

Predicting the outcome of regional development projects using machine learning

Jihad Satri¹, Chakib El Mokhi², Hanaa Hachimi¹

¹Advanced Systems Engineering laboratory, National School of Applied Sciences, Ibn Tofail University, Kénitra, Morocco

²Higher School of Technology, Ibn Tofail University, Kénitra, Morocco

Article Info

Article history:

Received Jul 27, 2023

Revised Oct 24, 2023

Accepted Nov 14, 2023

Keywords:

Artificial intelligence

Data mining

Data science

Machine learning

Prediction

Public decision making

ABSTRACT

Morocco, in its pursuit of inclusive and sustainable territorial development, initiated the advanced regionalization experiment over six years ago. The primary challenge facing government officials today is the management of a burgeoning number of regional development projects. In this article we developed a predictive model based on artificial intelligence and Machine Learning to predict the outcomes of regional development projects, in order to identify the risks associated with their potential failure, and anticipate their impact. To accomplish this, we implemented various data mining techniques and classification algorithms. We collected and analyzed data from past and ongoing regional development projects, considering diverse factors that influence their success or failure. Through rigorous experimentation, we assessed the effectiveness of different predictive models. Our findings reveal that the Random Forest classifier stands out as an efficient algorithm for predicting the outcomes of regional development projects. This research contributes to the broader discourse on the practical implementation of artificial intelligence in public policy and regional development, showcasing its potential to optimize resource allocation, and alleviate the burden of repetitive administrative tasks for organizations operating with limited resources.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Jihad Satri

Advanced Systems Engineering laboratory, National School of Applied Sciences, Ibn Tofail University

Kénitra, Morocco

Email: jihad.satri@uit.ac.ma

1. INTRODUCTION

To facilitate equal access for citizens to conditions that allow them to enjoy a healthy environment is a fundamental right, the regions, as well as the State and public institutions, are working to mobilize all possible means. Advanced regionalization in Morocco is fully in line with the framework of a decentralized democratic state and marks a qualitative leap in the process of democratization of society. Indeed, the choice of advanced regionalization in Morocco is not simply a matter of planning and territorial division. It reflects the Kingdom's determination to take advantage of the momentum of the modernization of States at the global level, as well as the will to amplify its sustainable development through the mobilization of local synergies. Since gaining independence, Morocco has put in place various approaches and strategies to address development challenges. Through an inclusive and coherent advanced regionalization, the country aspires to a harmonious economic development in all the regions of the Kingdom, thus encouraging the various local operators to carry out structuring projects and to reinforce the economic competitiveness of the regions.

Following the successful implementation of advanced regionalization, the need arises for advanced technological tools to empower public and political decision-makers. Artificial intelligence (AI) and machine

learning (ML) have revolutionized decision-making processes by offering rapid, precise, and data-driven insights. By leveraging these cutting-edge technologies, we can act swiftly and with greater accuracy. This is particularly crucial when it comes to regional development projects (DP). Since 2016, there has been a steady increase in the number of DPs. However, it's concerning to observe that nearly 40% of these projects face failure and never progress to the execution phase, even to the failure of execution. This is particularly troubling considering the substantial financial expenses and significant delays associated with the administrative processes related to these projects. These challenges cast a shadow over the progress of regionalization and pose a hindrance to the overall development of the region. By using AI and ML to determine the risks associated with a DP and predict its completion, we can make informed decisions that optimize outcomes and minimize potential setbacks. Ultimately, the use of AI and ML in DP enables us to operate more efficiently, cost-effectively, and with greater success.

In this article, a prediction of the outcomes of DP in the region of beni mellal khenifra (BMK) will be provided following the cross industry standard process for data mining (CRISP-DM) methodology (Figure 1). With the help of advanced prediction techniques, projects can be classified with a high degree of accuracy, their similarities detected, and insights gained from past experiences to enhance the overall project process. In practice, when a new DP is introduced into the system, it leverages previous project data to search for similar cases and make predictions accordingly. By drawing on these insights, we can optimize decision-making, increase project efficiency, and ultimately achieve greater success rates. With the use of predictive models, we can build upon our past experiences to create a stronger, more informed foundation for future projects.

In the BMK region, a staggering number of over 1000 regional DPs were initiated between 2016 and 2023, considering the presence of 164 local governments, each of which had its own set of DPs. This abundance of DPs within a single region presents a substantial challenge in terms of data collection. Additionally, it's important to recognize that Morocco consists of 12 distinct regions, each characterized by its unique distribution of local governments. This diversity further complicates the process of collecting data encompassing all DPs spanning these 12 regions. For our research objectives, we have chosen to focus on analyzing the DPs specific to the BMK Regional Council. To streamline data processing and reduce the complexity associated with managing the names of the 164 local governments within this region, we made the decision to consolidate them into a common category. While this dimensionality reduction approach has allowed for more efficient data management, it's essential to acknowledge that it might impose limitations on the scope and applicability of our analysis.

The purpose of this paper is to explore how ML algorithms can be used to identify patterns within various DPs carried out by the regional council of BMK. The paper is organized as follows: Section 2 gives a literature review of our research. In section 3, the structure of our proposed method is described. Section 4 discusses the results obtained. Finally, the paper is concluded in Section 5.

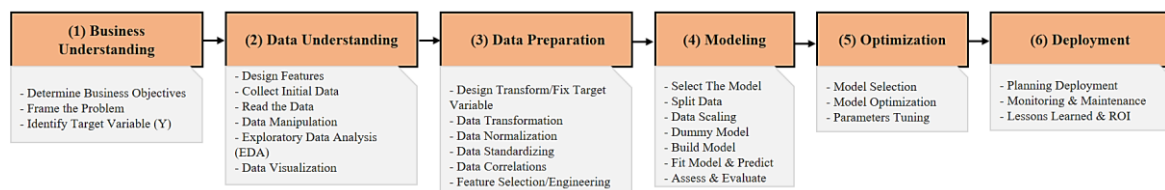


Figure 1. CRISP-DM methodology

2. LITERATURE REVIEW

2.1. About SRAT and PDR

The establishment of the new councils of the regions is an important moment in the consolidation of the process of implementing advanced regionalization, as this is a complex process that takes time to mature in practice. The conduct of planning and programming processes and the formalization of the tools that support it, namely the regional land use plan (SRAT) and the regional development program (PDR), must be built on renewed and innovative forms of territorial planning and programming tools and methods that are capable of giving visible content to the region as a major development player. The SRAT is a reference document that makes it possible to establish a vision of regional development and to define the region's orientations over a 25-year horizon [1]. It also establishes a general framework for sustainable and harmonious regional development in urban and rural areas, as well as proposals for territorial and structuring projects. This document aims to ensure the coordination of interventions of the State, local authorities and

private investors, and to support their strategic choices in terms of development and land use planning, and this through a broad consultation during its implementation, within the framework of the guidelines of the public policy of land use planning adopted at the national level, and in harmony with the strategies, programs and sectoral plans carried out at the regional and national level. The SRAT aims, in particular, to reach an agreement between the State and the Region on spatial planning measures and its upgrading, according to a strategic and prospective vision, so as to allow the definition of the orientations and choices of regional development.

