record selection. Consistent and complete data, including all the fields and records were needed to facilitate assessment.

Consumption control is associated with inventory cycles and can be carried out daily or weekly, depending on the warehouse. This is why we established weeks as control units, defined as Monday to Sunday, based on the merchandise replenishment cycle in the warehouses.

After data processing, we had a CSV file (separated by ;). UTF8 standard, encoding=ISO-8859-1, with the following nomenclature *Proposal phase_number of variables used_variable weighted and records used.csv*. Table I.

TABLE I. Description of File Name Composition

Phase	Phase of the proposal where the file is used
Var	Number of variables included in the file
05	Percentage of the weight of the main variable for the records of the file
FullReg	All records available
Esp	Records weighted by the consumption variable
Cos	Records weighted by the cost variable

Example. *Phase1_3Var_05_EspReg -> Phase1*, 3 variables used, weighting the registers at 5 percent of the consumption variable.

Data validation. Validation was carried out first from the business logic perspective to get an early approximation of the suitability of the data. The data can have a correct format but an incorrect value. To do this, graphic representation techniques, frequency distribution, and crossing variables in two and three dimensions were used. We used data from one year applying the *Sturges* rule that permits consistent scaling. This allowed us to identify errors and the business process where they originated, as well as the person responsible for correcting them. Fig. 2 shows a three-dimensional crossover of variables between the unit cost, sum of consumption - sum of cost, where we identified discordant elements requiring action in business processes. We needed the products to be correctly coded using the concept of generic product and its unit of measure. We also needed to identify the variables and values that did not provide value in controlling consumption and cost and that simply added noise to the data.

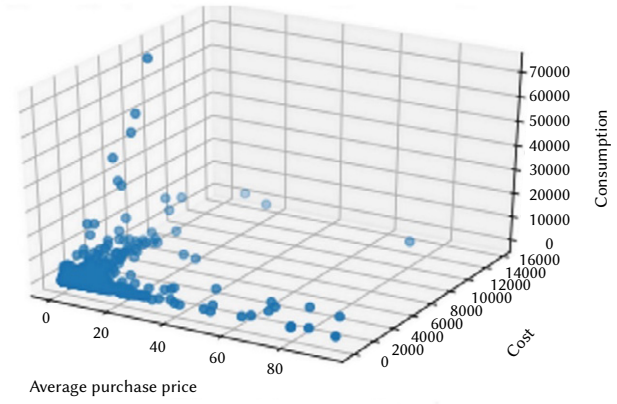


Fig. 2. Representation of variables.

Suitability of the data. The different HAC algorithms generated different solutions for evaluation by the panel of experts. The algorithms based their strategy for generating clusters on the minimum possible distance and maximum similarity between them. At first, to study the validity of the data, we used the single algorithm with Euclidean metric from the *linkage* package of *scipy.clusters.hierarchy* [41].

The three-dimensional representation of clusters and distances Fig. 3 shows us some examples of suitable values in green, and outliers in red. To see the relationships among the possible candidate products, the clusters formed, and the distances between the clusters, we used dendrograms to facilitate the interpretation of the data. A vertical dendrogram facilitates the analysis even more by representing the distances or heights on the "y" axis. An additional problem is the representation of data that does not add value to the study. To solve this problem, we decided to truncate the dendrograms and not graphically represent the values below a certain reference value. This improved the visual analysis Fig. 4. This figure shows the same data set with the weighting of records with the cost variable at 5 percent in the weight of the data set. The analysis allowed us to discard records with values close to zero with a weight in the variable of less than 5 percent of the total weight. It also made it possible to identify clusters linked at great distances that were susceptible to being analyzed directly.

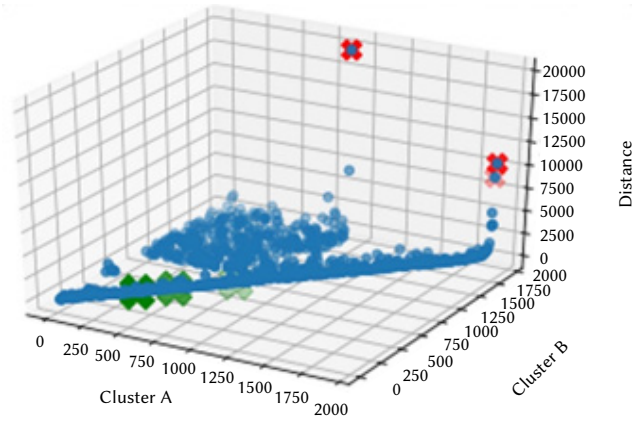


Fig. 3. Representation of clusters and distances.

2. Phase 1. Control Variables and Characterization Items

We continued with the application of the HAC algorithms and, specifically, with the single algorithm with Euclidean metrics to evaluate the possible combinations Table II. The variables under study were product consumption, product cost, and average purchase price. We used new data sets from the year 2017 that met the requirements of the previous stage.

TABLE II. Data Set and Variables Used

data set	consumption	cost	average purchase price	weighted variable
Phase1_ 3Var _FullReg	X	X	X	all records
Phase1_ 3Var _05_EspReg	X	X	X	consumption
Phase1_ 3Var _05_CosReg	X	X	X	cost
Phase1_ 2Var _05_EspReg	X		X	consumption
Phase1_ 2Var _05_CosReg		X	X	cost
Phase1_ 2Var _05_EspCos	X	X		consumption
Phase1_ 2Var _05_CosEsp	X	X		cost

For an objective evaluation, we made a table that reflected the results after applying the algorithm to the different data sets. This table was modified during the proposal by adding the results of each new phase. The data reflected initially were:

- Data sets. The data set used, which reflects the number of variables used and therefore prevails at 5 percent over the others.
- Weighted variable. Significant variable in the data set used.
- Algorithm clusters. Number of clusters generated by the default algorithm.
- Cophenetic coefficient. Cophenetic correlation coefficient.
- Maximum inconsistency value. To see how close it is to the fixed maximum value of 2.
- Number of elbow clusters. Number of clusters generated by observing the elbow method.
- Maximum Acceleration value.
- Minimum cutting distance.
- Maximum cutting distance.
- Number of candidate products. Number of candidate products identified after applying BAEA.

