Data science methodologies selection with hierarchical analytical process and personal construction theory

Karina B. Eckert [1-2] and Paola V. Britos[3]

[1] National University of Misiones, Posadas, Misiones, Argentina
[2] Gastón Dachary University, Posadas, Misiones, Argentina
[3] Applied Computer Lab, National University of Río Negro, El Bolsón, Rio Negro, Argentina
karinaeck@gmail.com, pbritos@unrn.edu.ar

Abstract. The amount of data currently available for Strategic Decision Making is substantial; which is why Data Science find itself in apogee in various areas where it can be applied. Expertise respecting the areas' methodologies is fundamental; which is why, the objective of this paper is to compare and ponder them, for which, Analytic Hierarchy Process, was utilized along with linguistic tags and Personal Construction Theory, with the purpose of establishing and prioritizing characteristics according to their degree of compliance in real validation cases. The sub-criteria were grouped in different levels, conforming a hierarchy for the present problem. The validation case consisted in determining causes for breakdowns in new automobiles as they are being transported from the factory to the concessionaires; in which the proposed model proved useful and MoProPEI could be identified as the most adequate methodology.

Keywords: Data Science Methodologies, Analytic Hierarchy Process, Personal Construction Theory, Linguistic tags, Criteria.

1 Introduction

Multiple Criteria Decision Making (MCDM) can be seen as a useful tool for Decision Making (DM) and of great potentiality for Systems Engineering (SE) processes. There exists a superposition on the multicriteria and systemic approaches on a conceptual and operational level. At a conceptual level, when the established objectives begin to conflict with each other and it is required to find an equilibrium or compromise. At an operational level SE can be understood as a sequence of steps in which it is necessary to evaluate and choose among different alternatives or criteria at all times. These methods allow to approach a problem subjacent of subjectivity in an organized and systemic manner, which helps rationalize a complex process. [1], [2].

A popular MCDM method is Analytic Hierarchy Process (AHP), created by Saaty [3], with the purpose of searching for a systemic practice to define priorities and support complex DM [4]. The advantages of using AHP lie, among other aspects, in that among MCDM techniques, it is one of the few that provides a theoretical axiomatic; from a practical point of view, it is characterized by its good performance; providing a flexible, adaptable, robust and easy to understand model [1], [5], [6].

In some scenarios decision makers have a very limited amount of information to specify their preferences on multiple pair comparisons; which is why a deeper analysis is required rather than a direct comparison [7]. Assessing and selecting Data Science (DS) methodologies is one of these scenarios.

Despite the fact that the majority of DS methodologies have been evaluated and validated by the community, these are not without flaw, for example the ones related to project management [8], [9]. Selecting a methodology can be complicated, especially for novices; while this is an essential task for experts in the area [10], [11].

The objective of this article is to determine which of the assessed DS methodologies (P^3TQ [12], CRISP-DM [13] y MoProPEI [14]) is the most robust for real applications. In order to achieve this, the first three stages of Personal Construction Theory (PCT) are used, which initially include a dialogue to determine how the expert thinks, and to identify which are their priorities and most important factors [15], [16], from that, linguistic tags are established, the sub-criteria involved in DS methodologies are defined y and the hierarchic structure of the problem is conformed. Sub-criteria are established based on a degree of compliance function, depending on the validation case, which is then integrated to AHP, to finally obtain the resulting ponderations for each methodology.

The present article is structured in the following way: Preliminary concepts referred to DS, AHP and PCT can be found in Section 2. Posteriorly the proposed model is described in Section 3; which is validated using a real-life case in Section 4. Finally, conclusions are presented in Section 5

2 Preliminary concepts

2.1 Data Science

Currently, informatics systems can generate and store a vast amount of data at a low cost ; which results in these growing exponentially and making them impossible to process using common methods [10], [11]. A substantial amount of attention must be given to the importance and implications of data for DM; given that they are a great advantage for it [10]. In consequence, there is an increasing number of companies that take decisions based on data, improving their performance in an operational and financial way [17].

Data Science was previously referred to as Data Mining or Information Mining; which over time changed its designation as it grew. The concept in this case is the extraction of knowledge from data and technology that incorporate these principles [11], [18].