Concerning the PDR which is the reference document that serves for the programming of projects and actions whose implementation is planned in the territory of the region, in order to promote integrated and sustainable development, involving, in particular, the improvement of the attractiveness of the territorial space of the region and the strengthening of its economic competitiveness [1]. Because of its importance, the development and implementation of the PDR must follow a long and precise procedure, whose contours are set by the legal texts on this issue.

2.2. About conventions and projects

The regional council of BMK holds ordinary and extraordinary sessions, during which a series of decisions are made resulting in numerous structuring projects to be implemented. These projects are often carried out through conventions between the council and various stakeholders in the relevant field (Figure 2). The development of such conventions involves several steps that the regional council of BMK must address.

The successful establishment of these conventions is crucial for the success of the DP, making it imperative for the council to execute and closely monitor these agreements. Additionally, the success of these projects serves as the foundation for the implementation of the BMK regional council PDR, which must be continuously monitored, updated, and evaluated to ensure its success. Thus, the regional council of BMK must place great emphasis on the effective execution of conventions and subsequent project implementation to achieve its long-term strategic objectives.

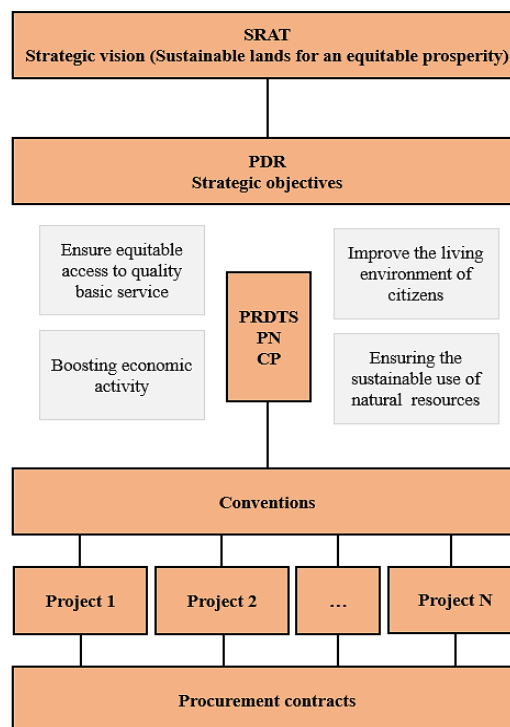


Figure 2. Structuring of the PDR

2.3. AI and ML in the public sector

AI and ML are becoming increasingly important as powerful tools for decision-making in the public sector [2], [3]. For example, they can help decision-makers make informed decisions by analyzing large amounts of data quickly and accurately [4]. By leveraging AI and ML technologies, decision-makers can gain insights into complex problems and develop data-driven strategies to address them [5]. Furthermore,

they improved the efficiency by automating routine tasks, such as data entry and analysis, freeing up employees to focus on higher-level tasks. By reducing manual workloads, public sector organizations can improve efficiency and productivity, and save time and money [6]. AI and ML can help decision-makers predict outcomes and identify trends. For example, predictive analytics can be used to forecast demand for public services, such as healthcare [7], construction [8]–[11], energy [12], and transportation [13], [14], allowing decision-makers to allocate resources more efficiently.

This kind of system could enhance risk management. For example, predictive analytics can be used to identify potential problems before they occur, allowing organizations to take proactive measures to mitigate the risk [15]–[17]. This tools also allow for an increased objectivity and make more objective decisions by removing human bias from the decision-making process. By using algorithms to analyze data [18], decision-makers can reduce the influence of subjective factors and make more informed decisions. Overall, AI and ML have the potential to transform the public sector, improving efficiency, productivity, decision-making, and service delivery.

2.4. Related work

While advanced regionalization is a pivotal component of Morocco's development strategy, there has been insufficient previous research on the integration of AI and ML systems in public decision-making processes to optimize regional development. Previous studies have largely focused on theoretical aspects or broader policy frameworks, neglecting the practical application of new technologies to enhance the advanced regionalization outcomes. Indeed, Ovsyannikova and Domashova [19] explores the application of ML techniques, specifically neural networks, in predicting outcomes of public procurement contracts within the pipe industry. A Python application was created to identify contracts that have a high likelihood of non-performance, utilizing a classification approach for public contracts. On the other hand, Domashova and Kripak [20] discusses the use of ML methods to predict unfavorable outcomes in public procurement procedures related to building and grounds maintenance. The study identifies risks such as collusion among suppliers, conspiracy between customers and suppliers, and inaccurate data in the unified information system. The authors use associative rules mining and clustering methods to identify sustainable groups of suppliers and potential violations. The study also identifies clusters that may be suspected of foul play, such as those with a large number of joint lots and a complete lack of competition. A classification model based on the decision tree was also built to identify the features with the greatest impact on the outcome of contract performance. The results may be of practical importance for authorities in the field of public procurement control. Moreover, Gallego *et al.* [21] discusses the use of ML models to predict inefficiency and corruption in public procurement in Colombia. It explores the challenges and tradeoffs involved in using ML for early warning methods to detect corruption, and highlights the importance of transparency and accountability in anti-corruption efforts. The article also discusses the relationship between information and communication technology (ICT), social capital, and corruption, and suggests that ICT and social capital can be used to combat corruption. Additionally, Rodríguez *et al.* [22] discusses the use of ML algorithms to detect collusion in public procurement auctions. The study compares the performance of eleven algorithms on six datasets from five countries and finds that ensemble methods such as extra trees, random forest (RF), and Ada Boost are the top-performing algorithms. The study also shows that even with limited data, ML algorithms can achieve satisfactory collusion detection rates.