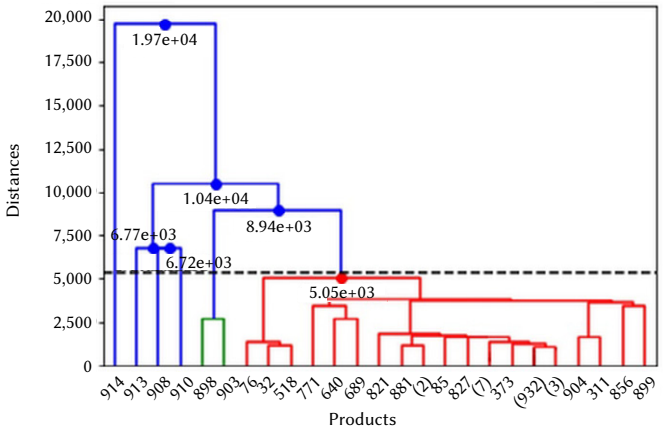


Fig. 4. Truncated vertical dendrogram.

Analyzing the data in Table III, we see that: the cluster number generated by the default algorithm ranges between 3 and 4; the number of candidate products is greater for two variables; the cophenetic coefficient is closer to a value of one for two variables; and the elbow method is not significant in this case. We see that the acceleration is less for two variables. At first, the algorithm behaves better for two variables than for three. At this point in the study, this is not significant because what is of interest is validating the usefulness of the HAC algorithms to propose candidate products to which BAEA can be applied.

3. Phase 2. Selection of the Algorithm

We studied the ability of single, complete, weighted, average, centroid, median, and Ward hierarchical agglomerative algorithms to propose candidate products using the variables consumption and cost. To do this, we used new data sets with two variables, one weighted by the consumption variable and the other by cost. We used new data from the year 2018 and corrected it as necessary. Finally, a maximum inconsistency of two was set.

To bring the research closer to the reality of the hotel, we included additional characteristics of the products from the kitchen and shopping requirements Table IV. We evaluated the combinations among the data set, algorithms, and new characteristics.

- Algorithm: Algorithm used.
- Acceleration cluster: Cluster where the acceleration is triggered.
- Distance difference: Difference between the maximum and minimum distance.
- Outliers: Number of outliers calculated by the default method.

We updated the characterization table V with the new items for all the algorithms, using the *Phase2_2Var_05Espcos* data set to illustrate it. This data set contained the products with a consumption incidence greater than 5 percent with respect to the cost variable. The algorithm that came closest to this hypothesis after applying BAEA was the average algorithm.

We analyzed the graph generated by applying the elbow method Fig. 5. We see the acceleration represented by the yellow line, marked with a light purple arrow where it begins to shoot, and in with a dark purple arrow where it reaches its maximum. The blue line reflects the distances, identifying the maximum and minimum (red and yellow, respectively). To find the interval identifying candidate products, we used the lower and upper cutoff distances. The minimum cutoff distance is after low cluster values (the light green arrow). The maximum cutoff distance is before the last cluster value (dark green arrow).

TABLE III. Machine Learning Assessment

data set	weighted variable	algorithm clusters	coefficient cophenetic	maximum inconsistency value	number of clusters elbow	maximum acceleration value	minimum cutting distance	maximum cutting distance	number of candidate products
Phase1_3Var _FullReg	all records	3	0.8672	1,9098	4	6	3,949	15,050	4
Phase1_ 3Var_05_EspReg	consumption	4	0.8523	1,1522	4	6	751	15,050	5
Phase1_ 3Var_05_CosReg	cost	3	0.8594	1,8418	4	6	258	22,755	5
Phase1_ 2Var_05_EspReg	consumption	3	0.9336	1,9599	4	5	751	15,050	7
Phase1_ 2Var_05_CosReg	cost	4	0.9231	1,9685	4	5	449	22,755	6
Phase1_ 2Var_05_EspCos	consumption	4	0.9260	1,8907	4	5	751	15,050	6
Phase1_ 2Var_05_CosEsp	cost	3	0.9596	1,8418	4	5	258	22,755	6

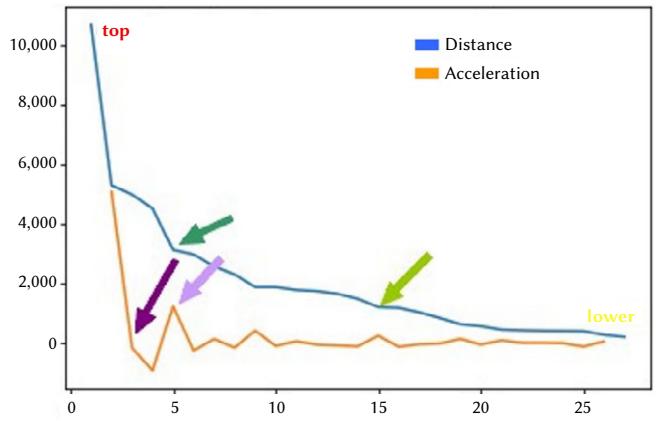


Fig. 5. Elbow method representation.

In the dendrogram Fig. 6, we observe the relationship between clusters, distances, and candidate products. We indicate the candidate products with a purple arrow, an example of two candidate products is shown with a blue circle, and two possible candidate products not valued because they fall outside the minimum distance set are shown with an orange circle.

