- ORIGINAL ARTICLE -

Proposed Extended Analytic Hierarchy Process for Selecting Data Science Methodologies

Propuesta del Proceso Analítico Jerárquico Extendido para la Selección de Metodologías de Ciencias de Datos

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

[1]*Faculty of Exact Sciences, Chemical and Natural, National University of Misiones, Posadas, Misiones, Argentina*

[2]*Department of Engineering and Production Sciences, Gastón Dachary University, Posadas, Misiones, Argentina*

karinaeck@gmail.com

[3]*Applied Computer Lab, National University of Río Negro, El Bolsón, Rio Negro, Argentina*

pbritos@unrn.edu.ar

Abstract

Decision making can present a considerable amount of complexity in competitive environments; where methods that support possess great relevance. The article presents an extension of the Hierarchy Analytical Process; complemented with Personal Construct Theory, which purpose is to reduce ambiguity when defining and establishing values for the criteria in a determined problem. In recent years, the scope for decision making based on data has considerably raised, which is why Data Science as a scientific field is rising in popularity; where one of the main activities for data scientists is selecting an adequate methodology to guide a project with this traits. The steps defined in the proposed model guide this task, from establishing and prioritizing criteria based on degrees of compliance, grouping them by levels, completing the hierarchical structure of the problem, performing the correct comparisons through different levels in an ascendant manner, to finally obtaining the definitive priorities of each methodology for each validation case and sorting them by their adequacy percentages. Both disparate cases, one referred to an industrial/commercial field and the other to an academic field, were effective to corroborate the extent of usefulness of the proposed model; for which in both cases MoProPEI obtained the best results.

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

Resumen

Los problemas de toma de decisiones son complejos en entornos competitivos; donde los métodos que ayudan a esta disciplina tienen gran relevancia. Este artículo presenta una extensión al proceso analítico jerárquico; complementado con la teoría de la construcción personal, con el propósito de disminuir la ambigüedad en la definición y valoración de criterios del problema. En los últimos años ha ido tomando mayor envergadura las decisiones tomadas a partir de los datos, es por ello que la ciencia de dato es una disciplina en pleno auge; donde una de las actividades principales de los científicos de datos es la elección de la metodología adecuada para guiar un proyecto de estas características. Los pasos definidos en el modelo propuesto guían esta tarea, desde el establecimiento y priorización de los criterios según el grado de cumplimiento, agrupándolos por niveles, completando la estructura jerárquica del problema, realizando las comparaciones pertinentes subiendo por niveles, hasta obtener las prioridades finales y ordenándolas según los porcentajes de adecuación de cada metodología para cada caso de validación. Ambos casos disimiles, uno referido al ámbito académico y otro al industrial/comercial, sirvieron para corroborar la utilidad del modelo propuesto; donde para ambos casos la metodología MoProPEI obtuvo mejores resultados.

Palabras claves: Etiquetas Lingüísticas, Metodologías de Ciencia de Datos, Proceso Analítico Jerárquico, Teoría de la Construcción Personal.

1. Introduction

Daily situations and contexts convey taking different types of decisions. Before taking any decisions, the context of the problem must be evaluated, then, the greatest amount of information available must be collected, as well as taking into consideration the knowledge and experience of the person entrusted with the task. Precisely, Decision Making (DM) problems are complex processes, where the person responsible of taking decisions must select among an several of alternatives, based on a set of criteria, from the problem itself; which conveys performing comparisons and appraisals of different aspects. In order to facilitate these processes, a set of tools must be used, said tools should satisfy, to the greatest degree possible, combining criteria and allowing selection of one of the alternatives without failure [1, 2].

Multiple Criteria Decision Making (MCDM) combine, in a generic manner, the performance of the alternatives of several qualitative and/or quantitative criteria and obtains a compromise solution as a result [3]. These methods are frequently applicable in numerous real-life problems, where several sets of alternatives for a decision are evaluated based on conflicting criteria [4].

One of the most utilized tools for MCDM by DM researchers is Analytic Hierarchy Process (AHP), proposed by Saaty [5]. Its popularity is due to its simplicity and performance; given that it is simple to understand and utilize in different scenarios, being docile enough for it; in turn it provides an axiomatic theory, which very few other methods provide [6, 7, 8, 9, 10, 11].

AHP as any other MCDM method, is not flawless, presenting both positive and negative aspects [6, 12]. One of its greatest inconvenients is the definition of criteria involved in complex problems, given their size, due to a lack of information or due to a certain amount of ambiguity presented by the criteria. The solution proposed for this inconvenient is integrating AHP with a knowledge elicitation theory called Personal Construction Theory (PCT) proposed by Kelly [13], which helps defining preferences and views for the experts.

Comparing and selecting methodologies may prove a difficult task, given certain inherent characteristics for each area; one notable case are Data Science (DS) methodologies; which are a primordial activity in these type of projects [14, 15, 16].

Albeit the scientific community, together with the industry, have validated several DS methodologies; these are not without flaw in terms of project management [17, 18].