In another way, İnan *et al.* [23] presents a study on the development of a ML model based on long-short term memory (LSTM) to forecast project cost. The proposed model uses a seven-dimensional feature vector, including schedule and cost performance factors and their moving averages as a predictor, and produces more accurate cost estimates when compared to the traditional earned value management (EVM) index-based model. The study aims to demonstrate how managers can benefit more from the data of completed projects by using ML. Similarly, Park *et al.* [24] proposes a two-level stacking heterogeneous ensemble algorithm that combines RF, SVM, and CatBoosting to estimate the cost of building construction projects at an early stage. The proposed method was evaluated using cost information data disclosed by the public procurement service in South Korea. The results showed that the two-level stacking ensemble model performed better than the individual ensemble models. The study suggests that this model can provide objective information for decision-making in the early stages of a construction project.

Apart from this, El Haddadi *et al.* [25] discusses the importance of sustainable public procurement (SPP) in reducing the environmental impact of information and communication technologies (ICTs) and the challenges of implementing it in Morocco. The study analyzes tenders for the purchase of IT equipment in the public administration in Morocco using text mining techniques and finds a low infiltration rate of environmental issues, indicating the need for more SPP efforts. The authors propose a tool for monitoring the promotion of SPP and provide the first green public procurement statistics in Morocco. The article highlights the need to update environmental criteria to adapt to the rapid evolution of the ICT sector and consider the

three pillars of sustainable development simultaneously. Also, Soyulu *et al.* [26] discusses the challenges of ensuring transparency and accountability in public procurement, and presents the authors' experience in Slovenia where they applied anomaly detection techniques to open procurement, company, and spending data sets integrated into a Knowledge Graph through a linked data-based platform called TheyBuyForYou. The article provides guidelines for publishing high-quality procurement data for better procurement analytics and highlights significant shortcomings in the quality of data being published. The authors argue that an open approach combined with data management and advanced analytics plays a critical role in ensuring transparency and accountability in public procurement.

As evident, despite extensive research efforts, there is a noticeable dearth of ML-based prediction models in the existing literature for regional DPs. These studies often fall short of delivering a comprehensive and detailed analysis of the prediction process, frequently limiting their scope to a handful of influential factors and employing rudimentary prediction tools. Consequently, the urgency of forecasting DPs' outcomes has become a critical concern demanding attention within the realm of regional development. Our research, on the other hand, focuses on public procurement contract analysis, a central component of our investigation. Our dataset encompasses a wealth of information, including data from public procurement contracts, in addition to others attributes pertaining to council decisions, conventions, and regional DPs during 7 years. Furthermore, the insights derived from this study can serve as a valuable contribution to more extensive research aimed at creating a decision support tool for assessing the risks linked to regional DPs across all Moroccan regional councils.

3. METHODS

It is widely acknowledged that a significant portion (around 80%) of the work in predictive data analytics projects is completed during the business understanding, data understanding, and data preparation phases of the CRISP-DM. The CRISP-DM is a well-established data mining process model that provides a structured approach to planning, developing, deploying, and maintaining data mining solutions (Figure 1). Originally published to standardize data mining processes across industries, the CRISP-DM has since become the most commonly used methodology for data mining, analytics, and data science projects [27]. Therefore, it is essential to prioritize these early phases of the process to ensure the success of any data-driven project. The implication is that the results are more likely to be aligned with the business goals and of higher quality, which justifies the initial choice of this approach.

3.1. Business and data understanding

Defining the scope and objectives of a data science project is critical to its success. In fact, the process of understanding the problem involves several crucial steps. Firstly, the project's objectives must be clearly defined by identifying what the business wants to achieve and why. Understanding the business problem is a key step in any data analytics project as it helps determine the kind of insights that a predictive analytics model can provide to help address the problem. This defines the analytics solution that the data scientist will set out to build using ML techniques. Defining the analytics solution is the most significant task in the business understanding phase of the CRISP-DM process. Therefore, it is essential to allocate sufficient time and resources to accurately define the scope and objectives of the project to ensure its success.

Once the objectives of a data science project have been defined, the next step is to identify the appropriate analytics approach, such as classification, regression, clustering, or anomaly detection. It is also essential to frame the problem by determining the best solution and reformulating the problem statement as an analytics problem. This involves building a bank of questions and understanding the stakeholders, as well as their expectations and requirements. By doing so, we can ensure that the analytics solution is aligned with the business objectives and meets the needs of all stakeholders.

As discussed in section 2 (about conventions and projects), the Regional Council of BMK engages in conventions with its partners to establish DPs. Following a comprehensive analysis of the DPs and discussions with stakeholders involved in the development and implementation process, two operational processes were identified within the organization. The first process involves the elaboration of conventions, while the second process pertains to the implementation of DPs via procurement contracts. The success of a DP hinges on the outcomes of conventions. Therefore, it is necessary to merge convention and DP data to conduct predictive analyses and accurately forecast the outcomes of the DP. This requires a classification model with a categorical target variable indicating either "FAIL" or "SUCCESS". The models are based on patterns extracted from historical datasets and provide insights to help the organization make better decisions to solve business problems, rather than solving the problems themselves.

Overall, to framing the problem statement (Figure 3), the model could be built to predict the likelihood of a project failing. This model could be used to assign each new DP a probability of failure, and those most likely to fail could be flagged for review by managers and decision-makers. In this way, the

limited time spent reviewing DPs could be targeted to those most likely to fail, thereby increasing the number of failed projects detected and reducing processing time. Once we have a good understanding of the problem, the next step is dedicated to the data understanding (Examine the data, identify problems in the data). In this phase, the relevant data is collected and explored to understand its quality, completeness, and structure.

In this article, a dataset (conventions and procurement contracts) extracted from a SQL server database is used to perform the DP (fail or success) analysis using ML classification algorithms (Table 1). The classification algorithms are ML algorithms that are used to classify or categorize data into predefined classes or categories [28]. These algorithms learn from labeled training data and use that knowledge to predict the class or category of new, unseen data. There are several classification algorithms, the most popular are: decision tree (DT), naive bayes (NB), support vector machines (SVM), k-nearest neighbors (KNN). Each classification algorithm has its own strengths and weaknesses, and the choice of algorithm depends on the specific problem being solved and the characteristics of the data.



