TAF Journal of Advances in Technology and Engineering Research JATER
2018, 4(6): 241-253

PRIMARY RESEARCH

Enhanced CRISP-DM model for SME coach operator

Siti Aishah Mohd Selamat [1*], Simant Prakoonwit [2], Reza Sahandi [3], Wajid Khan [4]

[1, 2] Creative Technology Department, Faculty of Science and Technology, Bournemouth University, Dorset, England

[3, 4] Department of Computing, Faculty of Science and Technology, Bournemouth University, Dorset, England

Keywords

Data mining methodology
Private coach hires
Transportation
SMEs
Data mining application
CRISP-DM

Received: 9 October 2018
Accepted: 6 November 2018
Published: 19 December 2018

Abstract

The surging call for SMEs in the private transportation domain to apply data analytics in its business is evidential. However, to date, there is a lack of specific data mining methodology or framework customized to meet the demand of the private transportation domain. Considered as a de-facto data mining methodology by the industry to date, the Cross-Industry Standard Process for Data Mining (CRISP-DM) consists of a six-phase process that is extendable in the framework – from generic to specialized tasks. The traditional CRISP-DM methodology exemplifies the DM application area, issues identifications, technical, tools and technique requirements. Through this study, an Enhanced CRISP-DM for SME Coach Operator (ECSMCO) methodology was developed. The extended methodology aims to curb the existing application limitation identified in the small and medium-sized enterprises (SMEs) by proposing an enhancement to the existing CRISP-DM activities. To evaluate the novel ECSMCO's acceptability, three UK SME coach operators were identified to apply and evaluate the ECSMCO methodology. The outcome of this paper will ascertain the users' acceptability outcome of the applied ECMSCO methodology.

I. INTRODUCTION

The emergence of technological advancement has accounted for the huge surge of data collected. As a result, it has become a point of criticality for organisations especially so for the SMEs to start adopting data analytic application in its business [1, 2, 3]. In the European (EU) continent alone, it was uncovered that the SMEs are the key economic drivers of growth, contributing close to 4 trillion euros to the EU economy in 2015 [4, 5]. In 2012, the transportation domain constitutes 5% of the 22.3 million non-financial economies [6]. In a further report by the IDC European Vertical Market, it was identified that an approximate of 49,000 transportation SMEs has yet to adopt data analytics in the business [7]. Despite the evidential validation that SMEs can reap up to 6% productive through data analytics adoption; the SMEs are still however reluctant it in approach [8, 7]. Through an extensive conducted it was uncovered that the hindrance factors of data analytics adoption are in the area of data management, knowledge management and data management [9]. Data Mining (DM) is crucial for SMEs, especially so, in the private transport domain as the organisation can inherently extract and process its dataset to uncover new insights to facilitate better decision for the business [10, 11]. In an extensive study of DM application by the large transportation enterprise and SMEs context at large, it was uncovered that the CRISP-DM methodology is the most popular applied DM model by the industry [9]. The other two leading DM methodology includes the Knowledge Discovery in Database (KDD) and 'Sample, Explore, Modify, Model and Assess' (SEMMA) models.

In this study, using the CRISP-DM model as the foundation methodology, a novel ECMSCO methodology is being proposed for SMEs in the private transportation domain. The aim of the novel methodology is to address the identified DM application challenges and limitations uncovered within the SMEs' context at large. Specifically formulated for the SMEs in the private coach hire within the transportation domain, the proposed novel methodology may bring potential benefits for the various multi-disciplinary SMEs' usages. The paper will cover the background overview of

*Corresponding author: Siti Aishah Mohd Selamat
†email: aishah@bournemouth.ac.uk

CRISP-DM methodology and followed by the issues to be considered when applying DM in the private transportation domain. The subsequent sections will outline the proposed ECMSCO for each phase and conclude with a proposed assessment method to evaluate the conformance of the methodology.

II. CRISP-DM

DM is defined as the art of extracting and processing new and unique insights from a large volume of the dataset using high computing data analysis [12, 13]. The overall outcome of DM is to churn out new information's or generate predictions from the raw dataset [14]. The CRISP-DM model was developed jointly in the mid-1990s by big organisations like Daimler-Chrysler, OHRA, SPSS and Teradata [15]. Originating from the initial KDD methodology, the CRISP-DM were developed with the aim to be industry independent and technologically neutral [15]. Therefore, CRISP-DM can be applied by the organisation from various sector and industries [16]. An online poll conducted by KDNuggets in 2014, reveals that the CRISP-DM model was voted as the most used and popular DM methodology for DM industry practitioners [17]. Unlike the first generation KDD methodology, the CRISP-DM methodology operates in an iterative process as depicted in Figure 1.

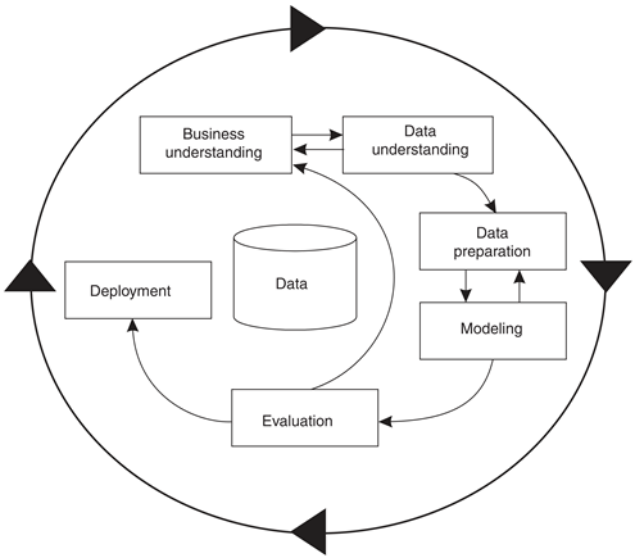

Fig. 1. CRISP-DM process methodology (Source: [15])

In overall, the CRISP-DM methodology consists of six sequential phases that are cyclical and non-restrictive in nature. Allowing the DM practitioner the flexibility to move backwards and forward in the DM process without any restraint. The detailed illustration of the CRISP-DM's six phases are as follows:

A. Business Understanding

The first phase consists of deriving the understanding of the DM project requirements and objectives–from the business perspective. This set of information will then outline into a set of DM problems that require addressing. The first phase will conclude putting together a preliminary plan in order to address the DM issues.

