Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute InformingScience.org

IJIKM.org

Volume 19, 2024

UNVEILING THE SECRETS OF BIG DATA PROJECTS: HARNESSING MACHINE LEARNING ALGORITHMS AND MATURITY DOMAINS TO PREDICT SUCCESS

Soukaina Mouhib*	Hassan II University, Casablanca, Morocco	soukaina.ine@gmail.com
Ossama Cherkaoui	Hassan II University, Casablanca, Morocco	ossama.cherkaoui@outlook.com
Houda Anoun	Hassan II University, Casablanca, Morocco	houda.anoun@gmail.com
Mohammed Ridouani	Hassan II University, Casablanca, Morocco	mohammed.ridouani@gmail.com

* Corresponding author

ABSTRACT

Aim/Purpose	While existing literature has extensively explored factors influencing the success of big data projects and proposed big data maturity models, no study has harnessed machine learning to predict project success and identify the critical features contributing significantly to that success. The purpose of this paper is to offer fresh insights into the realm of big data projects by leveraging machine-learning algorithms.
Background	Previously, we introduced the Global Big Data Maturity Model (GBDMM), which encompassed various domains inspired by the success factors of big data projects. In this paper, we transformed these maturity domains into a survey and collected feedback from 90 big data experts across the Middle East, Gulf, Africa, and Turkey regions regarding their own projects. This approach aims to gather firsthand insights from practitioners and experts in the field.
Methodology	To analyze the feedback obtained from the survey, we applied several algorithms suitable for small datasets and categorical features. Our approach included cross-validation and feature selection techniques to mitigate overfitting and enhance model performance. Notably, the best-performing algorithms in

Accepting Editor Geoffrey Z. Liu | Received: May 16, 2024 | Revised: July 27, August 4, 2024 | Accepted: August 5, 2024.

Cite as: Mouhib, S., Cherkaoui, O., Anoun, H., & Ridouani, M. (2024). Unveiling the secrets of big data projects: Harnessing machine learning algorithms and maturity domains to predict success. *Interdisciplinary Journal of Information, Knowledge, and Management, 19,* Article 22. https://doi.org/10.28945/5352

	our study were the Decision Tree (achieving an F1 score of 67%) and the Cat Boost classifier (also achieving an F1 score of 67%).
Contribution	This research makes a significant contribution to the field of big data projects. By utilizing machine-learning techniques, we predict the success or failure of such projects and identify the key features that significantly contribute to their success. This provides companies with a valuable model for predicting their own big data project outcomes.
Findings	Our analysis revealed that the domains of strategy and data have the most influential impact on the success of big data projects. Therefore, companies should prioritize these domains when undertaking such projects. Furthermore, we now have an initial model capable of predicting project success or failure, which can be invaluable for companies.
Recommendations for Practitioners	Based on our findings, we recommend that practitioners concentrate on developing robust strategies and prioritize data management to enhance the outcomes of their big data projects. Additionally, practitioners can leverage machine-learning techniques to predict the success rate of these projects.
Recommendations for Researchers	For further research in this field, we suggest exploring additional algorithms and techniques and refining existing models to enhance the accuracy and reliability of predicting the success of big data projects. Researchers may also investigate further into the interplay between strategy, data, and the success of such projects.
Impact on Society	By improving the success rate of big data projects, our findings enable organizations to create more efficient and impactful data-driven solutions across various sectors. This, in turn, facilitates informed decision-making, effective resource allocation, improved operational efficiency, and overall performance enhancement.
Future Research	In the future, gathering additional feedback from a broader range of big data experts will be valuable and help refine the prediction algorithm. Conducting longitudinal studies to analyze the long-term success and outcomes of Big Data projects would be beneficial. Furthermore, exploring the applicability of our model across different regions and industries will provide further insights into the field.

Keywords	big data projects, success prediction, key factors, maturity model, machine learning

INTRODUCTION

Despite the accelerated growth in the use of big data in recent years, particularly during the COVID-19 crisis, when the world made extensive use of big data to manage and control the pandemic (Alsunaidi et al., 2021; Bragazzi et al., 2020; Haleem et al., 2020), the implementation of big data projects is not always a success. According to Nick Heudecker, an analyst at Gartner (Asay, 2017), there are many solid reasons why big data projects fail, such as the difficulty of integrating big data with existing processes and applications, management resistance, internal politics, lack of skills, and security and governance challenges.

To help companies avoid this failure, many authors explore and expose big data success factors as fundamental keys to overcoming big data implementation challenges (Al-Sai et al., 2020; Gao et al., 2015; Soukaina et al., 2019). In addition, a few authors and software editors have proposed big data

maturity models to assess companies' big data maturity before starting implementation (Farah, 2017; Halper, 2020; Mouhib et al., 2020). Among all these attempts, we have not found a comprehensive study proposing a predictive model for the success of big data projects, which we aim to provide in this article.

In light of our research and experience in the data and analytics field, we have previously outlined the success factors that lead to the successful implementation of big data projects (Soukaina et al., 2019). Based on these success factors, we proposed the global big data maturity model (Mouhib et al., 2020), a comprehensive assessment model for measuring companies' ability to get their projects off the ground.

To complement previous work and take another step towards our big data adoption framework, we used the success factors to create a comprehensive questionnaire that we shared with big data experts from Africa, the Middle East, the Gulf, and Turkey. Thanks to the results of this survey, we were able not only to examine the factors or maturity domains that most influence the success of big data projects but also to create a model to help companies predict the success of their big data projects.

This study stands out from existing research in two key ways. First, it leverages specific characteristics aligned with our global maturity model, as proposed in prior work (Mouhib et al., 2020). Second, it is innovative by aiming to develop a predictive model using machine learning algorithms to predict big data success – an area that remains largely unexplored in the literature.

The subsequent sections of this paper are structured as follows. First, we delve into a comprehensive literature review on success factors and maturity domains. Next, we provide the context that underpins our work. Following that, we delve into the research design and methodology. Subsequently, we present and discuss the main findings and results. Finally, in the concluding section, we assess the value of this study, acknowledge its limitations, and outline potential directions for future research.

LITERATURE REVIEW

In the past few years, the importance of big data has been underscored, especially during the global pandemic of COVID-19. Organizations and governments around the world have relied heavily on big data analysis to understand better the spread of the virus, control its transmission, and predict its impact on mental health (Alsunaidi et al., 2021; Banna et al., 2023; Barbaglia et al., 2023).

However, the successful implementation of big data projects is challenging (Naeem et al., 2022). Organizations are required to deal with complexities, such as collecting, storing, processing, analyzing, and interpreting data to obtain relevant insights. In addition, determining how to improve the success of their big data initiatives is becoming a crucial topic today. In academic literature, we observe two primary trends related to big data project success. The first trend is related to critical success factors, and the second is related to maturity models.

Regarding success factors, researchers have linked the success of big data projects to specific critical success factors (CSFs), also known as key factors, adoption factors, or influencing factors. These factors are often identified through comprehensive literature reviews (Jonathan & Raharjo, 2024; Saltz & Shamshurin, 2016; Sun et al., 2018). Other authors tried to validate these factors through surveys (Lucas, 2019; Rahman, 2016; Santoso, 2023) or methods such as ABC analysis (Koronios et al., 2014). Other authors concentrate on identifying their categories (Al-Sai et al., 2020; Eybers & Hattingh, 2017; Surbakti et al., 2020), and others focus on a specific success factor such as organizational alignment (Chang et al., 2017; Kiron, 2013) or governance (Brous et al., 2020; Veeneman et al., 2018).