Figure 3. Framework for formulation

Table 1. Description of DP dataset

N°	Column	Type	Description
1	province	categorical	The province where the DP is carried out (Distinct 5)
2	program	categorical	The program to which the DP belongs (Distinct 4)
3	amount_project	numeric	The total amount of the DP to be implemented
4	delay_days	numeric	The DP duration in days
5	physical_advancement	numeric	The physical progress of the project completion in percent
6	financial_advancement	numeric	The financial progress of the project completion in percent
7	approval	boolean	Project approval by decision-makers
8	visa	boolean	Visa of the convention by the partners
9	os_start	boolean	Service order for the launch of the DP
10	order_stop	boolean	Service order to stop the execution of the DP
11	order_resume	boolean	Service order for the resumption of the execution of the project
12	provisional_receipt	boolean	The provisional reception of the DP
13	final_receipt	boolean	The final reception of the DP
14	cumulative_payments	numeric	Cumulative payment of the DP
15	completion	numeric	Completion of the project (no = 0, yes =1), the target variable

Now, a data collection was done while assigning a label (target variable) according to their past outcomes. The dataset related to conventions and procurement contracts will be used in order to look to this data from different corners, understanding the data, identifying the type, the shape, the structure, the quality of the data, understanding the statistical analysis to measure the quality of data from different perspectives. To this end, an exploratory data analysis (EDA) will be applied. The EDA is an approach to analyzing and summarizing data sets with the goal of discovering and understanding patterns, relationships, and trends in the data [29]. The purpose of EDA is to gain insights and develop an understanding of the data, which can be used to guide further analysis or to make decisions. EDA involves a variety of techniques, such as visualizations, summary statistics, and data transformations. Visualizations can be used to identify patterns and relationships in the data, while summary statistics can be used to provide a quick overview of the data's central tendencies and distributions. Data transformations, such as scaling or normalizing the data, can be used to make it easier to interpret or to prepare it for further analysis. EDA is typically performed at the

beginning of the data analysis process, before any formal modeling or hypothesis testing is conducted. It helps to identify potential outliers, missing values, or other issues with the data that may need to be addressed before further analysis can be done. Additionally, EDA can help to generate hypotheses or guide the development of more complex models.

In order to gain a deeper understanding of how DPs are distributed across various provinces, we can analyze the Figure 4. Based on our observations, it becomes clear that Azilal province has a higher number of DPs compared to the other provinces. This finding aligns with the logic of the PDR which is a positive indication. Azilal is a mountainous region with limited infrastructure and development. Therefore, it is crucial to pay special attention to its development and upgrade it compared to other provinces in order to reduce territorial disparities. Additionally, it is worth noting that the majority of the DPs are derived from conventions rather than other programs (Figure 4). Interestingly, there is a tendency for DPs resulting from these conventions to experience a higher rate of failure (Figure 5), this suggests that partnerships and collaborations play a significant role in the implementation of DPs across the different provinces. It also highlights the importance of fostering strong relationships and networks between various organizations and stakeholders in order to facilitate the execution of successful DPs. On the other hand, the distribution of the total cost of projects by province follows a normal distribution overall (Figure 6). Notably, the MAN program (upgrading of emerging centers) incurs a significant cost, highlighting its priority status among political decision-makers.

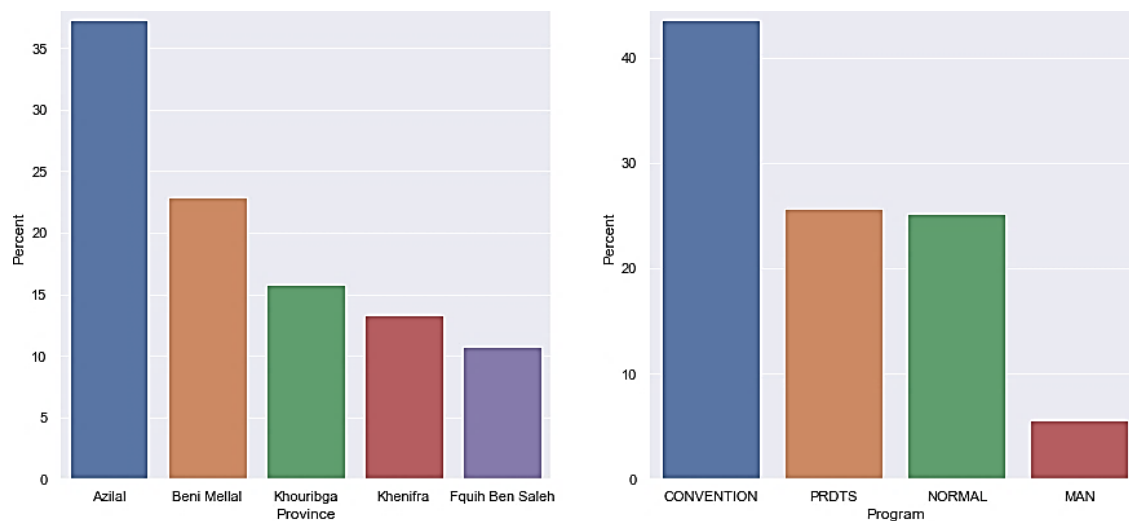


Figure 4. DPs by province and program

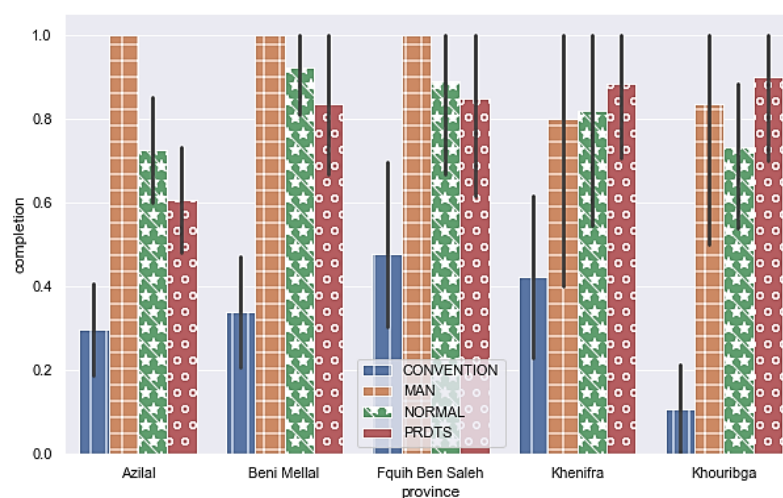


Figure 5. Completion of projects by province and program

3.2. Data preparation

Data preparation is a dynamic, iterative process, wherein the precise methods and procedures employed are contingent upon the characteristics of the data and the specific ML objective. This critical phase is instrumental in guaranteeing that the ML model is fed with data that is not only clean but also pertinent and correctly structured. Such meticulous data preparation is pivotal in enhancing the precision and dependability of the model's predictions or insights, making it an essential component of the ML pipeline.

