

Article

Machine Learning Analysis of Scientific Projects for Decision Support: Empirical Evidence from CORDIS FP6

Arailym Medetbek ^{1,*}, Aidos Mukhatayev ^{2,*} , Svitlana Biloshchytska ^{3,4}, Andrii Biloshchytskyi ³ and Dias Berlibek ⁵ 

¹ School of Software Engineering, Astana IT University, Astana 010000, Kazakhstan

² School of General Educational Disciplines, Astana IT University, Astana 010000, Kazakhstan

³ Department of Computational and Data Science, Astana IT University, Astana 010000, Kazakhstan; bsv@astanait.edu.kz (S.B.); a.b@astanait.edu.kz (A.B.)

⁴ Department of Information Technologies, Kyiv National University of Construction and Architecture, 03680 Kyiv, Ukraine

⁵ Information Security System, L.N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan; 021201550905@enu.kz

* Correspondence: arailym.medetbek@astanait.edu.kz (A.M.); aidos.mukhatayev@astanait.edu.kz (A.M.); Tel.: +7-707-546-9769 (A.M.); +7-702-631-2244 (A.M.)

Abstract

The increasing complexity of scientific research projects and growing competition for public funding require data-driven approaches to support evidence-based decision-making in research management. This paper presents an integrated machine learning-based analytical pipeline for scientific project analysis, positioning the contribution as methodological integration and empirical insight rather than the development of new algorithms. Using open data from the CORDIS database on projects funded under the Sixth Framework Programme (FP6), the study combines three complementary analytical tasks—structure identification, budget prediction, and funding-completeness classification—within a single empirical decision-support workflow. Unsupervised analysis based on dimensionality reduction and clustering reveals distinct structural patterns in project characteristics and identifies atypical large-scale projects characterized by substantially higher budgets and consortium sizes. Regression models predict total project costs, achieving $R^2 \approx 0.50$ on unseen data; the gap between explained and unexplained variance is interpreted in relation to latent contextual factors not available in administrative records. Classification models distinguish between fully and partially funded projects (threshold $r \geq 0.70$); the high discriminative performance ($AUC > 0.97$) is interpreted cautiously in light of target-variable construction and potential feature dependence. Association rule mining identifies interpretable funding patterns (strongest rule lift = 6.30). The study contributes a reproducible example of how mature machine learning techniques can be systematically integrated to generate domain-specific evidence for research funding governance.



Academic Editor: Paolo Renna

Received: 3 April 2026

Revised: 19 June 2026

Accepted: 24 June 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: machine learning; scientific projects; decision support; research funding; project analytics; data-driven management

1. Introduction

The increasing scale and complexity of scientific research projects, combined with growing competition for public funding, have significantly intensified the need for effective decision support in research governance. Modern research programmes typically

involve international consortia, diverse organizational structures, and substantial financial resources, making the evaluation and management of project proposals a complex and resource-intensive process [1–3]. Traditional assessment mechanisms, primarily based on expert review and qualitative judgment, remain central to funding decisions but face limitations in terms of scalability, consistency, and transparency when applied to large and heterogeneous project portfolios.

In recent years, the availability of large administrative and open datasets has created new opportunities for the application of data-driven methods in research management [4]. Machine learning techniques, in particular, have demonstrated considerable potential for extracting patterns from complex datasets and supporting evidence-based decision-making across various public-sector domains. Within the context of research funding, such approaches can complement expert evaluation by providing objective, reproducible insights derived from historical project data.

Despite these developments, existing analytical approaches to research funding remain largely fragmented. Prior studies have typically focused on isolated tasks, such as research performance evaluation, collaboration network analysis, or cost estimation, without offering an integrated project-level perspective on how project structure, financial configuration, and funding outcomes interact [5]. The present study therefore does not claim algorithmic novelty; instead, it addresses an application gap by combining mature analytical techniques into a systematic empirical workflow for research management. Project-level analyses are also constrained by data limitations, as information on rejected proposals is rarely available, complicating the formulation and interpretation of supervised learning tasks.

This study addresses these challenges by developing an integrated analytical pipeline that unifies three complementary tasks within a decision-support context: (1) identification of latent project structures via unsupervised learning; (2) prediction of project budgets via regression; and (3) classification of funding completeness via supervised learning. The contribution of the work lies in the systematic integration and interpretation of these tasks for scientific research management, rather than in introducing new algorithmic procedures. The three tasks are connected through a shared feature representation, and cluster membership from Stage 1 serves as an enriched input to Stage 3, forming a sequential pipeline that approximates a probabilistic model of funding outcomes under uncertainty.

The main objective of this paper is to assess whether structural and financial parameters of scientific projects contain sufficient information to support reliable, data-driven decision-making in research funding contexts. The central hypothesis is that project characteristics related to consortium composition, international collaboration, and budget configuration exhibit stable patterns that can be identified through unsupervised learning and can be used to predict budget scale and funding completeness with high accuracy.

Accordingly, the paper is positioned as an application-oriented study of methodological integration. To the best of the authors' knowledge, few project-level studies have systematically combined structure identification, budget prediction, funding-completeness classification, and association rule mining for scientific research management using participant-level and project-level CORDIS data.

To investigate this hypothesis, the study integrates three complementary analytical approaches. First, unsupervised learning methods are applied to identify latent project structures and typologies. Second, regression models are used to analyze the relationship between organizational features and total project budgets. Third, supervised classification and associative rule mining are employed to assess funding outcomes and extract interpretable patterns linked to successful and partially successful projects.