Regarding maturity models, they play a crucial role in assessing companies' ability to launch big data projects, consequently increasing their chances of success. Some models are designed by big data practitioners, who often propose challenging models. These models are typically supported by survey or assessment tools such as Transforming Data With Intelligence (Halper, 2020) and Hortonworks

(Dhanuka, 2016). Other models are proposed by authors, such as the value Base Maturity Model (Farah, 2017), the Zakat Maturity Model (Sulaiman et al., 2015) or the Temporal Maturity Model (Olszak & Mach-Król, 2018).

Today, several papers propose industry-specific models to address the assessment of Artificial Intelligence (AI) and big data implementations, particularly within the process industry. A recent study by Fornasiero et al. (2024) compares various maturity models relevant to AI and BD in the industry sector (Alsheiabni et al., 2019; Colangelo et al., 2022; Hausladen & Schosser, 2020; Hortovanyi et al., 2023). These maturity models also evaluate various maturity-related dimensions, such as strategy, governance, technology, and analytics, all within the context of the process industry.

In a previous work (Mouhib et al., 2020), we studied existing maturity models used across various sectors and proposed our Global Big Data Maturity Model (GBDMM). To illustrate the discrepancies between existing maturity models, Table 1 shows the global sub-domains in 15 of the existing maturity models.

Table 1. Global domains occurrence within 15 maturity models

Global maturity sub-domains	**Maturity models references**														
	(Schmarzo, 2013)	(van Veenstra et al., 2013)	(El-Darwiche et al., 2014)	(Radcliffe, 2014)	(Halper, 2020)	(Vesset & Xiong, 2015)	(Nott, 2022)	(Sulaiman et al., 2015)	(Comuzzi & Patel, 2016)	(Farah, 2017)	(Olszak & Mach-Król, 2018)	(Hausladen & Schosser, 2020)	(Mouhib et al., 2020)	(Dhanuka, 2016)	(Bond et al., 2013)
Strategy	x	x	x	x		x	x	x	x	x		x	x	x	x
Processes	x		x	x		x			x				x	x	
Data analytics		x	x	x	x		x	x	x		x		x	x	
Data management			x	x	x	x		x	x	x	x	x	x	x	x
Information technology				x			x		x			x	x		
IT infrastructure					x	x	x		x		x		x	x	x
People		x	x	x	x	x			x				x	x	x
Culture			x		x		x		x		x		x		
Governance				x	x		x		x	x			x		x
Methodology													x		

The main difference between all the models lies in the big data domains covered by each model, maturity levels, and assessment tools. Different models vary in the domains they address. Some offer comprehensive coverage across multiple domains (Comuzzi & Patel, 2016; Halper, 2020), while others focus on specific areas (van Veenstra et al., 2013). Maturity models typically have different levels (e.g., 4, 5, or 6) representing the organization's progression in adopting big data practices. These levels often range from initial (low maturity) to optimized (high maturity). Regarding the assessment tools, models like TDWI (Halper, 2020) and IDC (Vesset & Xiong , 2015) come with assessment instruments that help organizations evaluate their maturity. These tools automatically calculate a maturity score based on users' responses to specific questions.

The Global Big Data Maturity Model aims for thoroughness and includes project methodology as a critical maturity domain. It covers six global areas: Strategy Alignment, Data, People, Governance, Technology, and Methodology, and defines five maturity levels: Ad Hoc, Explore, Transformation, Adoption, and Maturity. It comes with an assessment framework that offers a global score and identifies areas for improvement to enhance companies' chances of success in big data initiatives (Mouhib et al., 2023b).

As a new step in our work, we intend to incorporate a prediction module into our framework. To the best of our knowledge, there is no other paper dealing with predicting big data success, hence the originality of this article in filling this research gap.

WORK BACKGROUND AND MOTIVATION

Previously, we aimed to create a framework to enhance big data adoption, providing companies with insights into their readiness for starting big data projects. We began our study by exploring and classifying the factors impacting big data projects (Soukaina et al., 2019). We identified six key categories: strategy alignment, data, people, governance, technology, and methodology. Then, we mapped these success factors to maturity domains within maturity models.

Through extensive analysis of literature and software vendor documents related to big data maturity models, we proposed a comprehensive model – the Global Big Data Maturity Model (Mouhib et al., 2020). This model allows for a thorough assessment of an organization's maturity, incorporating the methodology dimension as a critical factor for successfully driving big data projects (a dimension often missing in existing models). The high-level design of the GBDMM is illustrated in Figure 1.

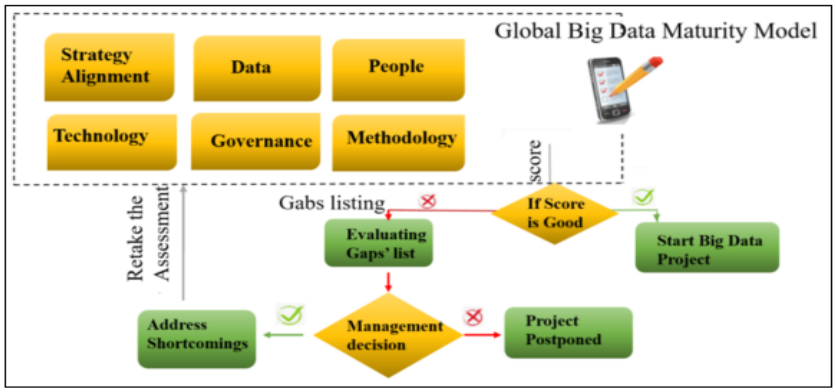

Figure 1. Global Big Data Maturity Model design (Mouhib et al., 2020)

Building upon this foundation, we developed a corresponding big data assessment framework. This framework provides final scores representing an organization's readiness to implement big data initiatives (Mouhib et al., 2023b) (see Figure 2 for a visual representation of the output of the assessment framework results).

To interpret the framework scores, we proposed Table 2, which presents the maturity level corresponding to each global score with a breakdown of scores per dimension (Mouhib et al., 2023a).

To obtain a more precise score and define the importance of each maturity area, we used the Analytic Hierarchy Process (AHP) technique in conjunction with big data professionals inputs (Mouhib et al., 2023a). The weightings for each area, in descending order, were as follows: data (0.278), governance (0.190), strategy alignment (0.173), methodology (0.166), people (0.128), and technology (0.064) (see Figure 3).

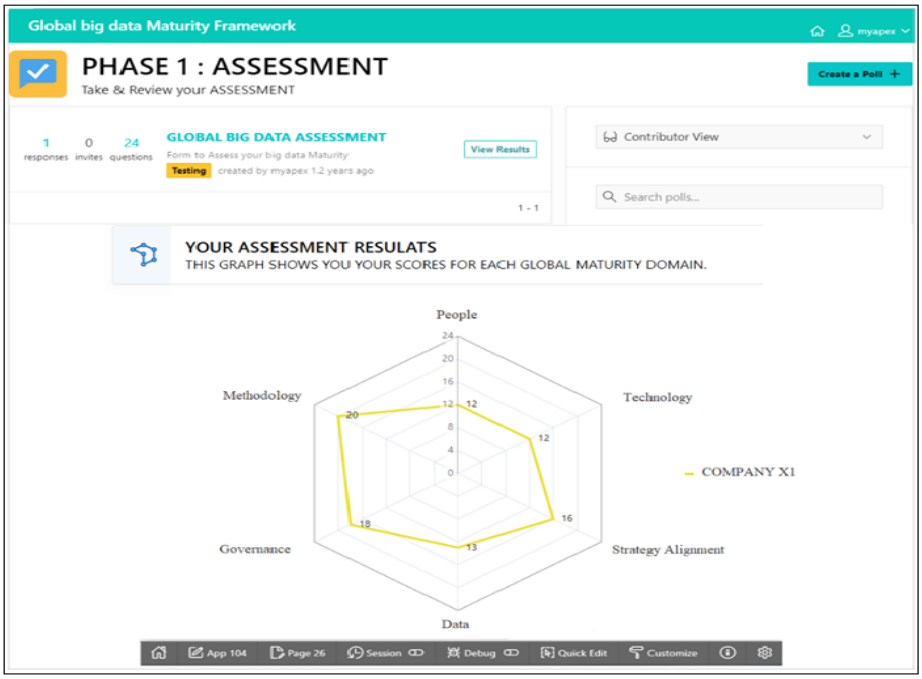

Figure 2. Global assessment framework outputs

Table 2. Maturity levels with corresponding global and domain scores

Maturity levels	Domain's score	Global score
Ad-hoc	=<4	=<24
Explore	>4 & =<8	>24 & =<48
Transformation	>8 & =<12	>48 &=<72
Adoption	>12 & =<16	>72 &=<96
Maturity	>16 & =<20	>96 & =<120