DS consists in a group of fundamental principles, guided by a specific methodology, which help and guide the extraction of knowledge from data; it includes several techniques, algorithms and tools which ease the exhaustive and automatic processing of data; allowing to identify useful knowledge which is not possible to be detected in plain sight [10], [11], [18]. In order to forecast results, areas such as statistics, math, behavioral science, computing and predictive analysis are included [19]. The objec-

tive is to obtained knowledge specialized for DM, from results represented as models or patterns [11], [20], [21].

There exists a diverse variety of techniques and algorithms for data processing and knowledge extraction applicable in this area; however, DS involves much more. DS provides professionals a structure and a group of principles that bring a framework in order to systematically treat knowledge extraction problems; where methods to treat data and methodologies utilized in these projects are transcendental [10], [18].

An amount of studies comparing existing methodologies exists shown in [20], [21], [22], [23], [24] y; from which methodologies such as Catalys (known as P3TQ) [12] and CRISP-DM [13] can be highlighted. Based on different studies, recommendations and our own study of the methodologies; the aforementioned where selected along with MoProPEI [14].

2.2 Analytic Hierarchy Process

AHP involves the following activities or steps [1], [5], [25], [26], [27], [28]:

Firstly, the decision problem must be modelled as a hierarchic structure. Situating the main objective in the upper level. In the level below, criteria such as attributes, secondary objective or parameters from which preferences are transformed or justified can be found. In some cases, criteria can be divided in sub-criteria forming another descending hierarchy. Finally, in the last level, alternatives are presented.

When establishing priorities using paired comparison, the objective is to define the relative weights for the criteria, said numeric values indicate the importance or relative priority between Ci and Cj as criteria, respecting the element in the immediately superior level. In order to achieve this, the fundamental scale proposed by Saaty was utilized; which ranges from an equivalent importance (value 1), with two criteria that contribute equally to achieve the objective, to an extreme degree of importance (value 9) where the evidence that benefits one criteria over another is the highest possible in the affirmation order. Numbers in the scale represent the importance proportion of an element respect another in relation to criteria or an objective which they share.

Making use of the fundamental scale, the decision maker must determine the assigned weight for each criterion, completing for this the matrix. For a matrix of these characteristics it is true that the maximum eigenvector λ_{max} is a positive real number and that there exists an eigenvector Z, which elements associated to this vector are positive. Posteriori the eigenvector must be normalized so that the summation becomes a unit.

An incoherence error in the pairing comparison process generates a matrix and an eigenvector which are unrepresentative; which in turn results in a contradiction since it violates the transitivity of the values, in order to correct this Saaty proposed the Consistency Ratio(CR) in order to evaluate coherence in the decisions made by the decision maker, which is shown in equation 1:

$$CR = CI/RI \quad (1)$$

CI is defined in equation 2, where λ_{max} is the maximum value in the matrix and n and its order. RI is a measurement utilized to improve the consistency of the decisions

accounting for the dimension of the matrix. Simulating 100.000 randomly generated inverse matrixes [29] , average RI was defined. For adequate consistency, Saaty indicated that CR must not be greater than 10% (CI <= 0,10); the closer to 0, the greater the consistency; in the opposite case, decisions must be further revised.

$$CI = (\lambda_{max} - n)/(n-1) \quad (2)$$

If there exist sub-criteria, their global weight must be calculated a priori associated to them; following the same procedure, but in this case the paired comparisons must be performed, in order to determine the relative importance to the criteria immediately above in the hierarchy (local priority). To calculate relative global importance, the product of the different weights of each one of the sub criteria and criteria is calculated, following the hierarchy from the most inferior part to the top of this one, this procedure is known as "Hierarchic composition"

Saaty proposed the use of a method known as pondered summation; which consists in finding the global priority vector p, which adds priorities obtained from criteria and alternatives. The pi components in the vector belong to total priorities associated to each alternative Ai, reflecting the total value which the decision maker has for each alternative, for the aforementioned, the following expression can be used: (3):

$$p_i = \sum_{j=1}^{n} (w_j \, . \, r_{ij}) \quad i = 1, 2, \dots, m \quad (3)$$

Where w_j corresponds to the associated weights from each of the considered criteria and r_{ij} are the components of the normalized matrix. In order to solve the decision problem and to determine the best alternative which will be the greater pondered summation, sorting the alternatives based on these values will be enough.