B. Data Understanding

The second phase of data understanding begins by carrying out the initial data collection in order to be accustomed and familiar with the data collected at hand. This is followed by identifying the data quality, discovering the first initial insights of the data and diagnoses any interesting subsets to derive a set of hypotheses.

C. Data Preparation

The third phase of data preparation involves constructing the raw dataset into the final dataset, which will be used in the modelling phase. The data preparation phase may require to be executed in several instances depending on the modelling phase outcome. The task in this phase includes attribute, record, and table selection. Last but not least, data transformation and cleaning of data for the modelling phase next.

D. Modeling

In the fourth phase, a selection of modelling techniques is chosen and is applied to the prepared dataset. The model's

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

parameter is calibrated accordingly in order to achieve the optimum performance of the model. Generally, a similar set of data mining issue type can have several modelling techniques. The data requirements of each technique may differ from one another. Therefore, this phase may experience the need to return back to the third phase of data preparation in order to achieve the data requirement of the techniques to be applied.

E. Evaluation

By the fifth phase, a perfected model (or models) would have been generated from the fourth phase. Before the model(s) could be finally deployed, it will be evaluated thoroughly in the evaluation phase in order to ensure that the model designed meets the business requirements–established in the first phase. The key objective of this phase is to review that all the business issue are being considered sufficiently. This phase would conclude whether the model is ready to be deployed and used by the business.

F. Deployment

The model created does not necessarily mark the last process of the DM project. When the model(s) create new and unique insights of the data, the knowledge created needs presented and reported to the organisation's management for strategic decision-making such as new product development or enhancement or creating a targetted marketing campaign. The last phase of deployment is crucial for the business to carry it out in order to reap the tangible benefits of the DM model(s) created through the CRISP-DM methodology.

III. ISSUES TO BE CONSIDERED WHEN APPLYING DM IN THE PRIVATE TRANSPORTATION DOMAIN

To date, there are little research and case studies conducted on DM application in the transportation domain. Therefore there are no evidential challenges available based on existing research. Nonetheless, the issues to be considered when applying DM in the private transportation domain are based on extensive studies carried out on the application of DM in the transportation sector by the larger enterprise and in the SMEs context [9]. It was uncovered that the CRISP-DM methodology is widely used by both the large enterprise in the transportation domain and the SME group as a whole. The distinctive list of CRISP-DM strength includes the methodology flexibility and yet structured approach that can be applied to organisation and businesses from different industry and operating size. The compiled challenges and list of limitation are as follows:

1. The CRISP-DM methodology entails a very extensive, precise and long process in each phase.
2. The end-user would require an in-depth knowledge concerning the DM domain application.
3. Determining the appropriate selection of data and attribute for the DM project.
4. The delay selection of DM technique that impacts the data formation during the data preparation phase–involving the end-user to go back and forth multiple times in between these two process.

From the above list of challenges, it suggests that the end-user is encountering an exhaustive means of operating the DM process in view of the unexpected outcome of each phase. This, therefore, requires the end-user to go back and forth in between each phase in order to derive a perfected DM model outcome for deployment. In a separate study carried out, the other area of limitation in data analysis includes in the area concern the data privacy and protection [9]. It was also highlighted that there is no specific DM-KM assessment method to evaluate the application of knowledge created. Last but not least, it was uncovered that the key data type used for analysis by the SMEs context at large is structured data.

IV. EXTENSION OF CRISP-DM METHODOLOGY FOR TRANSPORTATION DOMAIN

Based on the research conducted, and to the best of the authors' knowledge, there is no specific model or framework to date, to conduct DM analysis in the private transportation domain. As illustrated in Section 2, CRISP-DM methodology entails a hierarchical process that is expandable in its framework. According to the CRISP-DM methodology, the third and fourth layers are abstracted in order to map generic task to specialised task. The mapping is put in place in order to cater to the future extension of the generic and specialised task as per required the various pre-defined DM project [15]. Based on the discussion in Section 3, the append issues listed below will be considered in the application of DM in the private transport domain:

1. A systematic, concise and lucid DM process for SMEs practitioners.
2. Taking into consideration the data privacy and protection legislation constraints.
3.Pre-determining the data selection prior hand–structured data only.
4. Pre-determining the DM technique selections.
5. Incorporating an overall DM-KM assessment method.

A set of generic and specialised tasks will be derived in order to enhance the CRISP-DM methodology to be used for

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

the private transportation domain and in addition, address the above-listed issues.

V. ECSMCO METHODOLOGY

In this section, the formulated ECSMCO methodology is introduced. The enhanced ECMSCO generic and specialised tasks are marked with an asterisk (*) notation. The illustration of the changes will be described in the individual six phases. The overall list of generic tasks, specialised tasks, and deliverable can be found in Table 1.

A. Phase 1–Project Initiation & Planning

In the original CRISP-DM phase 1 and phase 2 of 'Business understanding' and 'Data understanding' is where the DM project expectations and conceptualisation are being outlined. The subsequent phases delve more into the implementation activities of the DM project. The implementation phases are driven entirely by the objectives set in the first and second phase. Unlike the implementation phases, which are distinctly iterative and incremental, the intended changes to be made in the first two phases will affect the entire DM project deliverables. Should there be a significant adjustment to the DM project phases an overall project restart would be required.

In the first phase of the ECMSCO methodology, the original 'Business understanding' was rephrased to 'Project Initiation & Planning'. This is done in order to give the private transportation the direct meaning and objective of Phase 1 at face value. Additionally, the first task of 'Determine Business Objectives' was renamed to 'Determine Project Objectives' in order to have the same consistency with Phase 1 header name. Under this task, two new activities were introduced–'Define Project Objectives' and 'Define Project Scope'–addressing the issue on the need for the DM process to be concise and more businesslike in its approach. In the second task of 'Access Current Condition', three new activities were introduced–'Business Data Availability', 'Project Resource Requirement' and 'Data Protection Regulation'. The first two activities address the issue on the need to have a systematic, concise and lucid DM processes for the SMEs end-user. The last activity on the, on the other hand, would address the issue and concern on the 'data privacy and protection legislation constraints'. In the last task of 'Produce Project Plan', a new activity of 'Identify data source' has been added to address the issue on 'pre-determining the data selection prior hand'. The described new additions are as displayed in Figure 2.