The present article attempts to validate the proposed model comparing three DS methodologies [14, 15], in order to determine which proves to be more robust, for which real cases are utilized; said methodologies are P^{3}TQ [19], CRISP-DM [20] and MoProPEI [21].

For this, PCT is utilized to define criteria in different levels with varying detail and conforming the hierarchical structure of the problem itself; linguistic labels are also specified based on the level of compliance for the different criteria evaluated in each case study, to posteriorly combine this with AHP, in order to perform a comparison of the criteria in different levels and therefore calculate the definitive ponderations for each methodology in each validation case.

The proposal originality lies in the combination of techniques (AHP and PCT), which has not been used previously; also in the testing context, which refers to SD methodologies and the flexibility it offers to compare various alternatives.

The present article is structured in the following manner: Section 2 presents an introduction to AHP, PCT and DS. Section 3 describes the proposed model. The validation for said model can be found in Section 4. Finally, in Section 5, conclusions and future research are thoroughly explained.

2. Preliminaries

2.1. Analytic Hierarchy Process

AHP is an efficient tool for complex DM and can help the subject in charge of DM to establish priorities take the most effective decision. This is accomplished by decomposing a complex problem in a hierarchic structure composed of several levels of abstraction (objective or goal, criteria, sub-criteria and alternatives) [6, 22, 23, 24, 25]. AHP can be summarized in a set of steps [6, 7, 26, 27]:

1. Problem definition, determining the type of knowledge that is being sought and hierarchic structuralizing; which is constituted by identifying the main goal or objective, followed by criteria and, if existing, sub-criteria definition in the middle level and alternative definition in the inferior level; this structure is sufficiently precise and detailed to include the main concerns the subject in charge of DM might possess.

2. Establishing criteria priorities and building comparison matrixes by pairs: Quoting Saaty [26], every element in any given level is utilized to compare the elements in the most immediately inferior level respect to said element. The comparison is made utilizing a defined scale which purpose is to display the degree of dominance or importance of any given element respects another, in relation to any criteria or characteristic for which they are being compared. Said scale is verbal and it is utilized to measure quantitative and qualitative criteria; while paired comparisons make the analysis more precise.

3. Establishing local and global priorities for each level, by means of calculating the relative weight of each element for each level: This calculation is performed in a series of steps, which in a

summarized manner includes, posterior to completing and normalizing the paired matrix, obtaining relative priorities for all criteria, to then evaluate the consistency of the matrix; if this results to not be true, the previously performed comparisons must be revised and repeat the aforementioned process.

4. Establishing total priorities associated to each alternative: This step consists in calculating total priorities associated to each alternative, which represent the importance of each alternative respect the objective or main goal; the best alternative is the one with the highest priority value.

5. Sensitivity analysis, verification and decision balancing.

2.2. Personal Construction Theory

PCT is among the indirect knowledge elicitation methods, which collaborate with the expert in order to be specific respecting their knowledge or mental processes, which are usually not clear [28].

The main strength that PCT possesses is the capability of modelling the internal vision of the world of any given person, without the necessity of establishing explicitly what that vision might be. Kelly holds that each person views the world in a different manner and believes that said differences can be expressed as personal representations [13, 28, 29].

This is considered a classification test, divided in five steps [13, 28]: The first step consists in identifying the representative elements from the evaluated domain. The second step is performed by identifying the characteristics, such as an attributed quality to the unidentified elements in the previous step. The third one, lies in designing the grid, a bidimensional matrix, that binds every element and identified characteristics, where columns represent the elements and rows represent the characteristics; the matrix is then completed using a defined scale from 1 to N; bipolar values are then placed in every extreme on the rows, where value 1 is opposite of N; the expert then assigns the corresponding value to each cell on the matrix, taking into account the defined scale and the intersection, element and characteristic that are being evaluated. The fourth step consists in formalization, where elements and characteristics are classified separately. Lastly, results are analyzed and interpreted.

2.3. Data Science

Following the increasing and immense demand for data that currently exists, knowledge and domain analysis must not be separated. DS can be defined, as the application of quantitative and qualitative methods to predict results and solve relevant problems. Precisely, a set of fundamental principles are considered, which guide and support the extraction of knowledge from data; including diverse methodologies, techniques, algorithms and tools that facilitate advanced and automated data processing; allowing to identify relevant and strategic information that cannot be detected in plain sight [14, 16, 30, 31].

DS provides a frame to solve knowledge extraction problems in a systemic way; where the methods used to work with data and the methodologies utilized to carry DS projects forward are transcendental [15, 16, 30].

Based on previous work such as, [32, 33, 34, 35] and [36], from the study of methodologies, expertise and suggestions from experts in the area; the following methodologies were chosen: P^{3}TQ (Catalyst) [19], CRISP-DM [20] and MoProPEI [21].