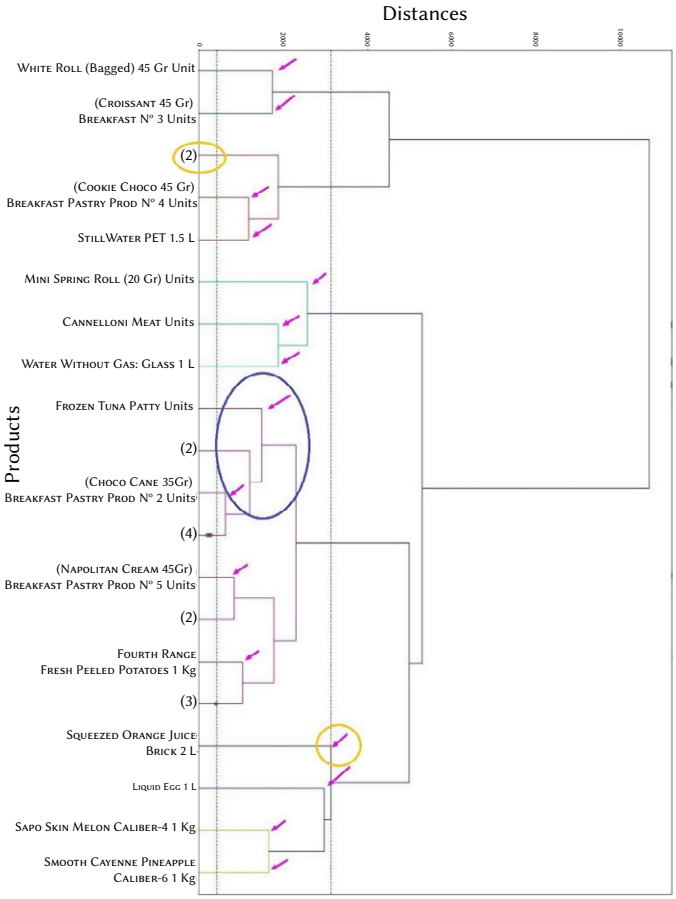


Fig 6. Relationship between clusters, distances, and candidate products.

Fig. 7 shows the relationship between the *cost* and *consumption* variables, identifying the same elements from the dendrogram in Fig. 6.

Low acceleration helps us identify the intermediate clusters since the closer we get to the value of one, the more significant it will be in the final part. A large difference between the distances helps us locate the clusters that are included and where the outliers will be identified.

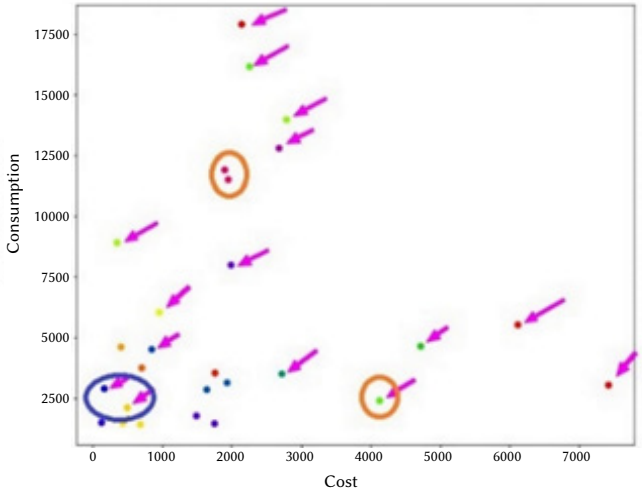


Fig 7. Relationship between variables and candidate products.

4. Intermediate Results

In this phase, we identified the algorithm that proposed more candidate products. We can see the characterization data in Table IV. The data that most interested us was the number of candidate products selected after applying BAEA that were delimited by the difference between the distances. There were other interesting results that helped us choose the algorithm. They are the following: having a cophenetic coefficient closest to one indicates a better correlation with the initial data set; low maximum acceleration at the beginning indicates that it will shoot up in the final part of the dendrogram; and the number of outliers indicates the number of elements that are left out and can be seen directly.

We know that when HAC is applied, there is not just one solution, and it depends on different factors, so we need to approximate an algorithm to each hypothesis. In Table IV, we see the data identified for each hypothesis. Blue is for cost and green for consumption. We highlighted the selected algorithm in bold for each color. The consumption hypothesis and average algorithm are in green, and the cost hypothesis and average algorithm are in blue.

5. Phase 3. Profile Variables

With the inclusion of the profile variables associated with client profiles, we wanted to see the relationship between client type and candidate products to facilitate decision-making based on the hotel's clients [42], [43]. Including financial variables made it possible to study the relationship between consumption and the two types of financial imputations studied, production in the F&B area and total hotel production. Considering the results of phase 2 on the applied algorithms, average for consumption, and median for cost, we addressed the profile variable. To do this, it was necessary to expand the data warehouse with the data for the period between 2018-07-01 and 2019-06-30. The variables were defined as:

- Customer profile.

a) Number of people = Breakfast (all guests have breakfast included).

b) Meals included = clients with lunches and dinners included in the hotel package purchased.

c) Occupation = Number of people + Number of meals included.

- Financial.

a) Billing of the A&B area.

b) Total hotel billing.