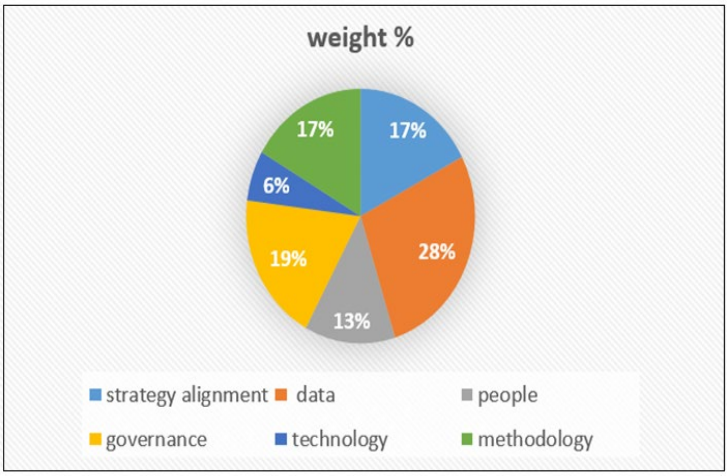

Figure 3. Global domains' weights percentage (Mouhib et al., 2020)

This approach allowed us to calculate accurate weighted scores. The AHP rankings revealed that the data dimension exerts the most significant impact on the success of big data projects, followed by strategy, methodology, governance, people then, and technology.

Taking this research a step further, we aim to expand the assessment framework and develop a machine-learning model capable of predicting the success or failure of companies in adopting their big data projects. Leveraging the global maturity domains and success factors, we conducted a comprehensive survey among experts from Africa, the Middle East, the Gulf, and Turkey. By utilizing machine-learning techniques, we not only identify influential features for project success but also empower companies to predict their likelihood of success in big data initiatives.

The driving force behind this initiative lies in the current lack of prediction models for big data success in the existing literature. Our proposed model aims to fill this gap, equipping organizations and stakeholders with data-driven insights to make informed decisions. Among the advantages of this approach are:

- Organizations can optimize resource allocation. Big data projects require substantial infrastructure, technology, and human resources investments. Consequently, accurately predicting the success of these projects may enable organizations to allocate strategically their resources, minimizing wasted efforts and maximizing return on investment.
- By gaining insights into big data success, organizations can create effective frameworks and guidelines for implementation. For example, if data governance significantly influences project success, prioritizing robust data governance frameworks enhances the chances of achieving overall success.

METHODOLOGY

Based on the Global maturity domains, we developed a machine-learning algorithm to predict the success of big data projects. To achieve that, we tested several machine-learning algorithms adapted to categorical features. We used feature selection and cross-validation techniques to remove irrelevant features and improve algorithm performance.

DATA UNDERSTANDING

In the context of machine learning implementation, data understanding plays a pivotal role. This step entails a comprehensive analysis and thorough familiarity with the data you will be working with. Following are the steps we went through.

Data collection

Based on the Global Big Data Maturity Model, we have designed a survey inspired by Global maturity domains, as illustrated in Table 3. We added two more questions about the company size and big data project outcome to complete our survey: What is your company size? Did you reach your big data project objectives?

We sent the survey to big data and analytics experts in the Middle East, Africa, the Gulf, and Turkey regions to investigate the status of current projects. We collected 90 responses that we used to train and run the first version of the predictive model.

Table 3. Global domains and corresponding survey questions

Global maturity models	Corresponding questions
Strategy Alignment	What level of support do you get from your management?
	Is Big Data part of your business strategy?
	Did you allocate the necessary budget to the project?
Methodology	Business and IT work together to identify use cases.
	Have all use case specifications and outcomes or deliverables been thoroughly identified?
	Do you use agile team methodology to implement big data?
	Do you use an iterative development approach to collect requirements and implement use cases?
Data	Is use cases' data available?
	Do you have a data lake that consolidates all data sources and types?
	Do you have well-defined data lifecycle management (collect, store, process and analyze data)?
Governance	Do you have well-known processes, and practices to protect and secure sensitive data?
People	Do you have a multidisciplinary team experienced in big data technologies? Do you use external resources to implement the project? Only internal ones?
	Does the company use analytics for daily operations to monitor and drive business growth?
Technology	Do you employ advanced analytics and utilize data exploration visualization?
	Do you have prebuilt Analytics solutions/ ML or custom development to implement the use cases?
	Do you use cloud services or on-premise solutions for your big data analytics platform?
	Do you ingest and process real-time data (sensors, events, media, and logs)?
	Does your current platform handle the data volume needed for the actual big data use cases and eventual growth?
	Do you use data quality tools to ensure the quality of data?
	Do you have data integration services or tools to support your different sources/targets?

Data Exploration

The first step in machine learning analysis is to obtain an overview of the collected data using a data visualization tool. Figure 4 illustrates the distribution of features within our dataset. For example, more than 50% of companies reported achieving more than 60% of their objectives; more than 70% of companies allocated a budget to their project, and the majority utilized a team methodology and processed sensitive data.