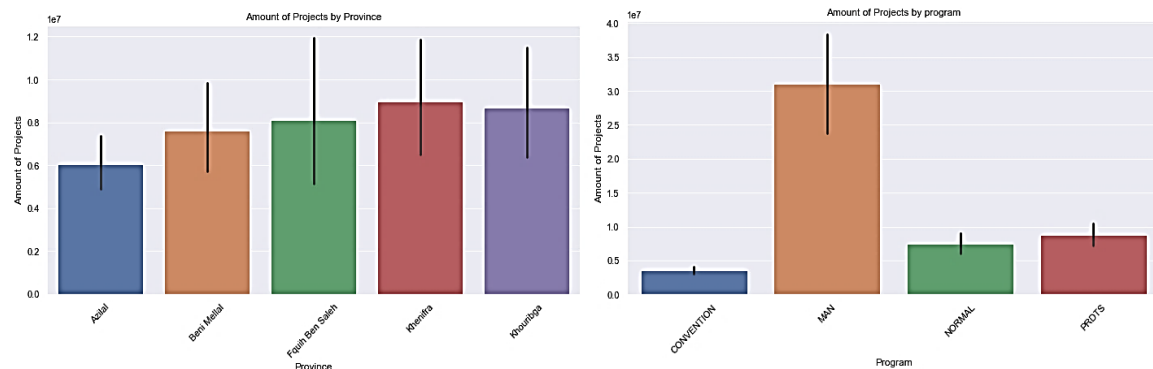


Figure 6. Number of projects by province and program

3.2.1. Target variable

After the data collection was conducted, during which labels were assigned based on their past outcomes, this process entails categorizing data points into distinct classes or labels based on their historical results. This approach is commonly employed in predictive modeling and ML tasks. By observing Figure 7 the target variable is evenly distributed among the classes, it indicates a balanced dataset. This is generally desirable as it allows the model to learn from an equal representation of each class. However, it's important to note that the ideal balance may vary depending on the specific problem and domain. At this stage our dummy model is: 57,97% (Success) - 42,02% (Fail) which means a naive model that serves as a baseline for performance comparison with more complex models. It provides a basic reference point against which the effectiveness of other models can be measured. The dummy model typically involves making predictions based on simple rules or random processes, without considering any relationships or patterns in the data. By comparing metrics such as accuracy, precision, recall, F1 score, or area under the receiver operating characteristic (ROC) curve (AUC-ROC), one can determine if the model performs significantly better than the dummy model. If the model fails to outperform the dummy model, it indicates that more sophisticated approaches or feature engineering are necessary to capture meaningful patterns in the data.

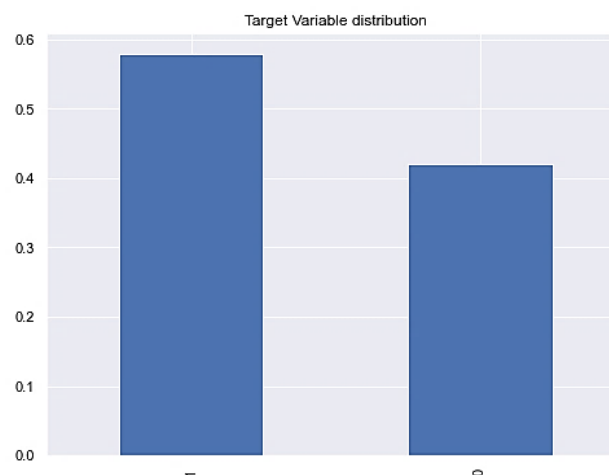


Figure 7. Target variable distribution

3.2.2. Data transformation

Data transformation is a vital process that modifies or converts the original data into a different representation, making it more suitable for analysis or modeling purposes. This involves employing various techniques to enhance data quality, extract valuable insights, and fulfill specific requirements of the analysis or modeling task. Data transformation encompasses a diverse range of techniques, such as handling missing values, outliers, and skewness, as well as performing scaling and normalization. The selection of techniques relies on the data's characteristics, the specific objectives of the analysis or modeling, and the algorithmic requirements. The ultimate aim of data transformation is to enhance data quality, interpretability, and performance for subsequent analysis or modeling.

When dealing with categorical variables in a dataset, a technique called "get_dummies" is often employed as part of the data transformation process. This technique specifically converts categorical variables into a numerical format by generating binary indicator variables for each category. It effectively incorporates categorical information into ML models, enabling them to make meaningful use of this type of data. By utilizing the "get_dummies" function, the categorical variables are transformed into a numerical representation that can be readily understood and utilized by ML algorithms. This enhances the effectiveness and accuracy of the models in incorporating categorical information during the analysis or modeling process.

3.2.3. Feature selection

Feature selection is the process of choosing a subset of relevant features from a larger set of available features in a dataset. In this task, a feature selection process is employed, leveraging stable prior knowledge. The selection process primarily involves the removal of IDs, dates, data with high cardinality, and constant values. After that a correlation analysis is utilized as a criterion for feature selection in ML. The first step is to compute the correlation between each feature (independent variable) and the target variable (dependent variable) in the dataset (Figure 8). This is typically done using a correlation coefficient, such as Pearson's correlation coefficient. The correlation coefficient measures the strength and direction of the linear relationship between two variables. Based on the correlation coefficients, we identify features that have a strong correlation with the target variable. Features with high positive or negative correlation values indicate a potentially significant relationship with the target. These features are more likely to contain useful predictive information. It's also important to examine the correlation between the independent variables themselves. Highly correlated features may provide redundant or overlapping information. In such cases, it is advisable to select only one representative feature from each highly correlated group. This helps in reducing multicollinearity and can improve model interpretability and stability. After assessing the correlation between features and the target variable, and considering the correlations among the features themselves, we proceed with selecting the final subset of features to be included in the ML model. The selected features should have a high correlation with the target variable while minimizing redundancy and multicollinearity.