Three work scenarios were proposed for the study Table V. One scenario was related to the control of consumption and the variables

TABLE IV. Algorithm Assessment

data Set	weighted variable	algorithm	algorithm clusters	coefficient cophenetic	maximum inconsistency value	number of clusters elbow	acceleration cluster	maximum value acceleration	minimum cutting distance	maximum cutting distance	distance difference	outliers	number of candidate products
Phase2_2Var_05_EspCos	consumption	single	6	0.860057	1.2960	3	3	724	208	3,054	2,846	1	7
Phase2_2Var_05_CosEsp	cost	single	3	0.932709	1.1154	4	4	930	19	4,319	4,300	1	8
Phase2_2Var_05_EspCos	**consumption**	**average**	**3**	**0.904901**	**1.1535**	**2**	**2**	**5,241**	**208**	**10,678**	**10,470**	**0**	**15**
Phase2_2Var_05_CosEsp	cost	average	3	0.942683	1.1546	2	2	6,178	19	12,317	12,298	0	8
Phase2_2Var_05_EspCos	consumption	weighted	3	0.902703	1.5356	3	3	2,188	208	9,813	9,605	0	8
Phase2_2Var_05_CosEsp	cost	weighted	3	0.959757	1.1535	2	2	2,76	19	9,462	9,443	0	7
Phase2_2Var_05_EspCos	consumption	centroid	3	0.904757	1.1539	2	3	5,726	208	10,477	10,269	0	8
Phase2_2Var_05_CosEsp	cost	centroid	3	0.942070	1.1542	2	3	1,589	19	12,22	12,201	0	8
Phase2_2Var_05_EspCos	consumption	median	3	0.904390	1.1537	3	3	4,87	208	9,647	9,439	0	7
Phase2_2Var_05_CosEsp	cost	median	3	0.959928	1.1538	2	3	2,152	19	9,293	9,274	2	8
Phase2_2Var_05_EspCos	consumption	ward	3	0.837265	1.1529	2	3	21,246	208	32,312	32,104	0	6
Phase2_2Var_05_CosEsp	cost	ward	3	0.919233	1.1496	2	2	2,692	19	42,928	42,909	0	6
Phase2_2Var_05_EspCos	consumption	complete	3	0.849150	1.1391	2	3	3,109	208	16,554	16,364	0	6
Phase2_2Var_05_CosEsp	cost	complete	3	0.936782	1.1349	2	3	3,337	19	17,989	17,970	0	5

TABLE V. Associated Profile Variables

number	data set	weighted variable	number people	number pensions	occupation	A&B billing	hotel billing	algorithm
1	Phase3_6Var_05EspCos_Ab_Hot	all records	X	X		X	X	average
2	Phase3_2Var_05Esp_Per	consumption	X					average
3	Phase3_2Var_05Esp_Pen	consumption		X				average
4	Phase3_2Var_05Esp_Ocu	consumption			X			average
5	Phase3_2Var_05Esp_Per_Pen	consumption	X	X				average
6	Phase3_6Var_05CosEsp_Ab_Hot	all records	X	X		X	X	median
7	Phase3_6Var_05CosEsp_Ab	cost				X		median
8	Phase3_6Var_05CosEsp_Hot	cost					X	median

TABLE VI. Profile Assessment

number	algorithm clusters	coefficient	maximum inconsistency value	number of clusters elbow	maximum value acceleration	minimum cutting distance	maximum cutting distance	cluster below cutoff	minimum distance	maximum distance	distance difference	outliers	possible products candidate	candidate products
1	2	0.909555	1.154697	2	4,395	519	5,249	6	152	19,622	19,470	1	17	14
2	2	0.947091	1.145795	3	2,870	54	2,102	11	16	19,537	19,521	2	16	5
3	2	0.947091	1.145795	3	2,870	54	2,102	11	16	19,537	19,521	2	16	5
4	2	0.947091	1.145795	2	2,870	54	2,102	11	16	19,537	19,521	2	15	5
5	2	0.947091	1.145795	2	2,870	54	2,102	11	16	19,537	19,521	2	13	5
6	3	0.925175	1.154313	3	6,914	190	2,528	14	52	17,369	17,317	3	22	12
7	3	0.902039	1.145677	3	1,051	95	1,401	26	1	4,549	4,548	0	16	4
8	3	0.902039	1.145677	3	1,051	95	1,401	25	1	4,549	4,548	0	17	4

people, meals included, and occupation. Another had to do with billing for both F&B and the total hotel, and a third covered all the possibilities, including these together with the data set and algorithms.

The characterization results are shown in Table VI. They associate the cost-control data, on a blue background, with the consumption data on a green background. The different hypotheses refer to the data obtained regardless of the variables and data sets used. The column showing the difference between distances illustrates how close together the results are. The column of possible candidate products indicates the number of clusters between the minimum and maximum cutoff distances. The inclusion of variables associated with customer profiles and billing improved the results, and these are related to the candidate products in Table VII.

TABLE VII. Evaluation Of Candidate Products by Hypothesis

candidate product \ number	1	2	3	4	5	6	7	8	Total
(Milk Buns 40Gr) Prod Breakfast Pastries Nº6 Unit	1	1	1	1	1				5
(Choco Cane 35Gr) Breakfast Pastry Prod Nº2 Unit	1	1	1	1	1				5
(Cookie Choco 45Gr) BreakfastPastry Prod Nº4 Unit	1	1	1	1	1				5
(Magdalena 30Gr) Breakfast Pastry Prod Nº1 Unit	1	1	1	1	1	1			6
(Miguelito 60Gr) Breakfast Pastry Prod Nº5 Unit	1	1	1	1	1				5
Frozen Trunk Tuna 1 Kg (LOINS)						1		1	2
COD FILLET T-1000g Frozen 1 Kg						1	1	1	3
Frozen Pork tenderloin 1 Kg						1			1
Frozen Tuna Patty Units	1	1	1	1	1				5
Swordfish Piece 10/30 Frozen 1 Kg							1		1
Cooked Shrimp 40/60 Frozen 1 Kg						1		1	2
Toad Skin Melon Caliber-4 1 Kg	1					1			2
Artisan Bread Rhombus 30Gr Uds	1								1
White Toasts 45Gr Uds	1	1	1	1	1	1			6
Integral Roll 40Gr Units	1								1
Fourth Range Fresh Peeled Potatoes 1 Kg	1								1
Monalisa Washed Potato 1 Kg	1								1
Smooth Cayenne Pineapple Caliber-6 1 Kg	1					1			2
Frozen Chicken Thigh Fillet "W/Skin" 1 Kg						1			1
Semi-cured Cheese Mixed 1 Kg						1			1
Mini Spring Roll (20G) Units	1	1	1	1	1				5
Fresh Salmon 5/6 1 Kg (LOINS)						1			1
Frozen Salmon 1 Kg							1		1
Processed Cuttlefish 05-1 Kg Frozen 1 Kg						1	1	1	3
Total	14	8	8	8	8	12	4	4	

III. Results

The research produced intermediate results, some of which have already been described, although we will summarize the phases carried out.