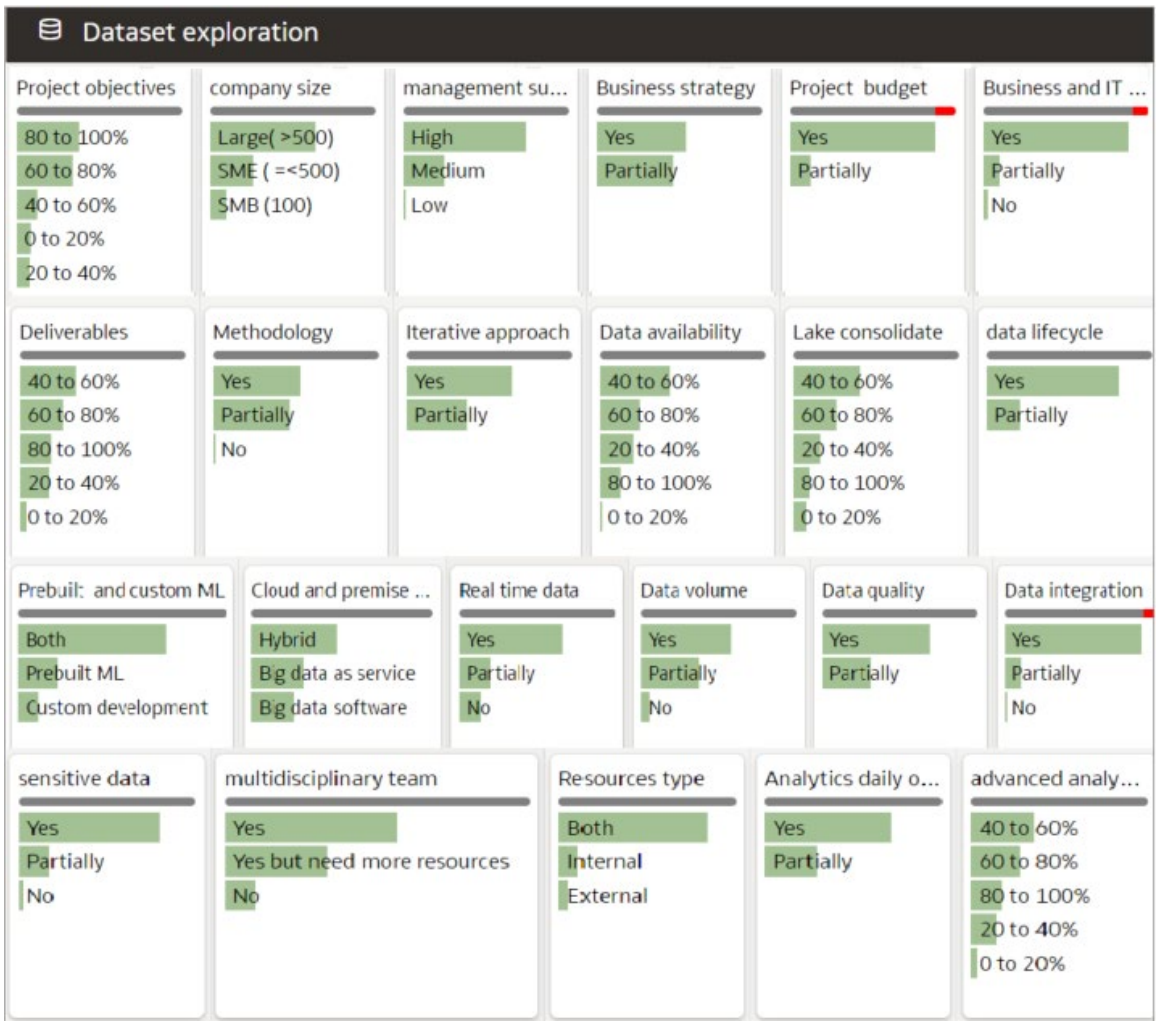

Figure 4. Global domains' weights percentage (Mouhib et al., 2023b)

For better data exploration, we created a visualization project with several graphs to better understand our target attribute: project objectives (Figure 5).

The pie chart (a) illustrates the distribution of project objectives values: 37.5% of companies achieve 80-100%, 33.75% achieve 60-80%, and 28.75% achieve less than 60% of their objectives. In addition, the bar chart (b) shows how the project objectives attribute is distributed relatively to the management support attribute.

The bar chart (b) illustrates how the project objectives attribute is distributed based on management support. Notably, companies with medium to high management support are more likely to achieve their big data project objectives. For instance, all companies with high management support have project objectives above 40%.

Additionally, graphics (c) and (d) show the distribution of the project objectives attribute according to methodology and multidisciplinary teams. Companies that either partially or fully adopt team methodologies and have multidisciplinary teams tend to be more successful in achieving their big data project objectives.

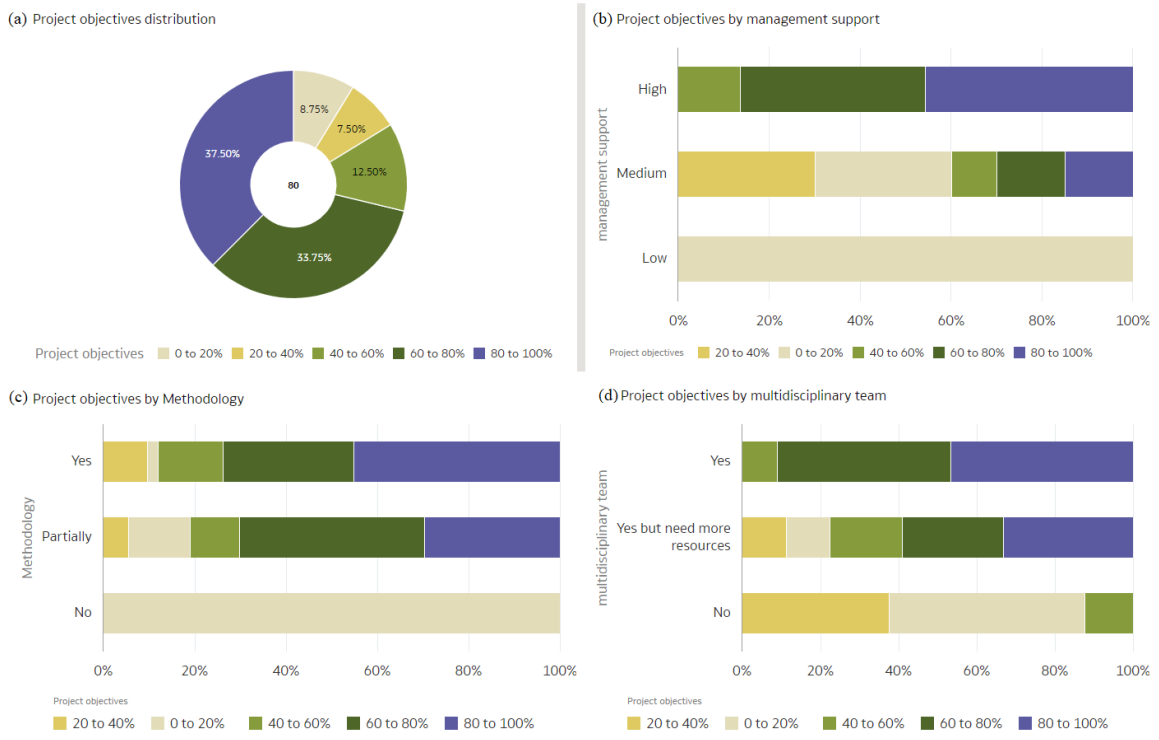

Figure 5. Project objective correlation with big data domains

DATA PREPARATION

The dataset used in this study resulted from a survey conducted with over a hundred big data specialists across Africa, Turkey, the Gulf, and the Middle East. Ninety responses were collected and used for training and testing the model.

In this analysis, we utilized 21 features, including those derived from global big data maturity domains and company size. The success level of the Big Data project serves as the prediction target. This success level is represented by the feature Project Objectives, which falls into the following percentage ranges: 0% to 20% (unsuccessful), 20% to 40% (moderately successful), 40 to 60% (successful), 60% to 80% (very successful), and 80% to 100% (exceptionally successful). Table 4 provides a comprehensive list of all the features used for prediction, along with their corresponding classes.

The input data is prepared based on the chosen classification algorithm. For decision tree-based algorithms, categorical features are converted into numerical values using an ordinal encoder. In contrast, Logistic Regression and Support Vector Machine classification methods necessitate transforming categorical data into binary vectors using one-hot encoding. However, the CatBoost algorithm can handle categorical data directly without any specific transformation.

The algorithms were tuned and trained on a random sample representing 70% of the data. The remaining 30% of the data serves as a test set, which remains untouched to facilitate performance comparisons between the algorithms.