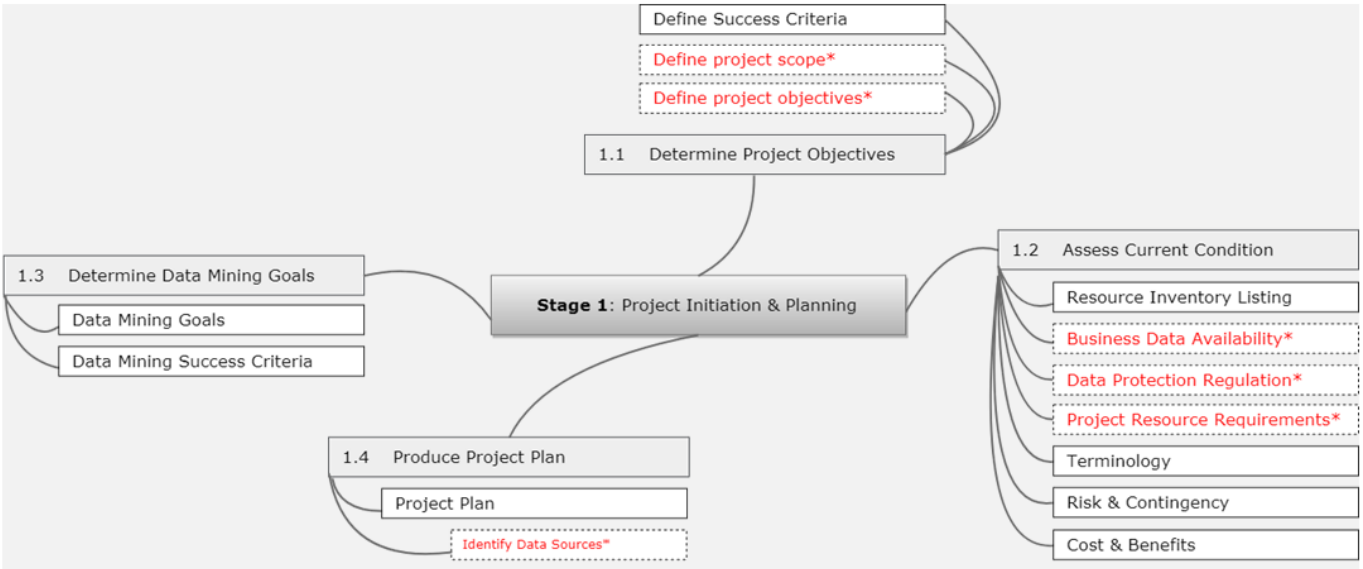

Fig. 2. ECMSCO phase 1

B. ECMSCO Phase 1

Next, the second phase of 'Data Understanding', a new generic task named 'Data Collection Preparation', with three new activities of 'Evaluate Data Source Availability', 'Define Data Collection Timeline', 'Define Data Handling Protocol'.

Additionally, another new generic task named 'Initial Data Discovery' with two new activities of 'Execute Exploratory Data Analysis' and 'Report Findings'. Last but not least, a new activity called 'Acquire Data From All Source' was also added to the pre-existing task named 'Collect Initial Data'.

In summation, the tasks and activities added will address

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

the issue on 'pre-determining the data selection prior hand'. This is key as with the full understanding of the existing attributes and characteristics, the next phase of 'Modelling' would be focus-driven and not ambiguous. The described new additions are as displayed in Figure 3.

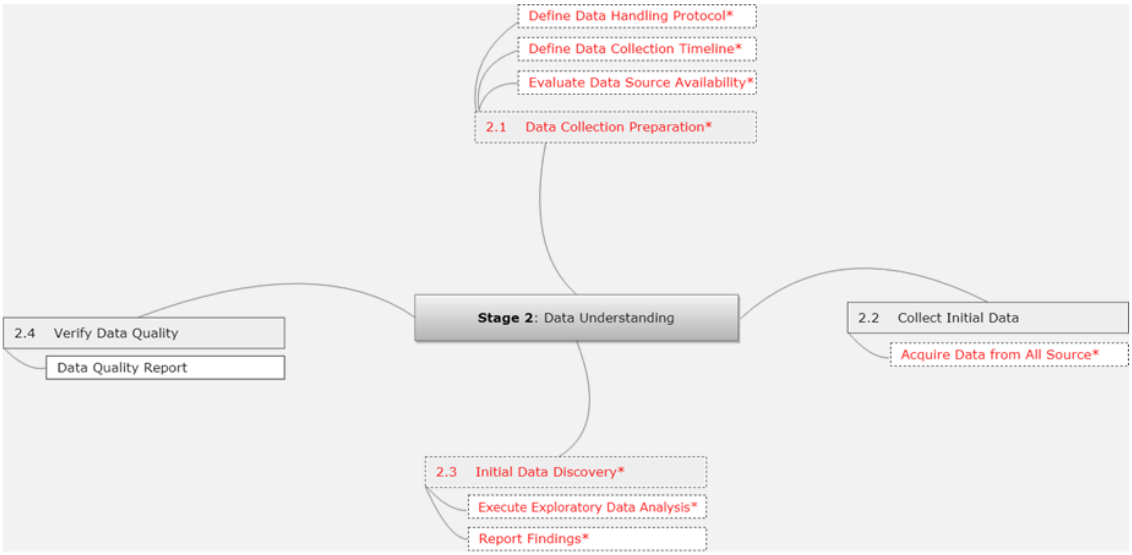

Fig. 3. ECMSCO phase 2

C. *Phase 3–Data Preparation*

In the 3rd phase of 'Data Preparation', a new generic task named 'Data Extraction' with two new activities of 'Extract Business Data' and 'Structured Data Only' was added. This is to address the issue on the need to 'pre-determine the data selection prior hand–focusing on structured data only'. Under the existing task of 'Data Cleaning', a new activity of 'Handling Outlier' was added. Apart from the customary issue of handling missing data, the detection of an outlier is key to detect data omission error caused by system or human error. Last but not least, under the existing task of 'Data Selection', two new activities under 'Data Sampling' were added. There are data samplings for 'Training' and 'Validation'. This step is crucial to prepare for the next modelling phase to test the performance of the formulated DM project model. Therefore, a separate set of training and validation dataset is required. The described new additions are as displayed in Figure 4.