2.3 Personal Construction Theory

Personal Construction Theory (PCT) was proposed by Kelly, it is a technique for the extraction of knowledge; which consists in becoming aware of the inconsistencies in the value scales, given that each person has their own view of the world. For a particular domain all of the aspects that the expert finds important must be included and must be represented as elements and their decomposition in bipolar characteristics; which are then evaluated. It is considered as a classification test divided in five stages [15], [16].

3 Proposed Model

The proposed model is divided in a series of steps:

1. Characteristics recognition

Based on methodology studies and experts' opinion, the criteria, sub-criteria and their hierarchic structure where selected. The first two stages of PCT were utilized in order to work with the experts: First, the elements are identified (DS methodologies) and secondly the characteristics (criteria y sub-criteria).

2. Linguistic tag definition from completeness level

The purpose of this step is to avoid ambiguities in the definition and the completeness of the criteria, a narrow scale ranging from 1 to 9 was utilized, where 1 indicated that the sub-criteria shouldn't be analyzed; 2 to 9 represent interval values in the form of percentages respecting the fulfillment of this aspect inside the case study. Depending on if the sub-criteria is a positive or negative aspect, the values in the scale defined between 2 and 9 can be inverted. For example, two sub-criteria are exposed in Table 1, the first one being Portability which is a positive aspect, the values(percentages) are positive and in an ascendant way, whereas for Transformation Costs, which is a negative aspect, the percentages are shown in descendent way.

Table 1. Linguists tags for sub-criteria

Portability		**Transformation Cost**	
1	Not analyzed	1	Not analyzed
2	1% to 13% portability	2	98% to 100% transformation cost
3	14% to 27% portability	3	84% to 97% transformation cost
4	28% to 41% portability	4	70% to 83% transformation cost
5	42% to 55% portability	5	56% to 69% transformation cost
6	56% to 69% portability	6	42% to 55% transformation cost
7	70% to 83% portability	7	28% to 41% transformation cost
8	84% to 97% portability	8	14% to 27% transformation cost
9	98% to 100% portability	9	1% to 13% transformation cost

3. Establishing a hierarchic structure

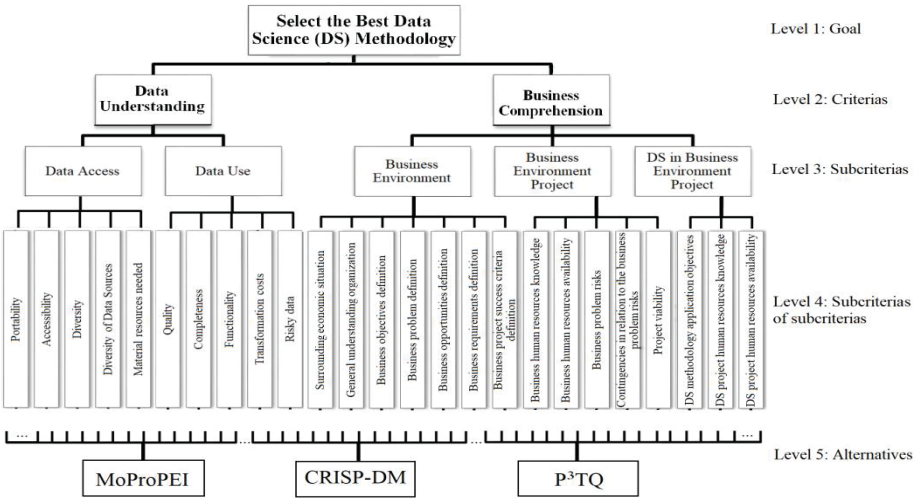

Fig. 1. Hierarchical structure to select the best DS methodology

The aforementioned steps allowed to identify the criteria and sub-criteria that must be taken into account when selecting a DS methodology; which were grouped in different levels, therefore establishing a hierarchy for the problem. In Fig.1. the obtained structure can be visualized; the main objective can be observed in the first level (Selecting the best DS methodology); the second level is where the two main branches from the methodologies and DS projects, the understanding of the data and the business comprehension are defined; in the third level data access and data usage can be found, with 5 sub-criteria for each in level 4 on one side; and on the other, business environment, project environment and DS in the business project on the third level, with seven, five and three sub-criteria respectively (level 4); the sub-criteria in the last level are compared based on each of the methodologies visible in level 5.