P^{3}TQ

Catalyst, also known as P^{3}TQ (Product, Place, Price, Time, Quantity) is composed by two models: Business Model (BM) and Data Exploitation Model (DEM) [20]. The former consists in a series of steps to build a model that allows to identify a problem or business opportunity. The latter provides a series of steps for utilizing DEM based on the previously identified model (BM) [24].

CRISP-DM

CRISP-DM (CRoss-Industry Standard Process for Data Mining) consists in a set of tasks described in four levels of abstraction, organized in a hierarchic manner: phases, general tasks, specialized tasks and process instances [19, 24]. The life cycle of a project is organized in six phases: Business comprehension, Data comprehension, Data preparation, Data evaluation and implementation [23, 24].

MoProPEI

MoProPEI (Process Model for Information Exploitation Projects) possesses a hierarchic structure composed of four levels, which are: Subprocesses, Phases, Activities and Tasks. The two main subprocesses are Management, aimed for project control and administration, which consists in five phases: Project initialization, Project planning, Support, Quality and Control Management, Delivery Management; and the second one is Development, aimed to technical activities, it is composed of six phases: Domain Comprehension, Data Comprehension, Data Preparation, Implementation, Evaluation and Presentation [21].

3. Proposed Model

The proposed model is divided in a series of steps which can be seen in the Fig. 1.

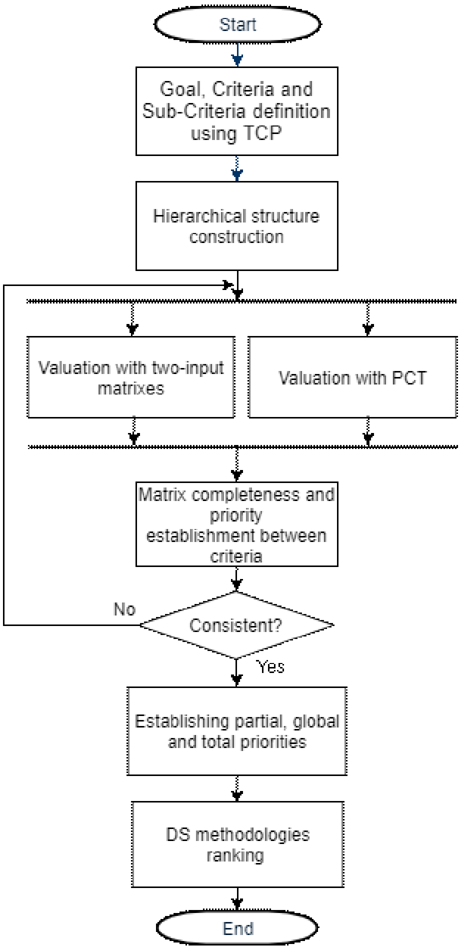
Start
Goal, Criteria and Sub-Criteria definition using TCP
Hierarchical structure construction
Valuation with two-input matrixes
Valuation with PCT
Matrix completeness and priority establishment between criteria
No
Consistent?
Yes
Establishing partial, global and total priorities
DS methodologies ranking
End

Fig. 1 Proposed model.

Step1. Goal, Criteria and Sub-Criteria definition using TCP

The first step consists in identifying the objective, criteria and sub-criteria which must be evaluated using each methodology. The aforementioned are established utilizing PCT working with experts in the area, the characteristics of all methodologies and previous studies

Step 2. Hierarchical structure construction

Following the previous step, utilizing PCT, the sub-criteria involved in DS methodologies are defined, to posteriorly be grouped by criteria in different levels, thus conforming the hierarchic structure seen in the proposed model, which can be seen in Fig. 2.; where level 1 contains the main goal or objective, which consists in selecting the most robust Data Science methodology from all three compared alternatives. The second level contains the main criteria inside projects with these characteristics; they are Data Comprehension (DC) and Business Comprehension (BC).

Subsequently, in the third level, the first criteria (DU) is decomposed in two sub-criteria which are data access and use of data; which are specified with five sub-criteria each; for Data Access the sub-criteria are Portability, Accessibility, Diversity, Data Source Diversity and Necessary Resources; and for the sub-criteria Data Usage, Quality, Completeness, Functionality, Transformation costs and Data Risks.