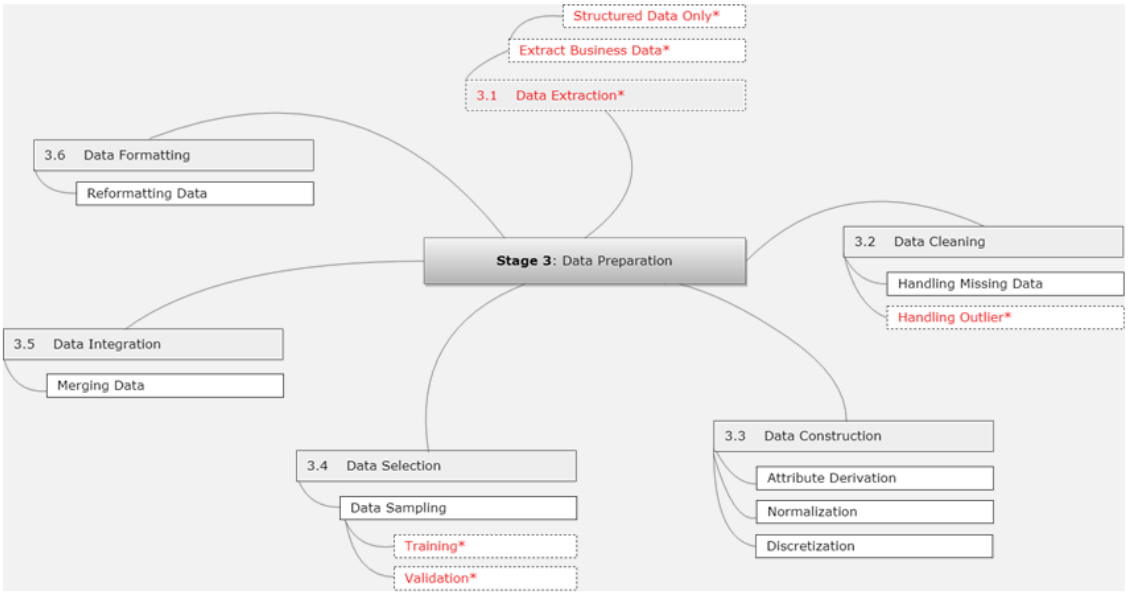

Fig. 4. ECMSCO phase 3

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

D. Phase 4–Modelling

In phase 4 of the 'Modelling' stage, a new task named 'Feature Engineering' with a new activity called 'Feature Selection' was added. The objective is to identify the most relevant variables that can give the best predictive modelling to the data. In other words, in this task, it could mean that the variable will be selected and de-selected in order to get the best performing model. This new task and activity would address the issue on the need to 'pre-determine the data selection prior hand'. The common feature selection includes stepwise regression, sequential feature selection, regularization and Neighborhood Component Analysis (NCA). A new generic task of 'Feature Transformation' was also added in, should there be a need to turn existing variable into the new set of variables. And the common feature transformation technique includes Principal Component Analysis (PCA) and factor analysis. Next, under the 'Select Modelling Technique' task, a new activity named 'Machine Learning Technique' was added to address the key issue on the need 'pre-determine the technique selection'. Next, under the existing task of 'Build Model', a new activity of 'Train Model' was added using the feature derived in the earlier task. Last but not least, under the existing task of 'Assess Models', two new activities of 'hyperparameter tuning' and 'prepare for next model iteration' were added. Hyperparameter tuning is an iterative process in order to find tunes the best model performance. The described new additions are as displayed in Figure 5.

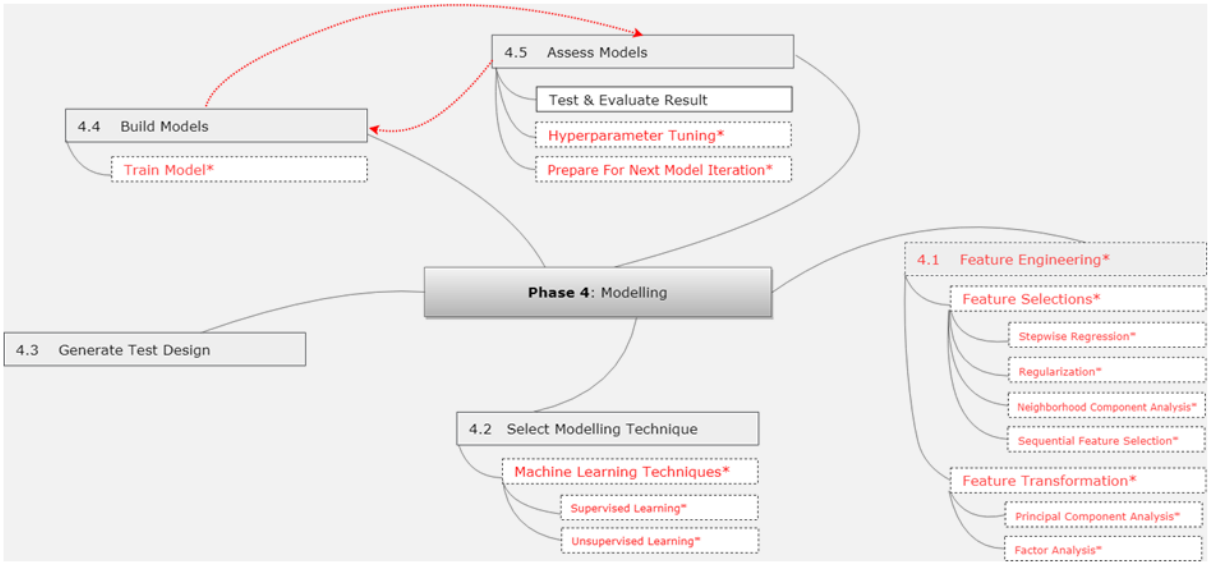

Fig. 5. ECMSCO phase 4

E. Phase 5–Evaluation

In the evaluation phase as depicted in Figure 6, remains with no amendments.

F. Phase 6–Deployment

Under the sixth phase 'Deployment', a new specialised task of 'DM-KM Assessment' was introduced in order to address the issue of the non-availability of an integrated DM-KM assessment. The added task will evaluate the applicability and usability of the knowledge created through the DM project. The described task is as shown in Figure 7.