4. Making arrays with the sub-criteria in the fourth level

For this step a grill-type matrix was made (PCT Stage 3), for which were set the extreme or bipolar values, the worst and best case, which represent values 1 and 9(See Table 1). Each expert filled the matrix with the corresponding values taking into account the previously established scale. For example, sub-criteria with their bipolar values in each extreme and their respectively assigned ponderation for each methodology can be found in Table 2.

Table 2. Compliance last level sub-criteria

	P^3TQ	**CRISP-DM**	**MoPro PEI**	
Data portability is not analyzed	8	5	9	98% to 100% data portability
Data diversity is not analyzed	3	4	2	0% to 13% data diversity

5. Paired comparison for criteria and sub-criteria

Starting from the values indicated by the expert in the previous step, matrixes were completed based on the sub-criteria in the fourth level; taking the difference between the assessments as absolute values, plus one (Ex: 5-5=0+1, both sub criteria have the same level of preference; 7-5=2+1, the first sub-criteria has a preference of 3 over the second one). The purpose of this was to complete the paired matrixes based on assessed matrixes with linguistic tags (see example in Table 2) fitting the values of the Saaty scale and placing them in their corresponding places inside the matrix

For the criteria in levels 2 and 3, other paired-comparison matrixes where made in order to ease the expert's choices; where the preference value respect another is marked with an X (based on the Saaty scale). The aforementioned is seen in Table 3, where the comparison between the criteria for data access and data usage is exemplified; for which the expert assigned a ponderation of 5, that is to say, that data usage possesses a great importance over data access. From these matrixes the corresponding ones where completed according to what Saaty proposed, defining their importance grouping them by criteria and sub-criteria based on the defined hierarchy.

Table 3. Criteria by pairs comparison

Data Access					**Data Use**			
Extreme importance: 9	Very strong importance :7	Strong importance :5	Moderate importance: 3	Equal importance: 1	Moderate importance: 3	Strong importance :5	Very strong importance :7	Extreme importance: 9
						X		

Posteriorly, their reciprocal values where incorporated, all the matrixes were normalized and the ponderations for each of them were defined. Table 4 shows the continuity of what was proposed in Table 3.

Table 4. Pairwise comparison matrix

	Data Access	**Data Use**	**Normalized Matrix**		**Weightings**
Data Access	1	1/3	0,25	0,25	0,25
Data Use	3	1	0,75	0,75	0,75

6. Coherence control

The previous step was followed by an assessment of the consistency of the decisions made by the experts based on the paired matrixes from levels 2 to 4; for this, the quotient for the matrixes was estimated along with the approximations of consistency for each of them, as indicated in equations 1 and 2 (Section 2).

7. Establishing final ponderations

The local and global priorities were established using the relative weight of the criteria for each level; this was followed by obtaining the total priorities associated to each alternative; using pondered summation (See equation 3).

4 Validation Cases

In order to verify the proposed model, two real validation cases were utilized, the first one has the objective of determining random breakdowns in new automobiles as they are being transported from factories to concessionaires and the second one, causes for college desertion. The results obtained for the first scenario are summarized below; for which experts chose the values when analyzing each methodology.

Fig. 2 shows the ponderations obtained for each methodology respect from the subcriteria in the third level. In this way, the preponderations for MoProPEI can be appreciated over the other two for these sub-criteria; it can also be noted that for the understanding of data, P3TQ obtained a better performance compared to CRISP-DM, however this was not the case in business comprehension, where in a general way CRISP-DM obtained better weights.