Table 4. Big data prediction model features

Feature	Description	Classes
Company size	Small, medium, or large company	SMB (100), SME (≤500), Large (>500)
Management support	Level of management support	Low, Medium, High
Business strategy	If big data is part of company's business strategy.	No, Partially, Yes
Project budget	If the project budget is allocated.	No, Partially, Yes
Business and IT collaboration	If business and IT work together to identify big data use cases.	No, Partially, Yes
Project outcomes	If use case specifications and deliverables are well-identified.	0-20%, 20-40%, 40-60%, 60-80%, 80-100%
Methodology	If an agile team methodology to implement big data is used.	No, Partially, Yes
Iterative approach	If an iterative development approach to implement big data is used.	No, Partially, Yes
Data availability	If use cases' data is available.	0-20%, 20-40%, 40-60%, 60-80%, 80-100%
Lake consolidate	If a data lake for data consolidate is used.	0-20%, 20-40%, 40-60%, 60-80%, 80-100%
Data lifecycle	If data lifecycle is defined.	No, Partially, Yes
Multidisciplinary team	If a multidisciplinary experienced team exists in the company.	No, Yes, but need more resources, Yes
Resources type	Type of resources that are going to implement the project.	Internal, External, Both
Analytics daily operations	If analytics is used for daily operations to drive business growth.	No, Partially, Yes
Advanced analytics	If advanced analytics and data exploration visualization are used.	0-20%, 20-40%, 40-60%, 60-80%, 80-100%
Prebuilt and custom ML	If prebuilt analytics ML is used or custom development.	Custom development, Prebuilt ML, Both
Cloud and premise solutions	Type of deployment of big data platform cloud/on-premise.	Big data software, Big data as service, Hybrid
Real-time data	If the company ingests and processes real-time data.	No, Partially, Yes
Data volume	If the platform can handle all use case data with the future growth.	No, Partially, Yes
Data quality	If data quality is used.	No, Partially, Yes
Data integration	If data integration tools are used.	No, Partially, Yes

Feature selection

In the present study, we employed feature selection techniques to eliminate redundant and irrelevant features from the dataset. The goal was to enhance the predictive performance of the model. In the literature, various approaches to feature selection have been proposed and can be classified into three main categories (Chandrashekar & Sahin, 2014).

Filter methods are applied before the training phase to identify correlations between features and rank them based on their dependence on the target variable. For numerical data, it is common to use variable dependency measures such as Pearson's correlation coefficient (Guyon &

Elisseeff, 2003) to detect highly correlated features that are considered redundant. Similar methods are employed to eliminate features that do not significantly influence the target variable. Specific measures are used for categorical features, which are based on statistical tests (Yang & Pedersen, 1997) or information theory criteria (Brown et al., 2012).

Wrapper methods [34] use classification performance as an objective function to select the optimal feature subset. With high-dimensional datasets, this becomes computationally very expensive as the number of features' subsets to explore increases exponentially. Wrapper methods can be applied sequentially or based on a heuristic to explore the space of all possible feature subsets (Chandrashekar & Sahin, 2014).

Embedded methods perform feature selection during the training phase. To achieve this, they integrate the feature selection criteria into the algorithm's objective function, which can lead to the elimination of features that are irrelevant to the optimization. Examples of this approach include tree-based algorithms such as Decision Tree and CatBoost, which perform embedded variable elimination and generate feature importance scores at the end of the training phase.

The present study employs the Mutual Information (Brown et al., 2012) filter method for feature selection, given that all features are categorical. The result of the feature selection stage is a ranking of the features according to their predictive power based on the assigned scores. Consequently, the top 10 features were utilized to conduct the tests. Figure 6 presents the feature ranking generated by the feature selection phase.

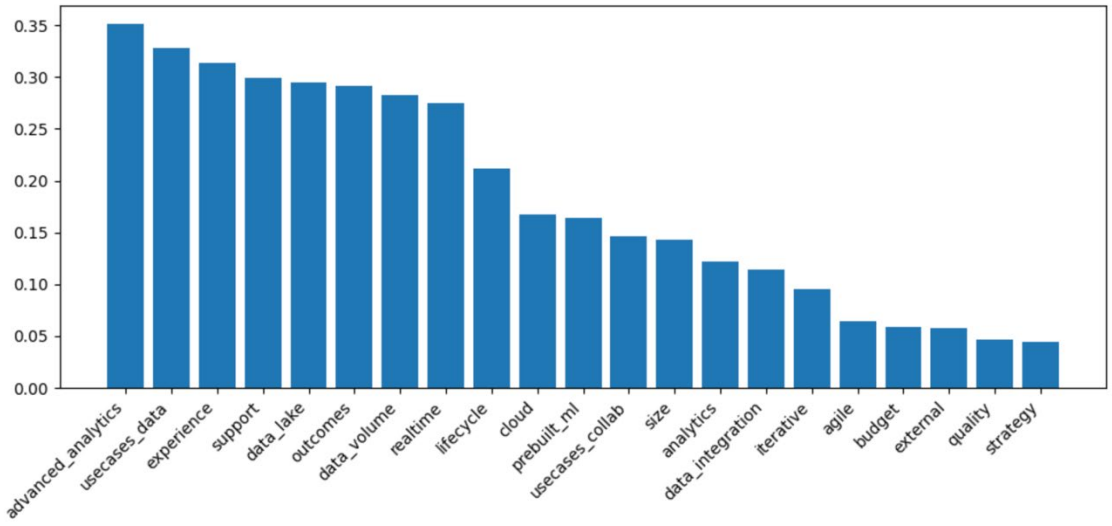

Figure 6. Mutual information feature selection ranking

Modeling – Algorithm Selection and Evaluation Metrics

Algorithm selection and tuning

In the present study, five classification algorithms that are commonly employed with categorical data were utilized to predict the success of Big Data projects. They are as follows: Decision Tree Classification (DT) (Günlük et al., 2021), Logistic Regression (LR) (Peng et al., 2001), Categorical Naive Bayes (CNB) (Maia et al., 2021), Support Vector Machine Classification SVM (Hancock & Khoshgoftaar, 2020), CatBoostClassifier (CB) (Dorogush et al., 2018), and Random Forest Classification (RF) (Breiman, 2001).

The hyperparameters of the tested algorithms were optimized using K-fold cross-validation. This technique involves iteratively partitioning the training data into disjoint subsets for training and validation. The goal is to identify optimal parameter values. Notably, this approach improved model performance and mitigated overfitting. Table 5 presents the optimal hyperparameter values for each algorithm after cross-validation, utilizing the grid search method with two folds (Pedregosa et al., 2011).

Table 5. Optimal hypermeter results for each algorithm

Model	Best parameter values
Decision Tree Classification	criterion: entropy, max_depth: 3, min_samples_leaf: 1
Logistic Regression	C: 10, max_iter: 1000
Categorical Naive Bayes	alpha: 0.1
Support Vector Machine Classification	C: 1, kernel: linear
CatBoost Classier	depth: 6, l2_leaf_reg: 3, iterations: 300, learning_rate: 0.5
Random Forest	criterion: entropy, max_depth: 6, min_samples_leaf: 1, n_estimators: 100

Evaluation metrics

The purpose of evaluation metrics is to assess the performance of classification algorithms and provide a basis for comparative analysis. In this study, we calculated the following performance metrics to evaluate the algorithms (Hossin & Sulaiman, 2015):

- **Accuracy** is the total success rate of the prediction. It is defined as the ratio between the number of correct predictions and the total number of instances.
- **Precision** is the ratio of correct predictions for a specific target class to the total number of predictions. It quantifies the degree of confidence associated with the prediction outcomes.
- **Recall** is the ratio of correct predictions for a specific target class to the number of instances of that class. It serves to quantify the efficacy of the classification algorithm in making predictions.
- **F1 score** is the harmonic mean of the precision and recall for a specific target class.

Except for accuracy, the metrics mentioned are target class-specific. To obtain an overall performance evaluation score, it is essential to calculate the average across all classes for each metric. The macro-average method computes the unweighted mean of individual metrics per class. In contrast, the micro-average method sums true positives, false negatives, and false positives across all classes to derive precision, recall, and F1 scores. The present study employs the micro-average method.

RESULTS

Two tests were conducted for all models: one with the top 10 features according to the Mutual Information feature selection method and one with all features. Table 6 illustrates the obtained performance metric values for each algorithm. The results demonstrate that superior performance was achieved when no feature selection was employed.