Phase 0. Data set. It was focused on obtaining the appropriate data set and selecting the work environment in which to carry out the research.

- We selected the hotel's internal IT systems, Warehouse Management System, Supply Chain Management, Property Management Systems, Point Of Sale, and Finance System to obtain the necessary data. We used MSQL Server Management Studio, SoapUI, Postman, and our own programs made with XML and JSON to extract the data.
- Creation and implementation of the generic product code within the eICT to have comparable data over time.
- We incorporated a procedure to create and debug the data sets. We defined the format and notation for the files (CSV separated by ;) standard UTF8, encoding=ISO-8859-1.
- Validation of the adequacy of the HAC algorithms to develop the research based on a single algorithm with Euclidean metric.

1. Phase 1. Control Variables and Item Characterization

- The variables product consumption and product unit price are set and weighted as basic for the data sets, discarding the average purchase price variable.
- After applying the single algorithm with Euclidean metric to the data sets, we obtained the first version of the valuation table characterizing the algorithms. The table reflects the items: data set, number of variables used, number of algorithm clusters, number of clusters generated by default by the algorithm, number of elbow method clusters, cophenetic coefficient, maximum inconsistency value, maximum acceleration value, maximum cutoff distance, minimum cutoff distance, and BAEA number.

2. Phase 2. Selection of the Algorithm

New items were obtained for the evaluation table, the algorithm used, the number of clusters where acceleration shoots up, the difference between the maximum and minimum distance, and the outliers.

We selected the algorithm for each hypothesis, the average algorithm to control based on consumption, and the median for control based on cost.

3. Phase 3. Profile Variables

In this phase, we obtained the first global comparison between consumption and cost management hypotheses, identified in Table VII on a green and blue background, respectively. To do this, we defined new variables associated with customer profiles, including: number of people, meals included, and occupancy. We also defined new variables related to F&B turnover and the hotel. These new variables generated new characterization items: hypothesis used, lower and upper cutoff distances, number of lower cutoff clusters, maximum acceleration difference, number of visual clusters, and number of candidate product clusters in the segment.

The final result is presented in Table VII as a summary. The candidate products generated by the consumption hypothesis with the median algorithm are in green, and those generated from the cost hypothesis with the average algorithm are in blue. The rows identify the candidate products and the columns, the hypothesis from which it is obtained. The last row and last column show the respective totals.

Some data are worth pointing out. When we used all the variables, more candidate products were generated (columns 1 and 6). The type of product and the number of candidate products varied depending on the hypothesis and the variables used. In green and bold we see that there are products, such as round cupcakes and white muffins, that are identified from the consumption hypothesis regardless of the number of variables, but they are only identified from the cost hypothesis if all the variables are used. In blue and bold, we see products such as cod and shrimp that are identified from the cost hypothesis, but not for all the variables. In brown and bold, we have the special case of pineapple and melon, which are only identified when all the variables are used, regardless of the hypothesis.

IV. Discussion

The intermediate results conditioned the viability of the work and the subsequent phases, making it necessary to modify the business logic of the eICT and gain access to the raw data to generate the data warehouse.

The initial results showed it was necessary to compare products from different manufacturers throughout time. These products, which are equivalent to each other, had to be used as if they were the same product. Therefore, we implemented a generic product code with its

unit of measure for each product, which allowed us to standardize the products. This implementation was vital to the research because, without it, the work would not have continued since it would not be possible to systematically purchase the same products over time.

The criteria for the generation of the data sets were established, excluding those non-significant products and defining a format and nomenclature for the files used. These files had to be easily usable in the different work environments.

The work used the ability of HAC algorithms to cluster by focusing on the elements that made up the cluster and using them as candidate products. The need to objectively compare the algorithms, variables, and results led to the creation of a working method and different tables of results. These tables reflect the needs that had to be met to objectively assess the candidate products from the business logic point of view. From this perspective, the panel of experts established the criteria that the candidate products had to meet, making those criteria part of the company's know-how. Throughout the study, the requirements established by the GDPR for the study of customer profiles were met.

The results reflected in the assessment tables reflect a double perspective: that of the panel of experts who, using business logic, established the criteria for the the candidate products and who assessed the results that became part of the company's know-how; and the one associated with business intelligence that was conditioned by the data available in the eICT and had to comply with the RGPD regulating customer profiles.

Compared to other works that address problems in the F&B area through food cost rates for a period of time or costs associated with hotel management, this work considered control from the product consumption management hypothesis and compared it to cost. Control based on consumption was more useful as it provided weekly details of raw material consumption according to the hotel's customer profiles. This made it possible to implement business processes quickly, improving customer management and satisfaction.

Selecting the algorithm closest to each hypothesis led us to expand the assessment items in each phase to align the HAC algorithms with the BAEA. The different algorithms gave different results, but all with a certain degree of validity, so it was not possible to speak of only one solution. There were factors, such as the type of distance used by the algorithm or the inconsistency value we set, that varied the results and represent possibilities for future work.