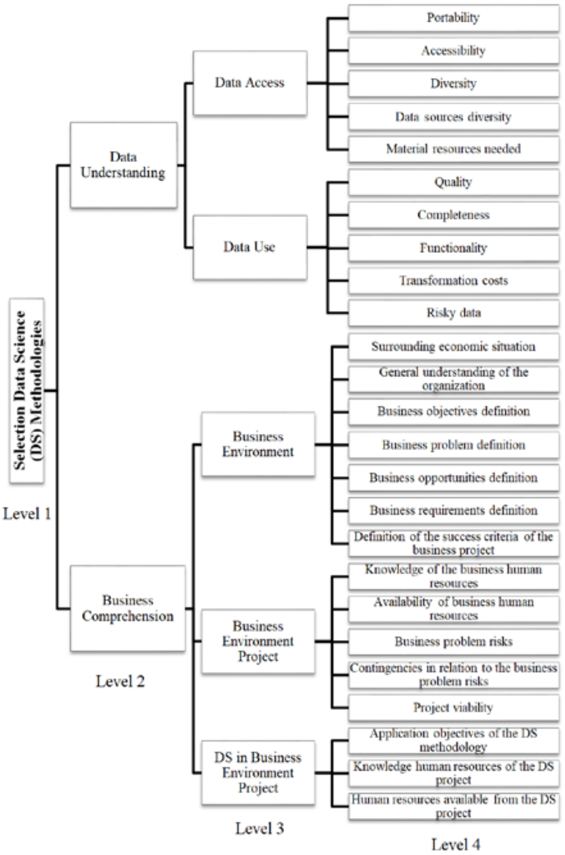
Selection Data Science (DS) Methodologies
Level 1
Data Understanding
Business Comprehension
Level 2
Data Access
Data Use
Business Environment
Business Environment Project
DS in Business Environment Project
Level 3
Portability
Accessibility
Diversity
Data sources diversity
Material resources needed
Quality
Completeness
Functionality
Transformation costs
Risky data
Surrounding economic situation
General understanding of the organization
Business objectives definition
Business problem definition
Business opportunities definition
Business requirements definition
Definition of the success criteria of the business project
Knowledge of the business human resources
Availability of business human resources
Business problem risks
Contingencies in relation to the business problem risks
Project viability
Application objectives of the DS methodology
Knowledge human resources of the DS project
Human resources available from the DS project
Level 4

Fig. 2 Hierarchical structure to selection DS methodologies.

In BC, there are three sub-criteria (level 3), Business Environment, Business Environment Project y Data Science in the Business Project; which simultaneously sub-divide in seven, five and three sub-criteria respectively in the fourth level.

The sub criteria for Business Environment are the current economic situation, the general comprehension of the organization, Business problem definition, Business opportunity definition, Business Requirement definition and Success criteria definition for the business project.

Human Resources (HR) business knowledge, HR availability, Business Problem Risks (concerning the project), contingencies related to the business problem and project viability are the sub-criteria for the sub-criteria Business Environment Project.

Conversely, the three sub-criteria for Data Science in the Business Project are Application Objectives for the DS methodology, HR knowledge on the DS project and HR available for the DS project.

The sub-criteria on the last level (fourth) are compared according to each methodology, meaning, P^{3}TQ, CRISP-DM and MoProPEI, being the three alternatives, which are to be evaluated.

Step 3. Valuation by levels

As it can be seen in Fig. 1, two types of appraisals are performed simultaneously, defined as steps, on one side, an amount of two-input matrixes are created, which are used for criteria and sub-criteria (Step 3.1), on the other, PCT is utilized to indicate the completeness of the sub-criteria in level 4 according to each methodology (Step 3.2).

Step 3.1. Valuation with two-input matrixes

Two-Input matrixes are utilized for the criteria in level 2 and the sub-criteria in levels 3 and 4, with the purpose of facilitate the experts taking a decision; where X is the corresponding preference value of one criteria respect another, based on the fundamental scale proposed by Saaty.

Table 1 exposes an example of said matrixes, where the main criteria are evaluated, in other words, Data Comprehension (DC) and Business Comprehension (BC), with the same assigned degree of importance.

Table 1 Criteria paired-comparison.

Data Understanding								**Business Comprehension**
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				

Step 3.2. Valuation with TCP

TCP is utilized to establish linguistic tags, defined in a scale from 1 to 9, where value 1 indicates that the sub-criteria is not analyzed; values from 2 to 9, represent the values in the interval expressed in percentages, referring to the fulfillment of that aspect concerning the case study. Depending on if the sub-criteria is either a positive or a negative aspect, the values in the scale between 2 and 9 may be inverted.

For example, Table 2 exposes the linguistic labels for two sub-criteria, the first one being Completeness, which is a positive aspect, and therefore the values (percentages) on the interval are organized in ascending order; the opposite is true for Data Risks (negative aspect), where the percentages in the scale can be found in a descending order.

Table 2 Linguists labels for sub-criteria.

Completeness		**Data risky**	
1	Not analyzed	1	Not analyzed
2	1% to 13%	2	98% to 100%
3	14% to 27%	3	84% to 97%
4	28% to 41%	4	70% to 83%
5	42% to 55%	5	56% to 69%
6	56% to 69%	6	42% to 55%
7	70% to 83%	7	28% to 41%
8	84% to 97%	8	14% to 27%
9	98% to 100%	9	1% to 13%

Following the establishment of the aforementioned definitions, a grid-like matrix is created, for which bipolar values or defined extremes for the aforementioned scale are defined; the best and the worst case, which represent 1 and 9 respectively.

The expert then completes the matrix with each corresponding value for the case study, considering the aforementioned scale. As an example, a variety of sub-criteria with their bipolar values on each extreme and their assigned appraisal for each methodology are exposed in Table 3.