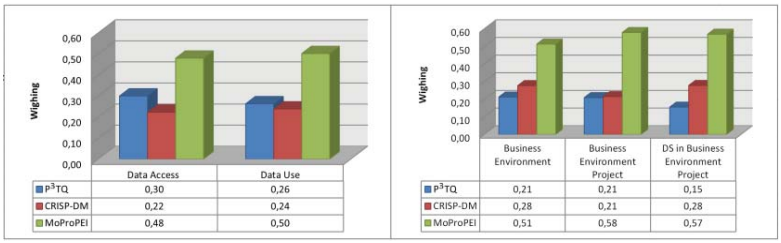

Fig. 2. Third level hierarchy weightings

Inside the understanding of data, the sub-criteria for data access have a ponderation of 0,25 whereas data usage has 0,75 respect from this criterion. Concerning business comprehension, the weights assigned to sub-criteria for business environment and project environment are of 0,20 each and for DS in business project it is 0,75 respect from this criterion. Taking into account the aforementioned ponderations and the ones obtained in level 3, the criteria located in the second level are shown in Table 5.

Table 5. Second level hierarchy weightings

	Data Understanding	**Business Comprehension**
P^{3}TQ	0,07	0,11
CRISP-DM	0,17	0,17
MoProPEI	0,26	0,23

Going up in the hierarchy taking into account that ponderations in inferior levels and that Data Understanding and Business Comprehension are equally important; Fig. 3. Shows the final ponderations obtained, which show that MoProPEI obtained a ponderation of 53%, followed by CRISP-DM with 25% and lastly P3TQ with 22%. This clearly indicates that given the assessed criteria and sub-criteria, MoProPEI is the most adequate and complete for the present case a, obtaining a ponderation larger than the other two combined.

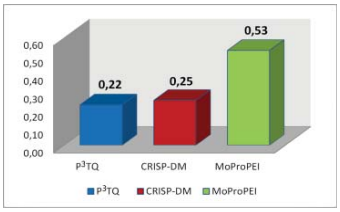

Fig. 3. Final weightings by methodologies

5 Conclusions

Based on the proposed model, the analysis and the results, it can be highlighted that the model that integrates AHP, linguistic tags and PCT is beneficial to identify the

primordial characteristics inside a DS project; as well as to establish de compliance level for each of the sub-criteria of each methodology using defined linguistic tags and grill-type matrixes used to complete these aspect therefore avoiding ambiguity. It was proven that these techniques can be integrated to obtain positive results.

Concerning determining causes for breakdowns in automobiles as they are being transported from factories to concessionaries, the proposed model obtained the ponderations for each methodology, where MoProPEI can be highlighted over the remaining two, being this one the selected one for the aforementioned validation case. For the other validation case, the results were similary.

As future research it is expected to develop software that implements the proposed model, validation for new DS project and even other areas.

References

1. García Cascales, M. del S.: Métodos para la comparación de alternativas mediante un Sistema de Ayuda a la Decisión (S.A.D.) y "Soft Computing," (2009).
2. Romero, C.: Análisis de las Decisiones Multicriterio. Isdefe, Madrid, España (1996).
3. Saaty, T.L.: The analytic hierarchy process. McGraw-Hill, New York (1980).
4. Forman, E.H., Gass, S.I.: The Analytic Hierarchy Process—An Exposition. Operations Research. 49, 469–486 (2001). https://doi.org/10.1287/opre.49.4.469.11231.
5. Russo, R. de F.S.M., Camanho, R.: Criteria in AHP: A Systematic Review of Literature. Procedia Computer Science. 55, 1123–1132 (2015). https://doi.org/10.1016/j.procs.2015.07.081.
6. Kou, G., Lin, C.: A cosine maximization method for the priority vector derivation in AHP. European Journal of Operational Research. 235, 225–232 (2014). https://doi.org/10.1016/j.ejor.2013.10.019.
7. Jalao, E.R., Wu, T., Shunk, D.: A stochastic AHP decision making methodology for imprecise preferences. Information Sciences. 270, 192–203 (2014). https://doi.org/10.1016/j.ins.2014.02.077.
8. Pytel, P., Britos, P., García Martínez, R.: Proposal and Validation of a feasibility Model for Information Mining Projects. Presented at the 25th International Conference on Software Engineering and Knowledge Engineering , Boston, USA.
9. Vanrell, J.Á., Bertone, R.A., García Martínez, R.: Modelo de proceso de operación para proyectos de explotación de información. Presented at the XVI Congreso Argentino de Ciencias de la Computación (2010).
10. Waller, M.A., Fawcett, S.E.: Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. Journal of Business Logistics. 34, 77–84 (2013). https://doi.org/10.1111/jbl.12010.
11. Eckert, K., Britos, P.V.: Modelo basado en la toma decisiones con criterios múltiples para la elección de metodologías de data science. Presented at the XX Workshop de Investigadores en Ciencias de la Computación (2018).
12. Pyle, D.: Business Modeling and Data Mining. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2003).

13. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0: Step-by-Step Data Mining Guide, http://tinyurl.com/crispdm, (2000).
14. Martins, S., Pesado, P., García Martínez, R.: Propuesta de Modelo de Procesos para una Ingeniería de Explotación de Información: MoProPEI. Revista Latinoamericana de Ingenieria de Software. 2, 313–332 (2014).
15. Britos, P., Rossi, B., García Martínez, R.: Notas sobre didáctica de las etapas de formalización y análisis de resultados de la técnica de emparrillado. Un Ejemplo. In: Proceedings del V Congreso Internacional de Ingeniería Informática. pp. 200–209 (1999).
16. García Martínez, R., Britos, P.V.: Ingenieria de Sistemas Expertos. Nueva Librería (2004).
17. McAfee, A., Brynjolfsson, E.: Big data: the management revolution. Harv Bus Rev. 90, 60–68 (2012).
18. Provost, F., Fawcett, T.: Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data. 1, 51–59 (2013). https://doi.org/10.1089/big.2013.1508.
19. Hazen, B.T., Boone, C.A., Ezell, J.D., Jones-Farmer, L.A.: Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics. 154, 72–80 (2014). https://doi.org/10.1016/j.ijpe.2014.04.018.
20. Rodríguez Montequín, M.T., Álvarez Cabal, J.V., Mesa Fernández, J.M., González Valdés, A.: Metodologías para la realización de proyectos de Data Mining. Presented at the VII Congreso Internacional de Ingeniería de Proyectos , Pamplona España (2003).
21. Moine, J.M.: Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo, http://hdl.handle.net/10915/29582, (2013).
22. Moine, J.M., Gordillo, S.E., Haedo, A.S.: Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. Presented at the XVII Congreso Argentino de Ciencias de la Computación (2011).
23. Giraldo Mejia, J.C., Jiménez Builes, J.A.: Caracterización del proceso de obtención de conocimiento y algunas metodologías para crear proyectos de minería de datos. Revista Latinoamericana de Ingeniería de Software. (2013).
24. Palacios, H.J.G., Pantoja, G.A.H., Navarro, A.A.M., Puetaman, I.M.A., Toledo, R.A.J.: Comparative between CRISP-DM and SEMMA for data cleaning of MODIS products in a study of land use and land cover change. In: 2016 IEEE 11th Colombian Computing Conference (CCC). pp. 1–9 (2016). https://doi.org/10.1109/ColumbianCC.2016.7750789.
25. Saaty, T.L.: Fundamentals of Decision Making and Priority Theory With the Analytic Hierarchy Process. RWS Publications (2000).
26. Saaty, T.L.: How to make a decision: The analytic hierarchy process. European Journal of Operational Research. 48, 9–26 (1990). https://doi.org/10.1016/0377-2217(90)90057-I.
27. Liu, B., Kong, F.: Research and application of sidewall stability prediction method based on analytic hierarchy process and fuzzy integrative evaluation method. Natural Science. 4, 142 (2012).
28. Saaty, T.L.: Decision making with the analytic hierarchy process. International Journal of Services Sciences. 1, 83–98 (2008). https://doi.org/10.1504/IJSSci.2008.01759.
29. Aguarón, J., Moreno Jiménez, J.M.: The geometric consistency index: Approximated thresholds. European Journal of Operational Research. 147, 137–145 (2003). https://doi.org/10.1016/S0377-2217(02)00255-2.