Carrying out the work in phases and including new items in each one made it possible to identify candidate products of higher quality and closer to customer profiles and consumption. This helped improve management and could lead to studying new variables.

When comparing management from consumption versus cost, Table 7 shows how management from the perspective of consumption identifies more candidate products, and how management from cost identifies those with the highest price per unit. Separating consumption management from financial management helps companies better control consumption and favors the identification of candidate products. This helps control products that are risky for the company and allows problems such as consumption and price fluctuations to be addressed more quickly than with other conventional cost-control systems. Developing a system that recommends candidate products based on business logic could be a future project.

The summary in Table VII shows the candidate products generated by both approaches. In Tables V, VI, and VII, the numbers correspond, and the hypotheses are differentiated by color. The data obtained from consumption with the median algorithm has a green background, and the data obtained from the average cost algorithm has a blue background.

V. Summary and Conclusions

The work proposes and validates an alternative system for managing the costs of raw materials based on consumption. To do this, different ML tools were evaluated to support a company's panel of experts in their decision-making. Specifically, the capabilities of HAC algorithms were used to generate clusters in an unsupervised way based on the similarities and differences between the elements. This produced candidate products that were studied against conventional analysis systems. The research was carried out during three years, using data from four years of the company's eICT.

For implementation, it was necessary to modify the company's eICT and provide it with the necessary items to feed the data warehouse. We had the collaboration of a panel of experts from the areas involved and a suitable tool, Anaconda, as a Python distribution.

To analyze whether cost control from consumption is feasible and comparable to financial control, it is necessary to study the variables and data associated with weekly product consumption control, customer profiles, and financial production in each phase of the study. Having data did not imply that they were adequate. It was necessary to validate them.

The main point of the study was to verify that the initial results generated by ML with the starting data sets allowed business experts to identify possible candidate products and thus help improve their business logic. For this reason, we refined the data set and defined the starting point to evaluate the algorithms using a table that included the results obtained at each moment. The subsequent study of each algorithm and the different data sets led to the expansion of the items evaluated and the characterization table. We could then establish the base algorithm for each hypothesis, the average algorithm for the consumption hypothesis, and the median algorithm for the cost hypothesis.

The evolution of the research led to studying how the variables associated with customer profiles and financial production influenced the results. To do this, the base algorithms of each hypothesis were analyzed with new data sets and this led to the inclusion of new items in the assessment table.

The comparison of the results of the consumption hypothesis and the cost hypothesis reflects that more candidate products are suggested from the consumption management perspective, and this makes it easier for the experts to replace some products with others that are more suitable at any given time. On the other hand, cost-based management is not as versatile, although it clearly identifies the most expensive products. Additionally, the inclusion of variables associated with client profiles will depend on the analysis that is required at each moment.

The proposal presents advances in aspects of both science and business:

- The creation and validation of the generic product concept and the standard consumption measure to equate products with equivalent properties is crucial. This permits processes to be automated and time series to be studied and analyzed. The generic product concept was included with a standard in the software functionalities of a software package of the sector following its validation in this study.
- The use of HAC algorithms is valid as a support tool for decision-making in an area that is difficult to manage, such as F&B in a hotel.
- This study points the way to new work in AI by identifying points of inquiry, such as recommending systems.

At the same time, the inclusion of HAC algorithms in raw material management and control provides a differential value for:

- Facilitating the early detection of anomalies in management and cost by allowing detailed, transparent control.

- Improving the quality of service and reducing costs by valuing products from the new approach provided by HAC algorithms.
- Facilitating and speeding up the work of the people involved.

The application of ML is a viable alternative for the management of costs in companies from the control of product consumption versus classic control. Panels of experts play an important role when implementing the system, identifying items, and validating results. The use of variables linked to different profiles, such as consumption, customer profiles, and financial production allows candidate products to be obtained using a new approach. This new decision-making support scenario makes it easier for experts to identify the items and algorithms that best suit their needs at all times.

References

[1] S. Coleman, "Data science in industry 4.0," *progress in industrial mathematics at ecmi 2018*. Springer, New York City, New York, USA, Springer Publishing Company, 2019, pp. 559-566, doi.org/10.1007/978-3-030-27550-1_71.

[2] P. Foster, F. Tom, *Data Science for Business: What you need to know about data mining and data-analytic thinking*. Gravenstein highway north Sebastopol, USA: O'Reilly Media, Inc, 2013.

[3] C. Flath, N. M. Stein, "Towards a data science toolbox for industrial analytics applications," *Computers in Industry*, vol 94, pp.16-25, 2018, doi.org/10.1016/j.compind.2017.09.003.

[4] M.A Waller, A.; S.E. Fawcett, "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management," *Journal of Business Logistics*, vol. 34, no 2, pp. 77-84, 2013, doi.org/10.1111/jbl.12010.

[5] S. Athey, "The Impact of Machine Learning on Economics," *The economics of artificial intelligence,* University of Chicago Press, Chicago, IL 60637 USA, pp. 507-552, 2019, doi.org/10.7208/9780226613475.

[6] Y.S. Reshi, R.A. Khan, "Creating business intelligence through machine learning: An Effective business decision making tool," *Information and Knowledge Management*, vol 4, pp. 65-75, 2014.

[7] M. Gopal, *Applied machine learning*, New York, USA, McGraw-Hill Education, 2018.

[8] S. Finlay, *Artificial intelligence and machine learning for business: a no-nonsense guide to data driven technologies*, Lancaster, UK, Lancaster University, 2021.

[9] K. R. Larsen, S. Becker, Automated *machine learning for business,* Oxford, UK, Oxford University Press, 2021.

[10] L. B. Akeem, "Effect of cost control and cost reduction techniques in organizational performance," International Business and Management, vol. 14, no 3, pp. 19-26, 2017, doi.org/10.3968/9686.