Table 3 Compliance for last level sub-criteria.

	P^{3}TQ	**CRISP-DM**	**MoPro PEI**	
Completeness is not analyzed	7	9	9	98% to 100% complete-ness
Data risky is not analyzed	5	8	9	0% to 13% data risky
...	…	…	…	…

Step 4. Matrix completeness and priority establishment between criteria

Based on the propositions brought forward by Saaty, the corresponding matrixes created in the previous step are completed, defining their importance by groups of criteria and sub-criteria according to the defined hierarchy.

For the case of the two-input matrixes, for the criteria in level 2 and sub-criteria in level 3, there is a direct transference of the evaluations performed by the experts to the corresponding values of the new paired-comparison matrixes.

Based on the obtained grids, for the sub-criteria in level 4, the paired matrixes are completed; for this, the difference in absolute values between the paired-appraisals plus one are taken, (for example, if two sub-criteria have the following appraisal 7: 7-7=0+1, both sub-criteria have the same preference or importance; another example, if the appraisal of one sub-criteria was of 9 but for another one was 5: 9-5=4+1, the first sub-criteria has a preference of 5 over the second one, in other words, it is considered more important); with the purpose of adjusting the defined values with the linguistic labels to the scale proposed by Saaty, locating them in their corresponding place inside the matrix.

In both cases depending on the appraisals established by the experts, linguistic labels are located in either the main or the secondary part of each paired matrix, which is completed with complementary values; to then normalize each matrix and define each of them.

Table 4 exposes the continuity of the comparisons previously shown in Table 1, where the assigned values in rows and columns 2 and 3 were located, from which the sum of said columns was obtained to then obtain the normalized matrix with the respective priorities for each level, which in this example is of 50% for each criteria.

Table 4 Pairwise comparison matrix.

	DU	BC	Normalized Matrix		Priorities
DU	1	1	0,50	0,50	0,50
BC	1	1	0,50	0,50	0,50
	2	2			

Step 5. Consistency evaluation

In order to evaluate consistency in the valuation emitted by the experts, the consistency of each of the obtained matrixes is evaluated using Consistency Ratio (CR), established by Saaty, which must be equal or less than 10% (0,10). For every matrix that results inconsistent, steps 3, 4 and the current one (5) must be repeated.

Since the example in Table 4 possesses only two criteria, it cannot present any inconsistency problems; consistency conflicts can only exist from 3 criteria onwards, which why they must be evaluated.

Step 6. Establishing partial, global and total priorities

Following the steps in AHP, partial and global priorities are established, through calculating the relative weight for the criteria on each level; from which the total priorities associated to each alternative are obtained utilizing the Weighed Sum Method.

As an example, Table 5 presents the priorities for the sub-criteria Data Usage, taking into account the 5 sub-criteria for this sub-criteria and Table 6 contains the total priorities associated to each alternative, in other words, the final weights for each of the evaluated DS methodologies.

Said tables can be found without any assigned values due to the fact that these are subject to every particular validation case.

Table 5 Sub-criteria prioritization.

	P^3TQ	CRISP -DM	MoPro PEI	Weight
Quality	0,00	0,00	0,00	0,00
...	…	…	…	…
Partial Priorities	**0,00**	**0,00**	**0,00**	

Table 6 Total priorities of alternatives.

	P^3TQ	CRISP-DM	MoProPEI
DU	0,00	0,00	0,00
BC	0,00	0,00	0,00
Total Priorities	**0,00**	**0,00**	**0,00**

Step 7. DS methodology ranking

Lastly, the weights obtained in the previous step are sorted in a descending manner, generating a ranking for each evaluated methodology.

4. Validation model

In order to confirm the proposed model, two real validation cases were utilized, the purpose of the first (VC1) is to determine student desertion in the Bachelor of Systems on National University of Río Negro for the period comprised from 2009 to 2015, and the second one (VC2), consists on identifying causes in breakdowns for 0KM automobiles as they are being transported from the factory to the concessionaire.

Thereupon the main results obtained by the model are presented for both validation cases. It is noteworthy to mention that the values were given by the experts as a result of their analysis of each methodology for every case; said experts being PhDs (highest academic degree), they possess a great amount of experience in numerous DS projects and offer DS consultancies.

Independently of the validation case for the study, the common steps for this type of study are, goal, criteria and sub-criteria definition using TCP (step 1), hierarchical structure construction (step 2), a part of validation by levels (step 3), using two-input matrixes (step 3.1.) and defining linguistic tags for the sub-criteria (step 3.2), without their appraisal which is characteristic of every study.

The final results obtained utilizing the proposed model for each validation case, are exposed below.

Student Desertion (VC1)

Taking only the left branch of the hierarchy proposed in Fig. 2, for VC1, the resulting percentages can be visualized in Fig. 3. As DU represents 50% relevance in DS projects, the global weights are of 26% for MoProPEI, 17% for CRISP-DM and 7% for P^3TQ.