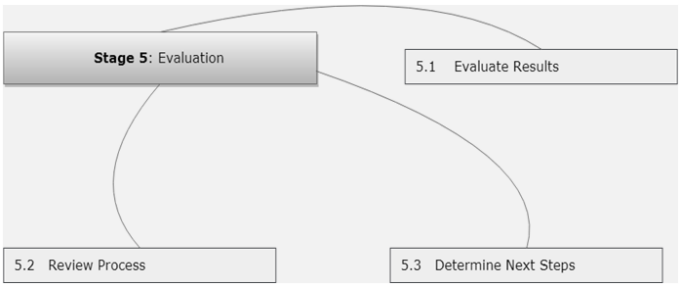

Fig. 6. ECMSCO phase 5

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

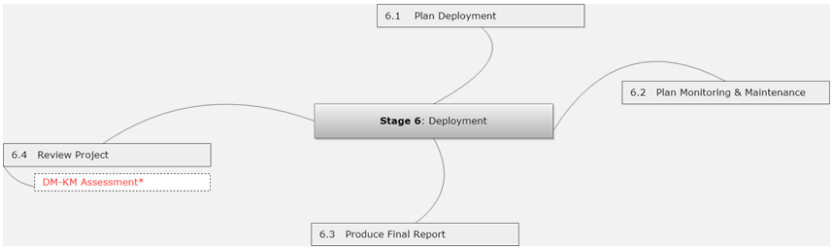

Fig. 7. ECMSCO phase 6

Last but not least, as depicted in Figure 8, an iterative process chain of the line has been drawn from the deployment phase (phase 6) to the project initiation & planning phase (phase 1). This substitute the traditional chain of the line marked from the evaluation phase (phase 5) to the first phase. With the new task of carrying out the DM-KM assessment in the last phase 6, it is practical to return to the first phase with a well-rounded evaluation to review the project initiation and planning phase before embarking on the next DM project. The overall ECMSCO process addresses the issue of the need to have a 'systematic, concise and lucid DM process for SMEs practitioners'. Further to this, the methodologies have rephrased the terms used to be more businesslike for the SMEs leaders and practitioners accustomed too.

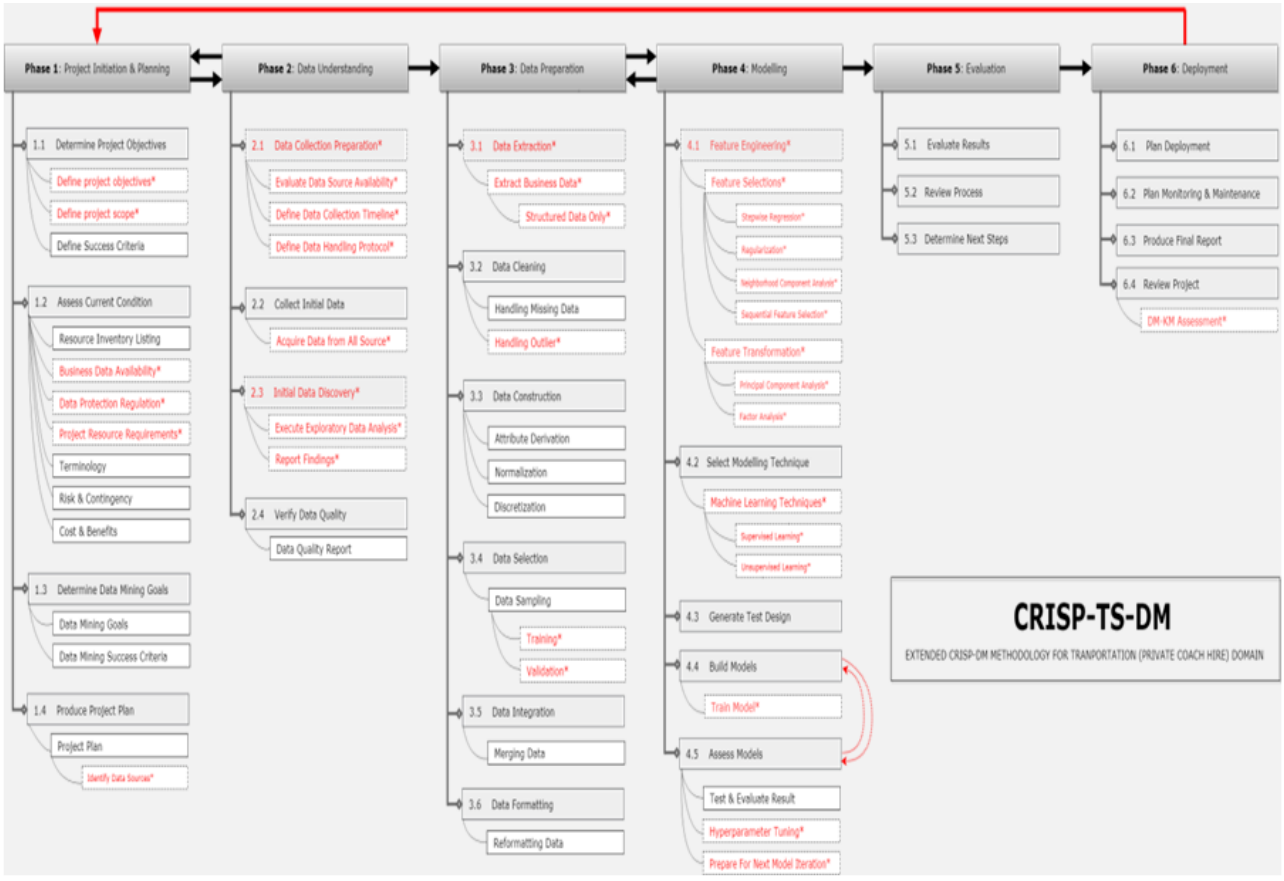

Fig. 8. Overall ECMSCO methodology

The overall list of ECMSCO tasks, deliverables and metrics in accordance to the first strategy is provided in Table 1.

VI. LIST OF TASK, ACTIVITIES AND DELIVERABLES

According to the CRISP-DM methodology, the classifications of tasks are segregated into 'generic' and 'specialised' task. Generic task refers to a set of task that encompasses the entire DM project. Meanwhile, a specialised task refers to specific tasks that are required to be delivered to a specific DM context. The last component of 'deliverables' refers to the output result of the task performed. As depicted in Table 1, the enhanced ECMSCO generic and specialised tasks are marked with an asterisk (*) notation.