[11] Y. Hamuro, et al, "A machine learning algorithm for analyzing string patterns helps to discover simple and interpretable business rules from purchase history," *Progress in Discovery Science,* Springer, Berlin, Germany, pp. 565-575, 2002. doi.org/10.1007/3-540-45884-0_43.

[12] Z. Guo, "Research on the cost control with hotel operation system based on cost management theory," *Journal of Computational and Theoretical Nanoscience,* vol. 13, no 12, pp. 9882-9885, 2016, doi.org/10.1166/jctn.2016.5945.

[13] Q. Y. Yan, H.J. Shen, "Assessing hotel cost control through value engineering: A case study on the budget hotels in a middle-sized city in China," *Asia Pacific Journal of Tourism Research,* vol. 21, no 5, pp. 512-523, 2016, doi.org/10.1080/10941665.2015.1063521.

[14] Z. Wu, P.D. Christofides, "Economic machine-learning-based predictive control of nonlinear systems," *Mathematics*, vol. 7, no 6, pp. 494. 2019, doi.org/10.3390/math7060494.

[15] E. Cengiz, et al, "Do food and beverage cost-control measures increase hotel performance? A case study in Istanbul, Turkey," Journal of Foodservice Business Research, vol. 21, no 6, pp. 610-627, 2018, doi.org/10.1080/15378020.2018.1493893.

[16] M. H. Rafiei, H. Adeli, "Novel machine-learning model for estimating construction costs considering economic variables and indexes," *Journal of construction engineering and management*, vol. 144, no 12, pp. 04018106, 2018 , doi.org/10.1061/(ASCE)CO.1943-7862.0001570.

[17] S. NosratabadiI, et al, "Data science in economics: comprehensive review of advanced machine learning and deep learning methods," *Mathematics*, vol. 8, no 10, pp. 1799. 2020, doi.org/10.3390/math8101799.

[18] J. Sun, "Analysis on Cost Control in Hotel Financial Management," *Destech Transactions on Social Science, Education and Human Science,* Huhhot, China, 2017, doi.org/10.12783/dtssehs/ssme2017/13011.

[19] A. Arbelo, P. Pérez-gómez, M. Arbelo-pérez, "Cost efficiency and its determinants in the hotel industry," *Tourism Economics*, vol. 23, no 5, pp. 1056-1068, 2017, doi.org/10.1177/1354816616656419.

[20] S. Coleman,et al, "How can SMEs benefit from big data? Challenges and a path forward," *Quality and Reliability Engineering International*, vol. 32, no 6, pp. 2151-2164, 2016, doi.org/10.1002/qre.2008.

[21] L. Oliveira, A. Fleury, M.T. Fleury, "Digital power: Value chain upgrading in an age of digitization," *International Business Review*, vol. 30, no 6, pp. 101850, 2021, doi.org/10.1016/j.ibusrev.2021.101850.

[22] O. D. Kazakov, et al, "Development of the concept of management of economic systems processes through construction and calling of machine learning models," *IEEE International Conference-Quality Management, Transport and Information Security, Information Technologies*-IEEE, Saint Petersburg, Russia, pp. 316-321, 2018, doi.org/10.1109/ITMQIS.2018.8524985.

[23] J. L. José, A.V. Borja, "*La adaptación al nuevo marco de protección de datos tras el RGPD y la LOPDGDD,*" Wolters Kluwer, Madrid, España, 2019.

[24] C. Batini, et al, "Methodologies for data quality assessment and improvement," *ACM computing surveys*, vol. 41, no 3, pp. 1-52. 2009, doi.org/10.1145/1541880.1541883.

[25] E. Parra, et al, "A methodology for the classification of quality of requirements using machine learning techniques," *Information and Software Technology*, vol. 67, p. 180-195, 2015, doi.org/10.1016/j.infsof.2015.07.006.

[26] Y. Jin, B. Sendhoff, "Pareto-based multiobjective machine learning: An overview and case studies," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 38, no 3, pp. 397-415, 2008, doi.org/10.1109/TSMCC.2008.919172.

[27] J. C. Chen, et al, "Off to the races: A comparison of machine learning and alternative data for predicting economic indicators," *Big Data for 21st Century Economic Statistics,* University of Chicago Press, Chicago, USA, 2019.

[28] N. Azarenko, "The model of human capital development with innovative characteristics in digital economy," *IOP Conference Series: Materials Science and Engineering,* IOP Publishing 2020, St. Petersburg, Russian Federation, pp. 012032, doi.org/10.1088/1757-899X/940/1/012032.

[29] E. G. Mitchell, et al, "From Reflection to Action: Combining Machine Learning with Expert Knowledge for Nutrition Goal Recommendations," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* pp. 1-17, Yokohama, Japan, doi.org/10.1145/3411764.3445555.

[30] K. D. Roe, et al, "Feature engineering with clinical expert knowledge: A case study assessment of machine learning model complexity and performance," *PloS one,* vol. 15, no 4, pp. e0231300, 2020, doi.org/10.1371/journal.pone.0231300.

[31] J. L. Loyer, et al, "Comparison of machine learning methods applied to the estimation of manufacturing cost of jet engine components," *International Journal of Production Economics,* vol. 178, pp. 109-119, 2016, doi.org/10.1016/j.ijpe.2016.05.006.

[32] H. Ahmed, et al, "Establishing standard rules for choosing best KPIs for an e-commerce business based on google analytics and machine learning technique," *International Journal of Advanced Computer Science and Applications,* vol. 8, no 5, pp. 12-24, 2017, doi.org/10.14569/ijacsa.2017.080570.

[33] F. Sánchez, Y. Hassan-Montero, "Visualization Design Dimensions for Data Science in Tourism and Transport," *Multidisciplinary Digital Publishing Institute Proceedings*. 13th International Conference on Ubiquitous Computing and Ambient "Intelligence UCAmI 2019, Toledo," Castilla la Mancha, Spain 2019, pp. 58, doi.org/10.3390/proceedings2019031058.