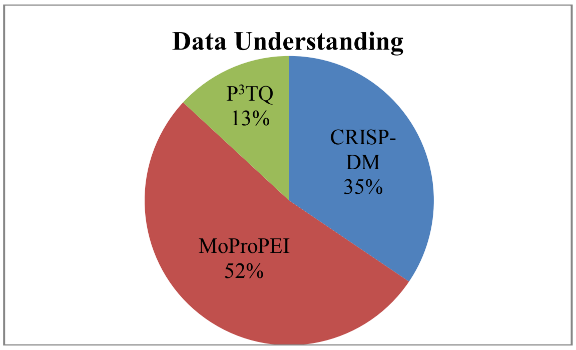

Fig. 3 Global ponderations for ED for VC1.

As BC represents 50% of transcendence when choosing a DS methodology. The calculated global appraisal indicates that MoProPEI represents 23%, CRISP-DM 17% and P^3TQ 11%; which expressed in 100%. Taking only BC into account, the obtained percentages are exposed in Fig. 4.

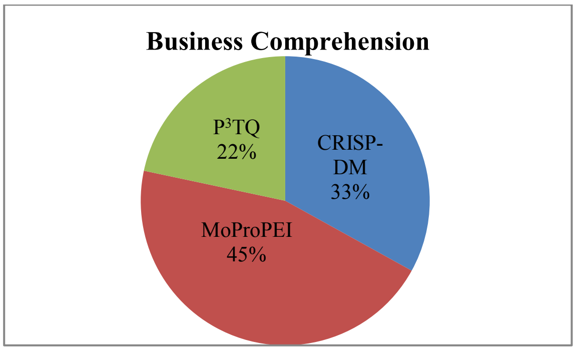

Fig. 4 Global Weightings of CN for CV1.

Finally, the total priorities for the model for each methodology were set, based on calculated priorities (local and global). Table 7 exposes the total priority for each methodology in the DS project utilizing CV1 as validation case; said table evidences the preference of the second alternative over the rest, given that MoProPEI was 49% adequate, followed by CRISP-DM with 34% and lastly P^3TQ with 17%.

Table 7 Total Priorities for VC1.

	DU	BC	Total Priorities
CRISP-DM	0,17	0,17	0,34
MoProPEI	0,26	0,23	0,49
P^3TQ	0,07	0,11	0,17

0KM Car Breakdowns (VC2)

Concerning VC2, with 50% importance when selecting or developing a DS project, DU obtained a global weighting of 25% for MoProPEI, 14% for P^3TQ and 12% for CRISP-DM. In an isolated manner, the left branch on the hierarchy reached the percentages presented in Fig. 5 (expressed in 100%).

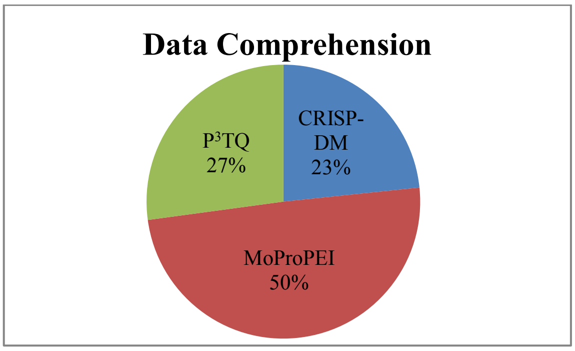

Fig. 5 Global weightings of ED for VC2.

Once local and global priorities are obtained in inferior levels, the global weighting for BC was calculated, for which again MoProPEI was the most adequate with 28%, followed by CRISP-DM with 13% and lastly P^3TQ with 9%. By analyzing only, the right branch of the hierarchy, BC obtained the percentages visible in Fig. 6.

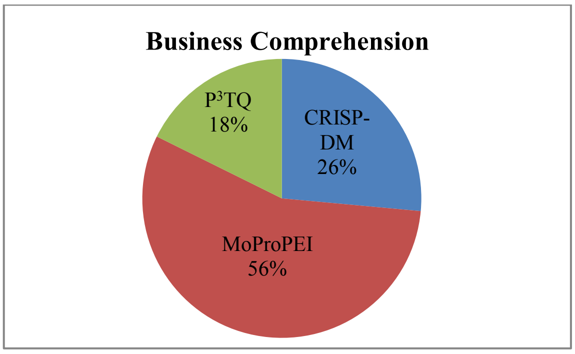

Fig. 6 Global Weighings for CN for VC2.

Total priorities obtained for VC2 are exposed in Table 8, where a clearly significant adequation for the case is presented by MoProPEI, given that a primacy del 53% was obtained, followed by CRISP-DM with 25% and lastly P^3TQ with 22%. The criteria on the second level (DU and BC) represent 50% each, which values are shown in said table.

Table 8 Total Priorities for VC2.