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

TABLE 1
ECMSCO'S GENERIC TASK, SPECIALISED TASK AND DELIVERABLES

Phase	Generic Task	Specialised Task	Deliverables
Phase 1: Project Initiation & Planning	GT1: Determine Project Objectives	Define project objectives*	Outlining the overall project objectives
		Define project scope*	Outlining the detailed description of the project
		Define Success Criteria	Outlining the overall success criteria of the project
	GT2: Assess Current Condition	Resource Inventory Listing	Outlining the available resource inventory listing
		Business Data Availability*	Outlining the types of data available that can be used for analysis
		Data Protection Regulation*	Ensuring that the project meets with the legal guidelines in collecting and processing personal information
		Project Resource Requirements*	Drawing out the resource listing requirements
		Terminology	Outlining the glossary of terminology used for the project
		Risk & Contingency	Outlining any risk and contingency matrix
		Cost & Benefit Analysis	Carrying out a cost benefit analysis to validate the viability of
	GT3: Determine Data Mining Goals	Data Mining Goals	Outlining the desired overall data mining goals for the project
		Data Mining Success Criteria	Outlining the overall list of data mining criteria of the project
	GT4: Produce Project Plan	Project Plan	Outlining the overall project plan–deliverables and timeline. Identifying the availability of data source point that can be used for analysis
		Identify Data Sources*	
Phase 2: Data Understanding	GT5: Data Collection Preparation*	Evaluate Data Source Availability*	Evaluating the data source point availability and usability
		Define Data Collection Timeline*	Outlining the data collection timeline of the project
		Define Data Handling Protocol*	Outlining the project data handling processes
	GT6: Collect Initial Data	Acquire Data from All Source*	Acquiring data from the identified sources
	GT7: Initial Data Discovery*	Execute Exploratory Data Analysis*	Carrying out the EDA
		Report Findings*	Documenting the exploratory data analysis findings
	GT8: Verify	Data Quality	Data Quality Report Documenting the data quality report findings

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

TABLE 1
CONTINUE...

Phase 3: Data Preparation	GT9: Data Extraction*	Extract Business Data* Structured Data Only*	Extracting business data for analysis Analysis of data focus on structured data only
	GT10: Data Cleaning	Handling Missing Data	Generating data cleaning report
		Handling Outlier*	Generating data cleaning report
	GT11: Data Construction	Attribute Derivation Normalization Discretization	Determining data attribute
	GT12: Data Selection	Data Sampling Training *	
		Validation *	Identifying data sampling size Selection of dataset size for training Selection of dataset size for validation
	GT13: Data Integration	Merging Data	To carry out merging of data–if required
	GT14: Data Formatting	Reformatting Data	Formatting of data in preparation for modeling process
Phase 4: Modelling	GT15: Feature Engineering*	Feature Selections* Stepwise Regression* Regularization* Neighborhood Component Analysis* Sequential Feature Selection*	Identifying the relevant variables for data modelling
		Feature Transformation* Principal Component Analysis* Factor Analysis*	Dimensionality reduction
	GT16: Select Modelling Technique	Machine Learning Techniques* Supervised Learning* Unsupervised Learning*	Identifying the relevant machine learning type to adopt
	GT17: Generate Test Design		Test design
	GT18: Build Models	Train Model*	Model training
	GT19: Assess Models	Test & Evaluate Result	Testing and evaluation
		Hyper parameter Tuning*	Hyper parameter tuning to identify the best model performance
		Prepare For Next Model Iteration*	Model iteration if required
Phase 5: Evaluation	GT20: Evaluate Results	Results understanding and interpretation	Evaluate overall success criteria
	GT21: Review Process	Review Data Mining Process	Review of process in accordance to results evaluation
	GT22: Determine Next Steps	Determine results application areas	Outline the next steps of actions
Phase 6: Deployment	GT23: Plan Deployment	Determine results application deployment	Outline the deployment plan
	GT24: Plan Monitoring & Maintenance	Developing the monitoring and maintenance plan	Deployment maintenance and monitoring process

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

TABLE 1
CONTINUE...

GT25: Produce Final Report	Draft report for management usage	Final report
GT26: Review Project	DM-KM Assessment*	Review entire project with DM-KM assessment

VII. EVALUATION AND RESULTS

In order to ascertain the proposed novel ECMSCO, three SME coach operators based in the United Kingdom were invited to evaluate the ECMSCO users' usability and acceptability. The SME's datasets will be acquired and applied to the ECMSCO method. A customised Data Mining-Knowledge Management Usefulness, Satisfaction, and Ease of Use (DM-KM USE) questionnaire were developed for this research in order to evaluate the users' usability and acceptability of the ECMSCO method in specific. Using Lund's USE questionnaire as the foundation evaluation instrument tool when deriving the DM-KM USE questionnaire, Lund's USE questionnaire measure the user's usability and acceptability from three key aspects of perceived Usefulness, Satisfaction and Ease of Use [18]. Deriving from the mentioned three key dimensions of USE questionnaire, four areas of evaluation component consisting of 'usefulness', 'ease of use', 'ease of learning' and 'satisfaction' are derived for the DM-KM USE questionnaire. The component on 'usefulness' will measure the quality of how useful the ECMSCO method has been for the users. The component on 'ease of use' will measure the quality of how easy the ECMSCO method has been for the users. The component on 'ease of learning' will measure the quality of how easy the ECMSCO method can be learned by the users. Lastly, the component of 'satisfaction' will measure the users' overall experience and acceptability of the ECMSCO method. The DM-KM USE Questionnaire was constructed to assess the four areas of evaluation component in a seven-point Likert rating scales. Participants are asked to rate agreement with the statements, ranging from strongly disagree (one point) to strongly agree (seven points). The total number of participants that participated in the research evaluation study amounts to 24 participants. Each company individually had a total number of eight participants from different division and position levels.

The overall evaluation result analysis of the three companies begins with the Likert data distribution by% of the list of questions in the area of 'Usefulness'. The seven-point Likert rating scale range from strongly disagree (with 1 point) and strongly agree (with 7 points). The coded lists of questions under Usefulness are as follows:

UF1: The model outcome provides effective results.
UF2: The model outcome is useful.
UF3: The result shown is useful in facilitating/determining business decisions.
UF4: The result shown is useful in facilitating/determining future business opportunities.
UF5: The result shown gives the company a better understanding of the company's data.
UF6: The result shown is applicable.
UF7: The extended model developed does everything I would expect it to do.
UF8: The extended model developed meets the end users' needs.