The algorithms demonstrating the best performance are the Decision Tree (F1 = 67%) and the CatBoost Classifier (F1 = 67%), followed by the Categorical Naive Bayes (F1 = 62%). The Logistic Regression, Random Forest, and Support Vector Machines algorithms exhibited an overfitting issue during the training phase, with F1 scores of 100%, 95%, and 93%, respectively.

The results indicate that feature selection does not enhance the accuracy of the predictions. In fact, the inclusion of all features yielded a superior performance. This finding aligns with the recommendations of previous researchers who suggest augmenting the data features to optimize model performance (Li & Liu, 2012).

Table 6. Model metrics values with/without feature selection

	Metrics /ml	Decision tree classification	Logistic regression	Categorical naïve bayes	SVM classification	CatBoost Classifier	Random forest
All features	Precision	**67%**	54%	62%	46%	**67%**	58%
	Recall	**67%**	54%	62%	46%	**67%**	58%
	Accuracy	**67%**	54%	62%	46%	**67%**	58%
	F1 Score	**67%**	54%	62%	46%	**67%**	58%
Mutual information feature selection (top 10 features)	Precision	58%	42%	50%	42%	58%	38%
	Recall	58%	42%	50%	42%	58%	38%
	Accuracy	58%	42%	50%	42%	58%	38%
	F1 Score	58%	42%	50%	42%	58%	38%

DISCUSSION

As previously mentioned in the results section, the most effective algorithms for this particular use case are the Decision Tree and CatBoost algorithms. The CatBoost algorithm is distinguished by its ability to perform integrated feature elimination and to provide feature importance scores. In this section, we present the findings from one of each of those algorithms.

CATBOOST PREDICTION FINDINGS

This algorithm has identified a ranking of features that are pivotal for project success. The top 10 features, ranked by impact, are as follows:

1. Lake consolidation
2. Multidisciplinary team
3. Project outcomes
4. Methodology
5. Real-time data
6. Management support
7. Prebuilt machine learning
8. Data volume
9. Company size
10. Resource type

This finding holds immense significance for companies. Recognizing these features is crucial, as they represent the essence of Big Data maturity domains to focus on. Companies are advised to cultivate advanced capabilities in specific areas (data lakes, multidisciplinary teams, project outcomes, methodologies, and others) to bolster their prospects for success. In other words, companies need to achieve a high maturity level in domains and subdomains related to data, analytics, people, methodology, strategy alignment, and infrastructure to increase their chances of succeeding in their big data projects.

Furthermore, this ranking validates our existing Global Big Data Maturity Model by emphasizing the importance of each global domain. Additionally, the results highlight a new dimension – Company Size – as a fundamental predictor of project outcomes. This insight not only reinforces the model's validity but also provides guidance for companies on where to focus their developmental efforts.

Decision Tree Prediction Findings

Although we initially trained the algorithm with all features, our hyperparameter tuning limited the number of levels in the decision tree, consequently reducing the number of features used. Decision tree offers transparent feature importance, providing a clear indication of the significance of different features or variables in the decision-making process. By observing the splits in the tree, we can identify which features had the most significant influence on predictions.

This understanding helps pinpoint key factors affecting the model's outcomes. Notably, decision trees reveal interesting conjunctions of features. Furthermore, decision trees excel in generating interpretable models, offering an intuitive, transparent, and easily explainable framework for decision-making. According to Figure 7, we can deduce the following rules:

1. If management support is Low or Medium (<=1.5) and advanced analytics adoption is less than 40%, then success likelihood is between 0% and 20%.
2. If management support is Low or Medium (<=1.5), advanced analytics adoption is greater than 40%, and project outcomes identification is less than 60%, then success likelihood is between 20% and 40%.
3. If management support is Low or Medium (<=1.5), advanced analytics adoption is greater than 40%, and project outcomes identification is greater than 60%, then success likelihood is between 60% and 80%.
4. For a small company, if management support is High (>1.5), the success likelihood is between 60% and 80%.
5. For a medium or large company, if management support is High (>1.5) and data availability is less than 40%, the success likelihood is between 60% and 80%.
6. For a medium or large company, if management support is High (>1.5) and data availability is greater than 40%, the success likelihood is between 80% and 100%.

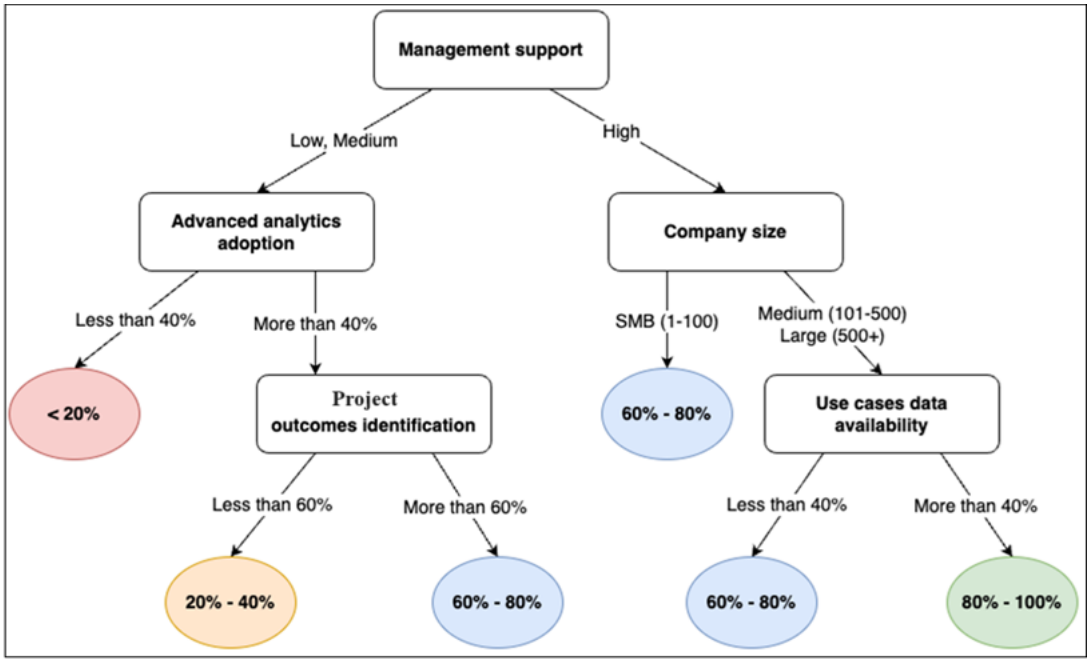

Figure 7. Decision tree trained with all Prediction Model features

This model shows that companies are most likely to succeed in implementing their big data projects when they have full management support for their big data project, and when they have the use case data available. Companies also stand a higher chance to succeed when advanced analytics are widely used across different company departments, and when use case specifications and results are well identified. In other words, the maturity domains that significantly influence the success of a big data project are strategy alignment, data, and methodology. Consequently, companies must focus on evolving in these areas to enhance their chances of success. The final crucial insight from the model is that success criteria and the domains to prioritize can vary based on company' size.

- For small companies, management support is sufficient for a good chance of success.
- For large companies, management support and the availability of use case data both matter to have the best chances of success in Big Data projects

CONCLUSION

This work represents a significant milestone toward our big data adoption framework. The study demonstrates promising results for the big data prediction model. Using decision trees and Catboost Classifier, we achieved acceptable performance and, most importantly, were able to identify a feature set that can drive companies toward successful implementations (Management support and use cases data availability).

Nevertheless, we acknowledge that due to the challenges associated with obtaining a larger dataset and the nature of the features, we were unable to reach the desired level of performance (67% accuracy and precision). This can be improved by employing techniques suggested by authors in the literature, such as the extension of the attribute information method, which is known to deliver superior classification performance (Li & Liu, 2012). Additionally, exploring different encoding techniques and testing various machine learning algorithms and ensemble methods could also lead to performance improvements.