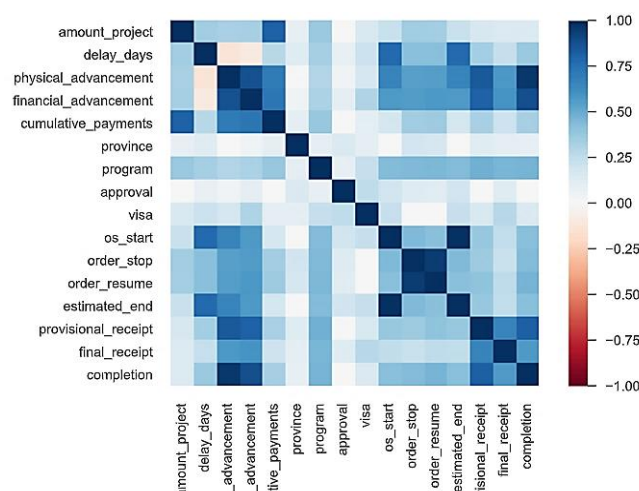


Figure 8. Correlation heatmap

3.3. Modeling

The Modeling step focuses on developing and training the predictive or descriptive models based on the prepared dataset. This dataset is split into training and testing sets. The next step is to choose suitable

models for training. Given that the problem at hand involves classification, appropriate classifiers need to be selected for conducting the classification task. The classification algorithms utilized in this analysis include DT, NB, logistics regression (LR), SVM, Kernel SVM, KNN, and RF.

Scaling the data prior to training our model is an essential step in the data preprocessing phase. It ensures that all features are normalized to a similar scale or range, addressing the sensitivity of many ML algorithms to feature scales. If features have varying scales, those with larger magnitudes can overshadow the learning process and disproportionately impact the model. By scaling the data, we establish parity among features, safeguarding against any single feature overpowering the learning algorithm.

Next, models are trained using the prepared dataset, with the target variable and relevant features appropriately defined. The training data is used to optimize the model's parameters and learn the underlying patterns and relationships in the data. Once the models are built, their performance needs to be evaluated. This is done by using evaluation metrics such as accuracy, precision, recall, F1 score, depending on the type of problem (classification and regression.). The models are typically assessed using evaluation techniques like cross-validation or holdout validation on a separate validation dataset.

3.4. Optimization

The model optimization focuses on fine-tuning models and algorithms to achieve the defined optimization goals. It involves adjusting model parameters, exploring alternative algorithms, employing ensemble techniques, and iteratively optimizing until the desired performance is attained. By conducting rigorous optimization, the models can deliver improved predictive accuracy, enhanced interpretability, and better alignment with business objectives.

3.5. Deployment

The Deployment step ensures that the developed models are effectively integrated into operational systems or decision-making processes. It involves planning the deployment strategy, preparing the infrastructure, implementing the models, testing and validating the deployment, monitoring and maintaining the models, documenting the processes, providing user training and support, and fostering continuous improvement. Successful deployment allows organizations to benefit from the predictive or descriptive capabilities of the models, enabling informed decision-making and driving business value.

4. RESULTS AND DISCUSSION

Table 2 presents the evaluation results depicting the performance metrics of the seven classification algorithms employed in this research, including measures such as accuracy, precision, recall, F1 score, and AUC. In this study, the predictive outcomes of the DP were determined using seven algorithms. The performance metrics of these algorithms were carefully evaluated to identify the most effective predictive model.

Accuracy represents the overall correctness of the model's predictions. RF has the highest accuracy (98.88%), while NB has the lowest (56.18%). Precision measures the accuracy of positive predictions. RF, KSVM, and DT have high precision values, indicating a high percentage of correctly identified positive cases. Recall focuses on the model's ability to identify positive instances. NB has the highest recall (100%), indicating that it can correctly identify all positive cases. SVM, RF, and DT also have high recall values. F1 Score is the harmonic mean of precision and recall, providing a balanced measure. RF has the highest F1 Score (99.01%), indicating an overall better balance between precision and recall. NB has the lowest F1 Score (71.94%), suggesting a trade-off between precision and recall. To further validate the performance of the seven algorithms and determine the optimal classifier for the dataset, the AUC measure was utilized (Figure 9). Considering the AUC metric, RF demonstrates the best overall performance with an AUC score of 0.99, indicating its effectiveness in predicting the outcomes of the DP.

Table 2. Performance of classification algorithms

%	Accuracy	Precision	Recall	F1 Score	AUC	Average
DT	96.63	96.08	98.00	97.03	0.96	77.74
NB	56.18	56.18	100	71.94	0.50	56.96
LR	96.63	96.08	98.00	97.03	0.96	77.74
SVM	96.63	94.34	100	97.09	0.96	77.80
KSVM	94.38	94.12	96.00	95.05	0.94	76.10
KNN	92.13	93.88	92.00	92.93	0.92	74.37
RF	98.88	98.04	100	99.01	0.99	79.30

In summary, the RF algorithm demonstrates strong performance across various metrics, including accuracy, precision, recall, F1 Score, and the AUC, achieving an impressive average score of 79.30. On the other hand, NB exhibits high recall but lower accuracy and F1 score. Both SVM and KSVM display high recall and reasonable accuracy. DT and LR classifiers demonstrate consistent performance across all metrics. However, KNN falls behind other classifiers in terms of performance. Considering these results, the RF classifier emerges as the preferred choice for the best classifier in this scenario.

The findings of this study hold significant implications for the regional council of BMK, providing valuable insights into anticipating the impact of project spin-offs, setting priorities, and minimizing administrative burdens. In the event of a predicted failure of a DP, decision-makers can proactively strategize at the early stage by drawing insights from previous projects. For instance, they can assess whether the project aligns with the partner's competencies and ensure its inclusion within the PDR. This includes scrutinizing the project's technical study, determining the land's nature and area designated for the project, and addressing its legal status. Moreover, it is paramount to verify that the project's partners possess the requisite financial resources to successfully execute it. These results offer an opportunity to enhance decision-making processes, optimize project efficiency, and ultimately improve success rates. By leveraging these findings, the regional council of BMK can make informed strategic decisions, streamline operations, and maximize overall project outcomes.