Figure 9 below provides an illustration of Likert data distribution by % of UF1 to UF8 in a smooth lined scatter plot. It can be observed that in totality, UF2 to UF8 portray a majority of Likert of 7 points with a % distribution ranging from 50.00% to 70.83%. UF5 and UF6 mark the highest % distribution value of 70.83%. UF1 distinguish from the rest with an equal % distribution value of 50.00% for both Likert rating of 6 points and 7 points.

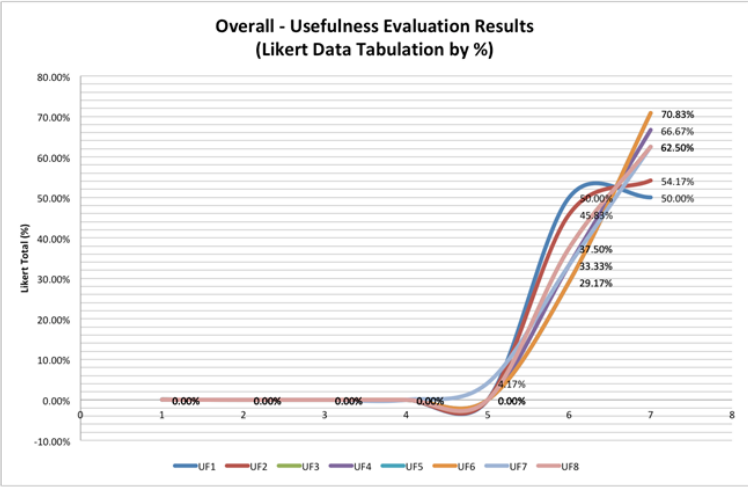

Fig. 9. Overall - usefulness likert data distribution by %

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

Next, the overall evaluation result for DM-KM Questionnaire list of question in the area of 'Ease of Use' will be covered. The coded lists of questions under 'Ease of Use' are as follows:

EU1: It is easy to use.
EU2: It is simple to use.
EU3: It is user friendly.
EU4: It requires the fewest steps possible to accomplish what I want to do with it.
EU5: It is flexible.
EU6: Using it is effortless.
EU7: I can use it without written instructions.
EU8: I don't notice any inconsistencies as I use it.
EU9: Both occasional and regular users would like it.
EU10: I can interpret and make sense of the results shown easily.
EU11: I can use it successfully every time.

Figure 10 present a smooth lined scatter plot of Likert data distribution by % for EU1 to EU11. The plot reflects a sporadic spread of % distribution of Likert ratings for EU1 to EU11. Firstly, EU6 and EU10 share a similar % distribution of 54.17% for Likert rating scale of 7 points. EU4, EU5, EU8, EU9 and EU11 has highest Likert rating scale of 6 points with a % distribution ranging from 50.00% to 75.00%. The remaining EU1 to EU3 has highest Likert rating scale of 5 points with a % distribution ranging from 58.33% to 62.50%. It is observed that only EU7 has the highest % distribution of 4 points valuing at 37.505%.

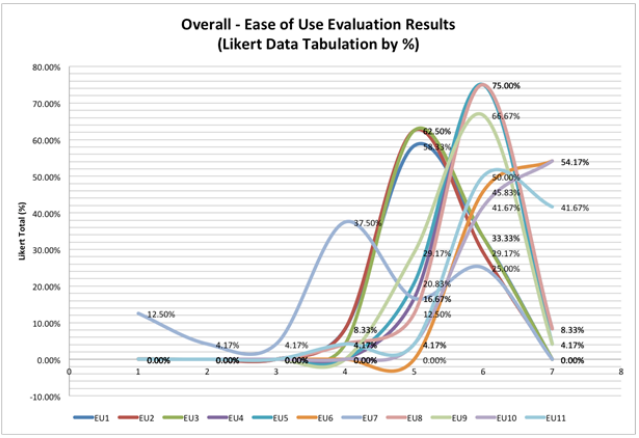

Fig. 10. Ease of use likert data distribution by %

Next, the overall evaluation result for DM-KM Questionnaire list of question in the area of 'Ease of Learning' will be covered. The coded lists of questions under ' Ease of Learning'are as follows:

EL1: I learned to use it quickly.
EL2 I easily remember how to use it.
EL3 It is easy to learn to use it.
EL4 I quickly became skilful with it.

Figure 11 reflects an illustration of Likert data distribution by % of EL1 to EL4 in a smooth lined scatter plot. It can be observed that there is a synonymous trend of Likert rating of 7 points being the highest % distribution for EL1 to EL4–where EL4 records the highest % distribution with a value of 79.17%, EL2 with 75.00% and 62.50% for both EL1 and EL3.

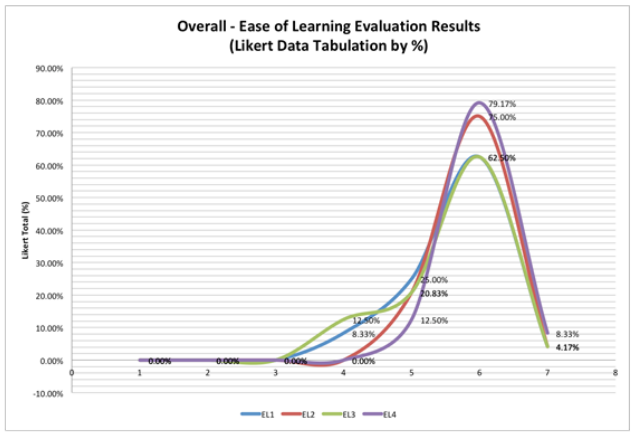

Fig. 11. Overall - ease of learning likert data distribution by %

Lastly, the overall evaluation result for DM-KM Questionnaire list of question in the area of 'Satisfaction' will be covered. The coded lists of questions under 'Satisfaction' are as follows:

EL1: I learned to use it quickly.
EL2 I easily remember how to use it.
EL3 It is easy to learn to use it.
EL4 I quickly became skilful with it.

Figure 12 below provides an illustration of Likert data distribution by % of ST1 to ST7 in a smooth lined scatter plot. First and foremost, it can be observed that there is a broad Likert data distribution by % as a whole–ranging from Likert rating of 3 points to 7 points. ST2 and ST6 have the highest Likert rating scale of 7 points with a % distribution ranging of 50.00% and 66.67% respectively. ST1, ST3 to ST5 on the other hand, has a Likert rating scale of 6 points with a % distribution ranging from 29.17% to 66.67%.