In future studies, we plan to investigate these enhancement aspects further and analyze the impact of these findings on our Global Maturity Model and Big Data Adoption Framework.

REFERENCES

Al-Sai, Z. A., Abdullah, R., & Husin, M. H. (2020). Critical success factors for big data: A systematic literature review. *IEEE Access*, *8*, 118940–118956. https://doi.org/10.1109/ACCESS.2020.3005461

Alsheiabni, S., Cheung, Y., & Messom, C. (2019, July). Towards an artificial intelligence maturity model: From science fiction to business facts. *Proceedings of the Pacific Asia Conference on Information Systems,Xi'an, China,* 46. https://aisel.aisnet.org/pacis2019/46

Alsunaidi, S. J., Almuhaideb, A. M., Ibrahim, N. M., Shaikh, F. S., Alqudaihi, K. S., Alhaidari, F. A., Khan, I. U., Aslam, N., & Alshahrani, M. S. (2021). Applications of big data analytics to control COVID-19 pandemic. *Sensors*, *21*(7), 2282. https://doi.org/10.3390/s21072282

Asay, M. (2017, November 10). *85% of big data projects fail, but your developers can help yours succeed.* TechRepublic. https://www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/

Banna, Md. H. A., Ghosh, T., Nahian, Md. J. A., Kaiser, M. S., Mahmud, M., Taher, K. A., Hossain, M. S., & Andersson, K. (2023). A hybrid deep learning model to predict the impact of COVID-19 on mental health from social media big data. *IEEE Access*, *11*, 77009–77022. https://doi.org/10.1109/ACCESS.2023.3293857

Barbaglia, L., Frattarolo, L., Onorante, L., Pericoli, F. M., Ratto, M., & Tiozzo Pezzoli, L. (2023). Testing big data in a big crisis: Nowcasting under Covid-19. *International Journal of Forecasting*, *39*(4), 1548–1563. https://doi.org/10.1016/j.ijforecast.2022.10.005

Bond, S., Grace, K., Ko, D., & Mcleod, R. (2013). *Big data maturity assessment tool.* Infotech. https://www.infotech.com/research/it-big-data-maturity-assessment-tool

Bragazzi, N. L., Dai, H., Damiani, G., Behzadifar, M., Martini, M., & Wu, J. (2020). How big data and artificial intelligence can help better manage the COVID-19 pandemic. *International Journal of Environmental Research and Public Health*, *17*(9), 3176. https://doi.org/10.3390/ijerph17093176

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brous, P., Janssen, M., & Krans, R. (2020). Data governance as success factor for data science. In M. Hattingh, M. Matthee, H. Smuts, I. Pappas, Y. K. Dwivedi, & M. Mäntymäki (Eds.), *Responsible design, implementation and use of information and communication technology* (pp. 431–442). Springer. https://doi.org/10.1007/978-3-030-44999-5_36

Brown, G., Pocock, A., Zhao, M.-J., & Luján, M. (2012). Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, *13*, 27–66.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chang, D.-I., Kim, J.-H., & Park, M.-J. (2017). A study on organizational design and operational planning of big data teams. *International Journal of Applied Engineering Research, 12,* 9835–9845.

Colangelo, E., Fries, C., Hinrichsen, T.-F., Szaller, Á., & Nick, G. (2022). Maturity model for AI in smart production planning and control system. *Procedia CIRP*, *107*, 493–498. https://doi.org/10.1016/j.procir.2022.05.014

Comuzzi, M., & Patel, A. (2016). How organisations leverage big data: A maturity model. *Industrial Management & Data Systems*, *116*(8), 1468-1492. https://doi.org/10.1108/IMDS-12-2015-0495

Dhanuka, V. (2016). *Hortonworks big data maturity assessment model: The strategic path to accelerating business transformations.* Hortonworks. https://hortonworks.com/wp-content/uploads/2016/04/Hortonworks-Big-Data-Maturity-Assessment.pdf

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: Gradient boosting with categorical features support.* arXiv:1810.11363. https://doi.org/10.48550/arXiv.1810.11363

El-Darwiche, B., Koch, V., Meer, D., Shehadi, R. T., & Tohme, W. (2014). Big data maturity: An action plan for policymakers and executives. *The Global Information Technology Report* (pp. 43-50). World Economic Forum. https://www3.weforum.org/docs/GITR/2014/GITR_Chapter1.3_2014.pdf

Eybers, S., & Hattingh, M. J. (2017, May). Critical success factor categories for big data: A preliminary analysis of the current academic landscape. *Proceedings of the IST-Africa Week Conference*, *Windhoek, Namibia.* https://doi.org/10.23919/ISTAFRICA.2017.8102327

Farah, B. (2017). A Value based big data maturity model. *Journal of Management Policy and Practice*, *18*(1), 11–18.

Fornasiero, R., Kiebler, L., Falsafi, M., & Sardesai, S. (2024). Proposing a maturity model for assessing artificial intelligence and big data in the process industry. *International Journal of Production Research.* https://doi.org/10.1080/00207543.2024.2372840

Gao, J., Koronios, A., & Selle, S. (2015). Towards a process view on critical success factors in big data analytics projects. *Proceedings of the Twenty-first Americas Conference on Information Systems*, *Puerto Rico.* https://core.ac.uk/download/pdf/301365683.pdf

Günlük, O., Kalagnanam, J., Li, M., Menickelly, M., & Scheinberg, K. (2021). Optimal decision trees for categorical data via integer programming. *Journal of Global Optimization*, *81*, 233–260. https://doi.org/10.1007/s10898-021-01009-y

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Haleem, A., Javaid, M., Khan, I. H., & Vaishya, R. (2020). Significant applications of big data in COVID-19 pandemic. *Indian Journal of Orthopaedics*, *54*(4), 526–528. https://doi.org/10.1007/s43465-020-00129-z

Halper, F. (2020). *TDWI analytics maturity model: Assessment guide.* TDWI Research. https://go.tdwi.org/rs/626-EMC-557/images/TDWI_Analytics-Maturity-Model-Assessment-Guide_2020.pdf

Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*, Article 28. https://doi.org/10.1186/s40537-020-00305-w

Hausladen, I., & Schosser, M. (2020). Towards a maturity model for big data analytics in airline network planning. *Journal of Air Transport Management*, *82*, 101721. https://doi.org/10.1016/j.jairtraman.2019.101721

Hortovanyi, L., Morgan, R. E., Herceg, I. V., Djuricin, D., Hanak, R., Horvath, D., Mocan, M. L., Romanova, A., & Szabo, R. Z. (2023). Assessment of digital maturity: The role of resources and capabilities in digital transformation in B2B firms. *International Journal of Production Research*, *61*(23), 8043–8061. https://doi.org/10.1080/00207543.2022.2164087

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2). https://doi.org/10.5121/ijdkp.2015.5201

Jonathan, B., & Raharjo, T. (2024). Big data project success factors: A systematic literature review. *AIP Conference Proceedings*, *3109*(1), 30018. https://doi.org/10.1063/5.0205495

Kiron, D. (2013). Organizational alignment is key to big data success. *MIT Sloan Management Review.* https://sloanreview.mit.edu/article/organizational-alignment-is-key-to-big-data-success/

Koronios, A., Gao, J., & Selle, S. (2014). Big data project success - A meta analysis. *Proceedings of the Pacific Asia Conference on Information Systems.* https://aisel.aisnet.org/pacis2014/376

Li, D.-C., & Liu, C. (2012). Extending attribute information for small data set classification. *IEEE Transactions on Knowledge and Data Engineering*, *24*(3), 452–464. https://doi.org/10.1109/TKDE.2010.254