	DU	BC	Total Priorities

CRISP-DM	0,12	0,13	0,25
MoProPEI	0,25	0,28	0,53
P^{3}TQ	0,14	0,09	0,22

Finally, the ranking for each methodology for each validation case with their respective percentages is shown in Fig. 7; it is noted that the distribution for the second case (C2) presented a larger bias towards MoProPEI than in the first one (C1).

For C1 the average difference between every position is approximately 16%; in C2 however, the difference between the first and last position is greater than for the other methodologies.

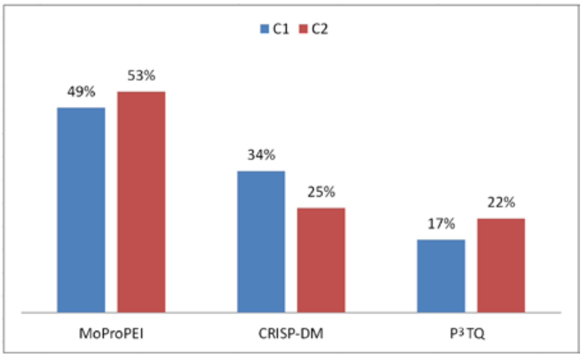

Fig. 7 Methodology ranking.

5. Conclusions

The proposed model provides a solid base for comparison and selection of methodologies utilizing by integrating AHP and utilizing PTC.

PCT was combined with AHP with the purpose of facilitating criteria and sub-criteria definition, as well as diminishing the subjectivity which said activity may present; it also contributes to the construction of a precise hierarchic structure in the different classificatory levels grouping the sub-criteria, achieved in collaboration with the experts.

It is also noteworthy to highlight the creation of two-input matrixes which facilitate criteria comparison; the same occurs for grid-like matrixes defined from TCP; these last being completed based on their degree of completeness or compliance for each sub-criteria in relation to each methodology for every particular validation case, taking into account their own defined linguistic labels which are then adapted to the scale proposed by Saaty.

Both validation cases allowed corroborating the effectiveness of the proposed model; said cases were real cases with dissimilar characteristics, which enriched the obtained results and validated their utilization in different application contexts.

For both cases, MoProPEI was the methodology with the highest degree of compliance and preference.

As future research, it is proposed to develop software that materializes the proposed model. On another hand comparing MoProPEI with agile DS methodologies or even utilization in other areas. It is currently being tested comparing MoProPEI with TDSP and ASUM.

It is also considered to integrate another MCDM method to the current model, with the purpose of correcting for some of the weaknesses that AHP can present.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

KE and VB conceived the general idea. KE defined and implemented the model. KE and VB con-ducted the experiments, analyzed the results and wrote the manuscript. All authors read and approved the final manuscript.

References

[1] M. Karanik, S. Gramajo, L. Wanderer, M. Giménez, and D. Carpintero, "Multi-Criteria Decision Model based on AHP and Linguistic Information," *Journal of Computer Science & Technology*, vol. 14, no. 1, pp. 16–24, Apr. 2014.

[2] J. C. Osorio Gómez and J. P. Orejuela Cabrera, "El proceso de análisis jerárquico (AHP) y la toma de decisiones multicriterio. Ejemplo de aplicación.," *Scientia et technica*, vol. 2, no. 39, Aug. 2008.

[3] A. Dadda and I. Ouhbi, "A decision support system for renewable energy plant projects," presented at the 2014 International Conference on Next Generation Networks and Services (NGNS), Casablanca, Morocco, 2014, pp. 356–362.

[4] E. Triantaphyllou and S. H. Mann, "Using the analytic hierarchy process for decision making in engineering applications: Some challenges," *International Journal of Industrial Engineering: Applications and Practice*, vol. 2, no. 1, pp. 35–44, Jan. 1995.

[5] T. L. Saaty, *The analytic hierarchy process*. New York: McGraw-Hill, 1980.

[6] M. del S. García Cascales, "Métodos para la comparación de alternativas mediante un Sistema de Ayuda a la Decisión (S.A.D.) y 'Soft Computing,'" Tesis de Doctorado, Universidad Politécnica de Cartagena - Departamento de Electrónica, Tecnología de Computadoras y Proyectos, Cartagena, Colombia, 2009.

[7] R. de F. S. M. Russo and R. Camanho, "Criteria in AHP: A Systematic Review of Literature," *Procedia Computer Science*, vol. 55, pp. 1123–1132, Jan. 2015.

[8] G. Kou and C. Lin, "A cosine maximization method for the priority vector derivation in AHP," *European Journal of Operational Research*, vol. 235, no. 1, pp. 225–232, May 2014.

[9] E. H. Forman and S. I. Gass, “The Analytic Hierarchy Process—An Exposition,” *Operations Research*, vol. 49, no. 4, pp. 469–486, Aug. 2001.

[10] O. S. Vaidya and S. Kumar, “Analytic hierarchy process: An overview of applications,” *European Journal of Operational Research*, vol. 169, no. 1, pp. 1–29, Feb. 2006.