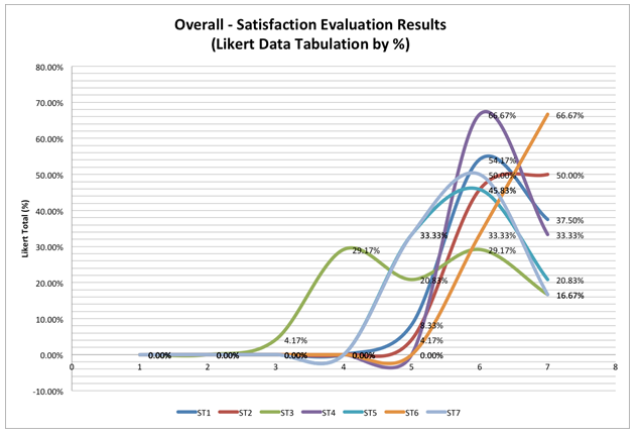

Fig. 12. Overall - satisfaction likert data distribution by %

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

VIII. CONCLUSION

The surging call for SMEs in the private transportation domain to apply data analytics in its business is evidential. However, through the research conducted and the authors' best of knowledge, there is a lack of specific DM methodology or framework specifically customised for the private transportation domain. Nor there have been evidential case studies available in this area. Therefore, there is a need to formulate a customised methodology fitting for the private transportation domain. The ECMSCO methodology was developed based on the foundational CRISP-DM methodology. In all, the proposed extended methodology outlined together with a total set of 26 generic tasks and up to 40 specialised tasks. The deliverables of the task are aimed to curb the following issues below:

1. A systematic, concise and lucid DM process for SMEs practitioners.
2. Taking into consideration the data privacy and protection legislation constraints.
3. Pre-determining the data selection prior hand–structured data only.
4. Pre-determining the DM technique selections.
5. Incorporating an overall DM-KM assessment method.

Additionally, an assessment and evaluation method known as DM-KM USE questionnaire was also formulated in order to evaluate the ECMSCO user's usability and acceptability. From the study of the overall evaluation and results collected using the DM-KM questionnaire, it was ascertained that the three SME coach operators rated the ECMSCO's 'usefulness', 'ease of use', 'ease of learning' and 'satisfaction' with average ratings of 6 to 7 Likert rating points. The overall evaluation outcome reflects that the novel ECMSCO has met with a positive response of the users' usability and acceptability. Despite the ECMSCO being synonymously accepted by the users' the research future work includes reviewing the area of improvement in the area of the method 'usefulness', 'ease of use', 'ease of learning' and 'satisfaction' which has a Likert score rating of 3 and below.

REFERENCES

[1] M. S. A. AL Khuja and Z. A. B. Mohamed, ``Investigating the adoption of e-business technology by small and medium enterprises,'' *Journal of Administrative and Business Studies*, vol. 2, no. 2, pp. 71-83, 2016. doi: https://doi.org/10.20474/jabs-2.2.3

[2] H. T. Derrick, ``The c-level is coming around on big data [infographic],'' 2012. [Online]. Available: https://bit.ly/2BQm618

[3] M. Iansiti and K. R. Lakhani, ``Digital ubiquity: How connections, sensors, and data are revolutionizing business,'' *Havard Business Review*, vol. 45, no. 2, pp. 12-39, 2014. doi: https://doi.org/10.2469/dig.v45.n2.8

[4] Eurostat, ``SMEs were the main drivers of economic growth between 2004 and 2006,'' 2009. [Online]. Available: https://bit.ly/34afmr9

[5] G. Musa, ``The role of accounting education towards the development of small and medium enterprises in Jigawa State,'' *International Journal of Business and Administrative Studies*, vol. 2, no. 4, pp. 96-102, 2016. doi: https://doi.org/10.20469/ijbas.2.10002-4

[6] Eurostat, ``Structural business statistics overview - statistics explained,'' 2018. [Online]. Available: https://bit.ly/32UKHhf

[7] European Commission, ``Big data analystical and decision making,'' 2013. [Online]. Available: https://bit.ly/2PsT1kH

[8] S. Boonvut, ``The quality financial statements of Small and Medium Enterprises Business (SME's) in view of the tax auditor,'' *International Journal of Business and Economic Affairs*, vol. 2, no. 6, pp. 335-340, 2017. doi: https://doi.org/10.24088/ijbea-2017-26002

[9] S. A. Mohd Selamat, S. Prakoonwit, R. Sahandi, W. Khan, and M. Ramachandran, ``Big data analytics-a review of data-mining models for small and medium enterprises in the transportation sector,'' *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 3, pp. 1238-1245, 2018. doi: https://doi.org/10.1002/widm.1238

[10] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, ``Big data: The management revolution,'' *Harvard Business Review*, vol. 90, no. 10, pp. 60-68, 2012.

[11] D.-W. Tan, W. Yeoh, Y. L. Boo, and S.-Y. Liew, ``The impact of feature selection: A data-mining application in direct marketing,'' *Intelligent Systems in Accounting, Finance and Management*, vol. 20, no. 1, pp. 23-38, 2013. doi: https://doi.org/10.1002/isaf.1335

[12] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ``Data mining with big data,'' *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97-107, 2013. doi: https://doi.org/10.1109/tkde.2013.109

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5

[13] S. Y. Coleman, ``Data-mining opportunities for small and medium enterprises with official statistics in the UK,'' *Journal of Official Statistics*, vol. 32, no. 4, pp. 849-865, 2016. doi: https://doi.org/10.1515/jos-2016-0044

[14] W. Fan and A. Bifet, ``Mining big data: Current status, and forecast to the future,'' *ACM IGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 1-5, 2013. doi: https://doi.org/10.1145/2481244.2481246

[15] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, ``CRISP-DM 1.0: Step-by-step data mining guide,'' 2000. [Online]. Available: https://bit.ly/36eXH3k

[16] G. Mariscal, O. Marban, and C. Fernandez, ``A survey of data mining and knowledge discovery process models and methodologies,'' *The Knowledge Engineering Review*, vol. 25, no. 2, pp. 137-166, 2010. doi: https://doi.org/10.1017/s0269888910000032

[17] K. Dnuggets, ``What main methodology are you using for your analytics, data mining, or data science projects?'' 2014. [Online]. Available: https://bit.ly/2pmks4T

[18] A. M. Lund, ``Measuring usability with the use questionnaire 12,'' *Usability Interface*, vol. 8, no. 2, pp. 3-6, 2001.

ISSN: 2414-4592
DOI: 10.20474/jater-4.6.5