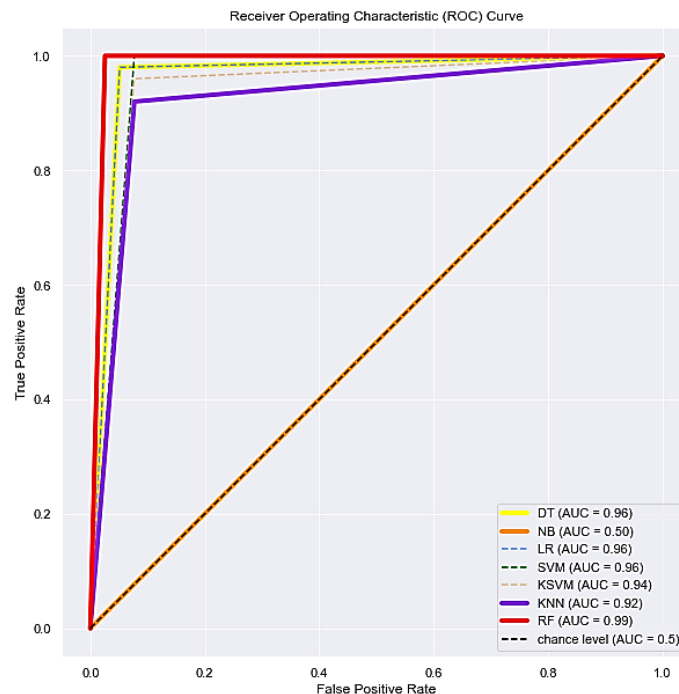


Figure 9. ROC curve

5. CONCLUSION

Selecting the most suitable algorithm for a specific data mining task can be a challenging. A recommended approach is to conduct a thorough evaluation of various algorithms to determine the most effective one that yields accurate outcomes. In this study, a comparative analysis was conducted on seven classification algorithms: DT, NB, LR, SVM, KSVM, KNN, and RF. The analysis employed a dataset (conventions and procurement contracts) extracted from a SQL server database, specifically focusing on the analysis of DP (fail or success) using ML classification algorithms. To obtain the optimal predictive model, an evaluation of various classification algorithms was conducted, considering key performance metrics such as accuracy, precision, recall, F1 score, and AUC. The purpose of this evaluation was to identify the algorithm that consistently delivers superior results across these performance measures. By assessing the algorithms based on these metrics, a comprehensive understanding of their effectiveness in predicting outcomes can be obtained, aiding in the selection of the most reliable and efficient predictive model. Based on the comprehensive analysis of performance metrics in this study, the RF classifier emerged as the most effective algorithm for predicting the outcomes of the DP. Outperforming all seven algorithms across all

evaluated metrics, RF showcased efficiency and accuracy in its predictions. These findings strongly indicate that the RF classifier is a highly reliable and robust choice for accurately forecasting DP outcomes.

This research adds a practical dimension to the ongoing discourse on integrating AI into public policy and regional development. It showcases how AI can be effectively applied to real-world challenges, offering concrete solutions rather than just theoretical discussions. In summary, the paper provides a practical illustration of how AI can significantly enhance public policy and regional development, demonstrating the tangible advantages of AI, including improved resource allocation and reduced administrative workload. These benefits are particularly valuable for organizations with limited resources.

As a suggestion for future research, there is room for enhancing data collection and recording processes, particularly concerning data sourced from various local governments, given that precise and all-encompassing data are paramount for accurate DPs predictions. Furthermore, the exploration of alternative algorithm models, such as AdaBoost, gradient boosting, LightGBM, and XGBoost, among other advanced algorithms, should be considered for further investigation.




REFERENCES

- [1] J. Satri, C. El Mokhi, and H. Hachimi, "Artificial intelligence and machine learning for a better decision making in the public sector," *8th International Conference on Optimization and Applications, ICOA 2022 - Proceedings*, 2022, doi: 10.1109/ICOA55659.2022.9934325.
- [2] S. J. Mikhaylov, M. Esteve, and A. Campion, "Artificial intelligence for the public sector: Opportunities and challenges of cross-sector collaboration," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128, 2018, doi: 10.1098/rsta.2017.0357.
- [3] A. Zuiderwijk, Y. C. Chen, and F. Salem, "Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda," *Government Information Quarterly*, vol. 38, no. 3, 2021, doi: 10.1016/j.giq.2021.101577.
- [4] M. Pourhomayoun and M. Shakibi, "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making," *Smart Health*, vol. 20, 2021, doi: 10.1016/j.smhl.2020.100178.
- [5] S. Rawat, A. Rawat, D. Kumar, and A. S. Sabitha, "Application of machine learning and data visualization techniques for decision support in the insurance sector," *International Journal of Information Management Data Insights*, vol. 1, no. 2, 2021, doi: 10.1016/j.jiime.2021.100012.
- [6] C. Wilson and M. van der Velden, "Sustainable AI: An integrated model to guide public sector decision-making," *Technology in Society*, vol. 68, 2022, doi: 10.1016/j.techsoc.2022.101926.
- [7] B. Van Calster, L. Wynants, D. Timmerman, E. W. Steyerberg, and G. S. Collins, "Predictive analytics in health care: how can we know it works?," *Journal of the American Medical Informatics Association*, vol. 26, no. 12, pp. 1651–1654, 2019, doi: 10.1093/jamia/ocz130.
- [8] O. Alshboul, G. Almasabha, A. Shehadeh, R. E. Al Mamlook, A. S. Almuflih, and N. Almakyeel, "Machine Learning-Based Model for Predicting the Shear Strength of Slender Reinforced Concrete Beams without Stirrups," *Buildings*, vol. 12, no. 8, 2022, doi: 10.3390/buildings12081166.
- [9] G. Almasabha, K. F. Al-Shboul, A. Shehadeh, and O. Alshboul, "Machine learning-based models for predicting the shear strength of synthetic fiber reinforced concrete beams without stirrups," *Structures*, vol. 52, pp. 299–311, 2023, doi: 10.1016/j.istruc.2023.03.170.
- [10] O. Alshboul, G. Almasabha, K. F. Al-Shboul, and A. Shehadeh, "A comparative study of shear strength prediction models for SFRC deep beams without stirrups using Machine learning algorithms," *Structures*, vol. 55, pp. 97–111, 2023, doi: 10.1016/j.istruc.2023.06.026.
- [11] G. Almasabha, A. Shehadeh, O. Alshboul, and O. Al Hattamleh, "Structural performance of buried reinforced concrete pipelines under deep embankment soil," *Construction Innovation*, 2023, doi: 10.1108/CI-10-2021-0196.
- [12] M. Zekić-Sušac, S. Mitrović, and A. Has, "Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities," *International Journal of Information Management*, vol. 58, 2021, doi: 10.1016/j.ijinfomgt.2020.102074.
- [13] A. R. A. Audu, A. Cuzzocrea, C. K. Leung, K. A. MacLeod, N. I. Ohin, and N. C. Pulgar-Vidal, "An Intelligent Predictive Analytics System for Transportation Analytics on Open Data Towards the Development of a Smart City," *Advances in Intelligent Systems and Computing*, vol. 993, pp. 224–236, 2020, doi: 10.1007/978-3-030-22354-0_21.
- [14] A. Mahpour and T. El-Diraby, "Application of Machine-Learning in Network-Level Road Maintenance Policy-Making: The Case of Iran," *Expert Systems with Applications*, vol. 191, 2022, doi: 10.1016/j.eswa.2021.116283.
- [15] B. W. Wirtz, J. C. Weyerer, and B. J. Sturm, "The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration," *International Journal of Public Administration*, vol. 43, no. 9, pp. 818–829, 2020, doi: 10.1080/01900692.2020.1749851.
- [16] J. Satri, C. El Mokhi, and H. Hachimi, "Road Accident Forecast Using Machine Learning," 2023, pp. 701–708. doi: 10.1007/978-3-031-26254-8_102.
- [17] A. Shehadeh, O. Alshboul, R. E. Al Mamlook, and O. Hamedat, "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression," *Automation in Construction*, vol. 129, 2021, doi: 10.1016/j.autcon.2021.103827.
- [18] D. Kolkman, "The usefulness of algorithmic models in policy making," *Government Information Quarterly*, vol. 37, no. 3, 2020, doi: 10.1016/j.giq.2020.101488.
- [19] A. Ovsyannikova and J. Domashova, "Identification of public procurement contracts with a high risk of non-performance based on neural networks," *Procedia Computer Science*, vol. 169, pp. 795–799, 2020, doi: 10.1016/j.procs.2020.02.161.
- [20] J. Domashova and E. Kripak, "Application of machine learning methods for risk analysis of unfavorable outcome of government procurement procedure in building and grounds maintenance domain," *Procedia Computer Science*, vol. 190, pp. 171–177, 2021, doi: 10.1016/j.procs.2021.06.022.