Lucas, A. (2019). Critical success factors for corporate data quality management. In Á. Rocha, H. Adeli, L. P. Reis, & S. Costanzo (Eds.), *New knowledge in information systems and technologies* (pp. 630–644). Springer. https://doi.org/10.1007/978-3-030-16181-1_60

Maia, M. R. de H., Plastino, A., & Freitas, A. A. (2021, December). An ensemble of Naive Bayes classifiers for uncertain categorical data. *Proceedings of the IEEE International Conference on Data Mining*, *Auckland, New Zealand*, 1222–1227. https://doi.org/10.1109/ICDM51629.2021.00148

Mouhib, S., Anoun, H., Ridouani, M., & Hassouni, L. (2020, October). Towards a global big data maturity model. *Proceedings of the Fourth International Conference on Intelligent Computing in Data Sciences*, *Fez, Morocco,* 1–5. https://doi.org/10.1109/ICDS50568.2020.9268720

Mouhib, S., Anoun, H., Ridouani, M., & Hassouni, L. (2023a). Analyzing the global big data maturity model domains for better adoption of big data projects. *International Journal of Information Science and Management*, *21*(4), 83–102. https://doi.org/10.22034/ijism.2023.1977940.0

Mouhib, S., Anoun, H., Ridouani, M., & Hassouni, L. (2023b). Global big data maturity model and its corresponding assessment framework results. *IAENG International Journal of Applied Mathematics, 53*(1).

Naeem, M., Jamal, T., Diaz-Martinez, J., Butt, S. A., Montesano, N., Tariq, M. I., De-la-Hoz-Franco, E., & De-La-Hoz-Valdiris, E. (2022). Trends and future perspective challenges in big data. In J.-S. Pan, V. E. Balas, & C.-M. Chen (Eds.), *Advances in intelligent data analysis and applications* (pp. 309–325). Springer. https://doi.org/10.1007/978-981-16-5036-9_30

Nott, C. (2022). *A maturity model for big data and analytics.* Whitehall Media. https://whitehallmedia.co.uk/blog/2015/12/29/a-maturity-model-for-big-data-and-analytics/

Olszak, C., & Mach-Król, M. (2018). A conceptual framework for assessing an organization's readiness to adopt big data. *Sustainability*, *10*(10), 3734. https://doi.org/10.3390/su10103734

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, C.-Y. J., Manz, B. D., & Keck, J. (2001). Modeling categorical variables by logistic regression. *American Journal of Health Behavior, 25*(3), 278–284. https://doi.org/10.5993/AJHB.25.3.15

Radcliffe, J. (2014). *Leverage a big data maturity model to build your big data roadmap*. Radcliffe Advisory services. https://web.archive.org/web/20170802005853/http://www.radcliffeadvisory.com/research/download.php?file=RAS_BD_MatMod.pdf

Rahman, N. (2016). *Factors affecting big data technology adoption*. Student Research Symposium, Portland State University. https://pdxscholar.library.pdx.edu/studentsymposium/2016/Presentations/10

Saltz, J. S., & Shamshurin, I. (2016, December). Big data team process methodologies: A literature review and the identification of key factors for a project's success. *Proceedings of the IEEE International Conference on Big Data, Washington, DC, USA*, 2872–2879. https://doi.org/10.1109/BigData.2016.7840936

Santoso, L. W. (2023). Big Data: Identification of critical success factors. In P. Vasant, M. Shamsul Arefin, V. Panchenko, J. J. Thomas, E. Munapo, G.-W. Weber, & R. Rodriguez-Aguilar (Eds.), *Intelligent computing and optimization* (pp. 342–349). Springer. https://doi.org/10.1007/978-3-031-50158-6_34

Schmarzo, B. (2013). *Big data business model maturity index*. https://www.dasca.org/world-of-datascience/article/big-data-business-model-maturity-index

Soukaina, M., Anoun, H., Ridouani, M., & Hassouni, L. (2019, October). A study of the factors and methodologies to drive successfully a big data project. *Proceedings of the Third International Conference on Intelligent Computing in Data Sciences, Marrakech, Morocco,* 1-6. https://doi.org/10.1109/ICDS47004.2019.8942257

Sulaiman, H., Cob, Z. C., & Ali, N. (2015, August). Big data maturity model for Malaysian zakat institutions to embark on big data initiatives. Proceedings of the *4th International Conference on Software Engineering and Computer Systems*, *Kuantan, Malaysia,* 61–66. https://doi.org/10.1109/ICSECS.2015.7333084

Sun, S., Cegielski, C. G., Jia, L., & Hall, D. J. (2018). Understanding the factors affecting the organizational adoption of big data. *Journal of Computer Information Systems*, *58*(3), 193-203. https://doi.org/10.1080/08874417.2016.1222891

Surbakti, F. P. S., Wang, W., Indulska, M., & Sadiq, S. (2020). Factors influencing effective use of big data: A research framework. *Information & Management*, *57*(1), 103146. https://doi.org/10.1016/j.im.2019.02.001

van Veenstra, A. F. E., Bakker, T. P., & Esmeijer, J. (2013). Big data in small steps: Assessing the value of data. In the proceedings of *ECP-Jaarcongres 2013*. https://repository.tno.nl/islandora/object/uuid%3A6e0e4f90-13ce-4069-a365-5c787518270a

Veeneman, W., van der Voort, H., Hirschhorn, F., Steenhuisen, B., & Klievink, B. (2018). PETRA: Governance as a key success factor for big data solutions in mobility. *Research in transportation Economics, 69*, 420–429. https://doi.org/10.1016/j.retrec.2018.07.003

Vesset, D., & Xiong, S. (2015). *IDC MaturityScape benchmark: big data and analytics in the United States*. Database trends and applications. https://www.dbta.com/Readers/Subscriber.aspx?Redirect=https://www.dbta.com/DBTA-Downloads/WhitePapers/IDC-MaturityScape-Benchmark-Big-Data-and-Analytics-in-the-United-States-6494.pdf

Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. *Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, USA*, 412-420.

AUTHORS

Soukaina MOUHIB is a doctoral student at Hassan II University, Morocco. She received an Engineering degree in software engineering from Institut National des Postes et Telecommunication (INPT). She worked in the Analytics and Big Data domain for more than ten years with hundreds of customers from Europe, the Middle East, and Africa. She is now Cloud Adoption Manager. Her research interest is Big Data frameworks and methodologies.

Ossama CHERKAOUI is a doctoral student at Hassan II University in Morocco. In 1997, he obtained a Computer Science Engineer degree from Ecole Nationale d'Informatique et d'Analyse des Systèmes (ENSIAS). His research areas are anomaly detection algorithms and the application of machine learning techniques for fraud detection in healthcare insurance.

Pr. Houda ANOUN is a professor in the Department of Computer Science at Ecole Supérieure de Technologie (Hassan II University), where she has been since 2009. She received her degree in Software Engineering from ENSEIRB Bordeaux in 2003 and her Ph.D. in computational linguistics from Bordeaux I University in 2007. Her current research interest lies in the area of artificial intelligence, especially machine learning, deep learning, and Big Data.

Mohammed RIDOUANI is an associate professor and is currently in the Computer Sciences Department at the High School of Technology, Hassan II University, Casablanca, Morocco. He received his engineering degree from the National Institute of Post and Telecommunications (Rabat, Morocco). He has a doctorate degree from Hassan I University, is a trainer (AUF, LPI & CISCO), and is a LINUX expert in networks, systems & security. He serves as a reviewer for prestigious international journals and has TPC member roles at many international conferences. His current research interests are data sciences and smart cities, especially machine learning applications for healthcare, Telecommunication, and IOT security.