[11] J. Mayor, S. Botero, and J. D. González-Ruiz, “Modelo de decisión multicriterio difuso para la selección de contratistas en proyectos de infraestructura: caso Colombia,” *Obras y proyectos*, no. 20, pp. 56–74, Dec. 2016.

[12] G. Islei and A. G. Lockett, “Judgemental modelling based on geometric least square,” *European Journal of Operational Research*, vol. 36, no. 1, pp. 27–35, Jul. 1988.

[13] P. Britos, B. Rossi, and R. García Martínez, “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*, 1999, pp. 200–209.

[14] K. Eckert and P. V. Britos, “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.

[15] K. B. Eckert and P. V. Britos, “Data science methodologies selection with hierarchical analytical process and personal construction theory,” presented at the XXV Congreso Argentino de Ciencias de la Computación (CACIC), Río Cuarto, Córdoba, Argentina, 2019.

[16] M. A. Waller and 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.

[17] P. Pytel, P. Britos, and R. García Martínez, “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, pp. 33–88.

[18] J. Á. Vanrell, R. A. Bertone, and R. García Martínez, “Modelo de proceso de operación para proyectos de explotación de información,” presented at the XVI Congreso Argentino de Ciencias de la Computación, 2010.

[19] D. Pyle, *Business Modeling and Data Mining*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[20] P. Chapman *et al.*, “CRISP-DM 1.0: Step-by-Step Data Mining Guide.” Edited by SPSS, 2000.

[21] S. Martins, P. Pesado, and R. García Martínez, “Propuesta de Modelo de Procesos para una Ingeniería de Explotación de Información: MoProPEI,” *Revista Latinoamericana de Ingenieria de Software*, vol. 2, no. 5, pp. 313–332, 2014.

[22] T. L. Saaty, “How to make a decision: The analytic hierarchy process,” *European Journal of Operational Research*, vol. 48, no. 1, pp. 9–26, Sep. 1990.

[23] T. L. Saaty, “Analytic Hierarchy Process,” in *Encyclopedia of Operations Research and Management Science*, S. I. Gass and M. C. Fu, Eds. Boston, MA: Springer US, 2013, pp. 52–64.

[24] P. T. Harker, “The Art and Science of Decision Making: The Analytic Hierarchy Process,” in *The Analytic Hierarchy Process: Applications and Studies*, B. L. Golden, E. A. Wasil, and P. T. Harker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1989, pp. 3–36.

[25] L. Vera Montenegro, “Aplicación y Comparación de Metodologías Multicriterio (AHP y Fuzzy Logic) en la Selección de Tecnologías Postcosecha para Pequeños Productores de Cacao,” Tesis de Doctorado, Universidad Politécnica de Valencia, Valencia, España, 2014.

[26] T. L. Saaty, “Decision making with the analytic hierarchy process,” *International Journal of Services Sciences*, vol. 1, no. 1, pp. 83–98, Jan. 2008.

[27] T. L. Saaty, *Fundamentals of Decision Making and Priority Theory With the Analytic Hierarchy Process*. RWS Publications, 2000.

[28] R. García Martínez and P. V. Britos, *Ingenieria de Sistemas Expertos*. Nueva Librería, 2004.

[29] T. Butt, *George Kelly: The Psychology of Personal Constructs*. Macmillan International Higher Education, 2008.

[30] F. Provost and T. Fawcett, “Data Science and its Relationship to Big Data and Data-Driven Decision Making,” *Big Data*, vol. 1, no. 1, pp. 51–59, Feb. 2013.

[31] T. Schoenherr and C. Speier-Pero, “Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential,” *Journal of Business Logistics*, vol. 36, no. 1, pp. 120–132, 2015.

[32] J. C. Giraldo Mejia and J. A. Jiménez Builes, “Caracterización del proceso de obtención de conocimiento y algunas metodologías para crear proyectos de minería de datos,” *Revista Latinoamericana de Ingeniería de Software*, 2013.

[33] J. M. Moine, S. E. Gordillo, and A. S. Haedo, “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.

[34] J. M. Moine, “Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo,” Tesis de Maestría, Facultad de Informática, 2013.

[35] H. J. G. Palacios, G. A. H. Pantoja, A. A. M. Navarro, I. M. A. Puetaman, and R. A. J. Toledo, “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)*, 2016, pp. 1–9.

[36] M. T. Rodríguez Montequín, J. V. Álvarez Cabal, J. M. Mesa Fernández, and A. González Valdés, “Metodologías para la realización de proyectos de Data Mining,” presented at the VII Congreso Internacional de Ingeniería de Proyectos, Pamplona España, 2003, pp. 257–265.

Citation: K. B. Eckert and P. V. Britos. *Proposed Extended Analytic Hierarchy Process for Selecting Data Science Methodologies.* Journal of Computer Science & Technology, vol. 21, no. 1, pp. 49-58, 2021.

DOI: 10.24215/16666038.21.e06

Received: March 14, 2020 **Accepted:** December 1, 2020.

Copyright: This article is distributed under the terms of the Creative Commons License