- [21] J. Gallego, G. Rivero, and J. Martínez, "Preventing rather than punishing: An early warning model of malfeasance in public procurement," *International Journal of Forecasting*, vol. 37, no. 1, pp. 360–377, 2021, doi: 10.1016/j.ijforecast.2020.06.006.
- [22] M. J. García Rodríguez, V. Rodríguez-Montequín, P. Ballesteros-Pérez, P. E. D. Love, and R. Signor, "Collusion detection in public procurement auctions with machine learning algorithms," *Automation in Construction*, vol. 133, 2022, doi: 10.1016/j.autcon.2021.104047.
- [23] T. Inan, T. Narbaev, and Ö. Hazir, "A Machine Learning Study to Enhance Project Cost Forecasting," *IFAC-PapersOnLine*, vol. 55, no. 10, pp. 3286–3291, 2022, doi: 10.1016/j.ifacol.2022.10.127.
- [24] U. Park, Y. Kang, H. Lee, and S. Yun, "A Stacking Heterogeneous Ensemble Learning Method for the Prediction of Building Construction Project Costs," *Applied Sciences (Switzerland)*, vol. 12, no. 19, 2022, doi: 10.3390/app12199729.
- [25] T. El Haddadi, O. El Haddadi, T. Mourabit, A. El Allaoui, and M. Ben Ahmed, "Automatic analysis of the sustainability of public procurement based on Text Mining: The case of the Moroccan ICT markets," *Cleaner and Responsible Consumption*, vol. 3, 2021, doi: 10.1016/j.clrc.2021.100037.
- [26] A. Soylyu *et al.*, "Data Quality Barriers for Transparency in Public Procurement," *Information (Switzerland)*, vol. 13, no. 2, 2022, doi: 10.3390/info13020099.
- [27] F. Martínez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3048–3061, 2021, doi: 10.1109/TKDE.2019.2962680.
- [28] J. A. Talingdan, "Performance comparison of different classification algorithms for household poverty classification," *Proceedings - 2019 4th International Conference on Information Systems Engineering, ICISE 2019*, pp. 11–15, 2019, doi: 10.1109/ICISE.2019.00010.
- [29] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, "Exploratory data analysis using python," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 4727–4735, 2019, doi: 10.35940/ijitee.L3591.1081219.

BIOGRAPHIES OF AUTHORS






Jihad Satri    is currently a research student at Ibn Tofail University of Kenitra - Morocco with Master degree in database and system integration from the university of Nice sophia antipolis in France. His research is in fields of computer science, artificial intelligence, and digital nudging. He has published papers in international conferences and journals. He can be contacted at email: jihad.satri@uit.ac.ma.



Chakib El Mokhi    is an assistant Professor for electrical engineering, renewable energy and computer science at Ibn Tofail University of Kenitra - Morocco. He studied electrical and information engineering in Germany and then worked there for several years in the automotive industry. His research work focuses on the optimization of energy production and the energetic efficiency using metaheuristic optimization algorithms. His research interests as well in the fields of Big Data and decision-oriented informatics. He is co-editor of the « International Journal on Optimization and Applications » (IJOA). He is member of the organizing committees of the « International Conference on Optimization and Applications » and the International Competition of Innovation « Let's Challenge ». He reviewed several academic research articles for the MDPI publisher and many books for CRC Press. He can be contacted at email: chakib.elmokhi@uit.ac.ma.



Hanaa Hachimi    Ph.D. in Applied Mathematics & Computer Science and a Ph.D. in Mechanics & Systems Reliability, former Secretary General of Sultan Moulay Slimane University in Beni Mellal. President of the Moroccan Society of Engineering Sciences and Technology (MSEST). She is Associate Professor at Ibn Tofail University, National School of Applied Sciences of Kenitra, Morocco. She is the Editor in Chief of "The International Journal on Optimization and Applications" (IJOA). She is Director of the Systems Engineering Laboratory (LGS) and IEEE Senior Member, precisely she is affiliated at the Big Data, Optimization, Service and Security (BOSS) team at USMS. She is Lecture and Keynote Speaker of the courses: Optimization & Operational Research, Graph Theory, Statistics, Probability, Reliability and Scientific Computing. She is Member of the Moroccan Society of Applied Mathematics (SM2A). She is the General Chair of "The International Conference on Optimization and Applications" (ICOA) & the International Competition of Innovation (Let's Challenge). Lions Club Member and UNESCO UIT-Chair Member. She can be contacted at email: hanaa.hachimi@uit.ac.ma.