[34] A. Cook, P. Wu, K. Mengersen, "Machine learning and visual analytics for consulting business decision support," *2015 Big Data Visual Analytics (BDVA),* Hobart, Tasmania, Ausatralia, 2015, pp. 1-2, doi.org/10.1109/BDVA.2015.7314299.

[35] I. Lee, Y.J. Shin, "Machine learning for enterprises: Applications, algorithm selection, and challenges," *Business Horizons*, vol. 63, no 2, pp. 157-17, 2020, doi.org/10.1016/j.bushor.2019.10.005.

[36] Python Software Foundation. Python Language Reference, version 3.1. Available at http://www.python.org.

[37] P. Berkhin, "A survey of clustering data mining techniques," *Grouping multidimensional data,* Springer, Berlin, Heidelberg, Germany, pp. 25-71, 2006, doi.org/10.1007/3-540-28349-8_2.

[38] A. Fernandez, J. Preciado, A. Prieto, F. Sánchez-Figueroa, J. Gutiérrez, "Compare ML: A Novel Approach to Supporting Preliminary Data Analysis Decision Making," *International Journal of Interactive Multimedia and Artificial Intelligence*, 2021, doi.org/10.9781/ijimai.2021.08.001.

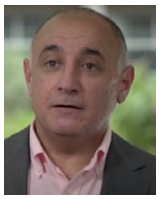
[39] S. Balakrishna, et al, "Incremental hierarchical clustering driven automatic annotations for unifying IoT streaming data," *International Journal Of Interactive Multimedia And Artificial Intelligence*, 2020, doi.org/10.9781/ijimai.2020.03.001.

[40] A. A Navarro, P. M. Ger, "Comparison of clustering algorithms for learning analytics with educational datasets," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 2, pp. 9-16, 2018, doi.org/10.9781/ijimai.2018.02.003.

[41] The SciPy community, SciPy documentation, https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

[42] P. Talón-ballestero, et al, "Using big data from customer relationship management information systems to determine the client profile in the hotel sector," *Tourism Management*, vol. 68, pp. 187-197, 2018, doi.org/10.1016/j.tourman.2018.03.017.

[43] C. Kim, K. Chung, "Measuring Customer Satisfaction and Hotel Efficiency Analysis: An Approach Based on Data Envelopment Analysis," *Cornell Hospitality Quarterly*, 2020, doi.org/10.1177/1938965520944914.



Fulgencio Sánchez Torres

Fulgencio Sánchez Torres is currently CIO at Garza Real Hotels. He is a doctoral student in computer science at the University of Alicante and an approved collaborator of the School of Industrial Organization. He has a Master's degree in Big Data and massive data visualization from UNIR and a degree in computer science. He has been an external expert for ICUAL of the Ministry of Education of Spain. He is a researcher at the ICT Technology Center, a researcher at the University of Castilla La Mancha, and at the National Technological University of Argentina, where he conducted a project financed by the International Cooperation Agency, Ministry of Foreign Affairs.

Iván González

Iván González is assistant professor and researcher at the Castilla-La Mancha University (UCLM). He received his M.Sc. (2015) and Ph.D. (2018) degrees in Advanced Computer Technologies from the same University. Member of the MAmI (Modelling Ambient Intelligence) Research Group since 2013, Dr. González has been involved in several research and development of International and National projects and contracts. He has participated in International conferences with 18 publications to date and he is author of 11 JCR research contributions. Member of research networks and scientific platforms related to Ubiquitous Computing and Ambient Intelligence (UBIHEALTH, AIAm, RedAmITIC and GITCE-UTP). He is currently performing research efforts focused on Quantitative Gait Analysis (QGA), Frailty assessment and Mild Cognitive Impairment (MCI) screening through mobile technologies and embodied sensors. His research interests also include Ubiquitous Computing, Smart Health, Smart Environments, Artificial Intelligence, IoT and Sensor Networks. Dr. González has years of experience organizing International conferences and R&D+I activities being one of the main organizers of UCAmI annual conference (since 2014). He has participated in the scientific committee of International conferences (7) and as a regular external reviewer of impact journals from research publishers (MDPI, Springer, Hindawi, SAGE, etc.). Also, he has been guest editor of 2 JCR-indexed special issues and Volume Editor of the UCAmI 2019 MDPI Proceedings. In the Educational field, Dr. González is coordinator of the Computer Engineering Degree at UCLM. His teaching covers the following subjects: Programming Fundamentals I and II, Operating Systems, Concurrent and Real-Time Programming, Multimedia and Human-Computer Interaction.

Cosmin C. Dobrescu

Cosmin C. Dobrescu is PhD candidate in the MAmI research group at the University of Castilla–La Mancha (UCLM). He completed the master's degree in Systems and Control Engineering in 2021 at the National University of Distance Education (UNED). He has also a degree in Computer Engineering with a specialization in computing in 2018 at the UCLM. Constantin obtained a competitive contract as a research support technician in the WeCareLab laboratory at the Institute of Information Technologies and Systems (ITSI), co-financed by the Fondo Social Europeo AEI 2018 (PEJ2018) in 2019. His teaching in 2021/22 includes the practices of Interactive Systems Design subject in the Degree in Computer Engineering. He is primarily interested in the design and development of IoT wearable medical devices aimed at prevention and rehab. His research is specialized in the development of firmware for IoT devices with a variety of sensors and integrated System on a chip SoC. The main challenge in this technology is to achieve maximum energy efficiency when acquiring, preprocessed and storing the generated data. By analyzing data using artificial intelligence, conclusions can be drawn on how to prevent diseases or improve the rehabilitation of people.