The results demonstrate that scientific projects form a limited number of stable structural configurations and that financial and organizational parameters play an important role

in both budget determination and funding completeness. By integrating standard machine learning paradigms into a single empirical workflow, this study provides application-oriented evidence on the role of project structure in shaping funding outcomes and offers practical insights for the design of intelligent decision-support tools in research governance.

2. Related Work

The application of data-driven methods in research evaluation and science policy has expanded significantly over the past decade, driven by the growing availability of large-scale administrative and bibliometric datasets and the increasing complexity of research systems. Traditional approaches to assessing scientific activity have historically relied on expert judgment, peer review procedures, and descriptive statistical indicators. While these mechanisms remain central to research governance and funding allocation, their effectiveness is increasingly challenged in environments characterized by large project portfolios, international collaboration, and multidimensional performance criteria [6,7]. In particular, qualitative evaluation processes often struggle to scale efficiently and to provide consistent, reproducible insights when applied to heterogeneous sets of research projects with diverse organizational and financial configurations.

In response to these challenges, computational and analytical methods have gained prominence as complementary tools for supporting research management and funding decisions. Within the broader field of scientometrics and research analytics, statistical modeling and machine learning techniques have been widely adopted to analyze publication output, citation patterns, and collaboration dynamics [8–11]. These approaches have enabled the identification of structural regularities in scientific production and have contributed to evidence-based policy discussions regarding research performance, excellence, and impact. Network-based analyses, in particular, have emphasized the role of collaboration structures, co-authorship networks, and international cooperation in shaping scientific productivity and knowledge diffusion [12]. Such studies consistently demonstrate that structural characteristics of research systems influence outcomes, suggesting that similar analytical perspectives may be valuable at the level of research projects [13,14].

Despite these advances, much of the existing literature remains focused on output-oriented indicators, such as publications, citations, or journal metrics, rather than on the organizational and financial properties of scientific projects themselves [15]. Project-level analysis presents distinct methodological challenges, as projects represent temporary, goal-oriented organizational forms that combine financial resources, institutional capacities, and collaborative arrangements. Several strands of research have addressed specific analytical tasks at this level. Cost estimation and budget forecasting models have been developed to support project planning, often drawing on methods from engineering management and operations research. Separately, predictive models have been proposed to estimate research success or impact using proxy measures, such as publication volume or citation counts, as dependent variables [16]. While these studies provide valuable insights, they typically address isolated problems and do not offer an integrated analytical perspective that spans exploration, prediction, and evaluation.

A further limitation concerns the treatment of anomalous projects: many empirical studies exclude projects with exceptionally large budgets, risking biased conclusions regarding strategically important large-scale initiatives [17]. However, such anomalous projects often correspond to strategically important initiatives, including large collaborative infrastructures or mission-oriented research programs, which are of particular interest to funding agencies and policymakers [17]. The systematic exclusion of these cases risks producing biased conclusions and obscuring important governance challenges associated with large-scale scientific investments.

Recent contributions in the field of data-driven public administration and digital governance have argued for the development of intelligent decision-support systems capable of integrating heterogeneous data sources and analytical methods to enhance transparency, efficiency, and accountability in public-sector decision-making [18]. Within this perspective, machine learning is increasingly viewed not merely as a predictive tool, but as a means of generating structured, interpretable insights that can inform strategic and operational decisions. Nevertheless, there remains a lack of comprehensive methodological frameworks combining unsupervised exploration, predictive modeling, and evaluative analysis in a coherent manner.

This study builds on prior research by addressing these limitations through an integrated machine learning-based analysis of scientific projects. The approach combines unsupervised learning for structure identification and anomaly detection, regression models for budget prediction, classification models for funding-completeness assessment, and association rule mining for interpretable pattern extraction. All technical tools used in the study are mature and well-established; accordingly, the contribution is positioned as methodological integration and domain-specific empirical insight for research governance rather than as the proposal of a new machine learning algorithm.

3. Data and Methods

The empirical basis of this study is formed by open data on scientific projects funded under the European Union's Sixth Framework Programme (FP6). The data were obtained from the Community Research and Development Information Service (CORDIS), which provides detailed administrative records on EU-funded research projects, including financial, organizational, and temporal characteristics. FP6 was selected for analysis due to the completeness and consistency of its project-level information, as well as its relevance as a large-scale international research funding programme.

The initial dataset comprised information on 8536 projects and 74,524 participating organizations. The data are organized into two interrelated tables: a project-level table containing general project attributes and a participant-level table describing organizations involved in each project. These tables were linked using a unique project identifier, enabling the aggregation of organizational information at the project level.

Following data cleaning and preprocessing procedures, the final analytical dataset consisted of 5278 projects launched between 2002 and 2008, involving organizations from 50 countries. The reduction in dataset size was primarily due to the removal of projects with incomplete or inconsistent financial information, which is essential for reliable budget modeling.

The data are organized into two interrelated tables linked by a unique project identifier (projectId). Table 1 describes the project-level table; Table 2 describes the participant-level table. Both tables were joined to aggregate organizational information at the project level.

Preprocessing was performed in five sequential steps. The rationale and parameters for each step are described below.

Step 1—Temporal encoding: Project start dates were converted into numerical year values (2002–2008).

Step 2—Financial variable standardization: Budget values were converted to uniform floating-point format; non-numeric characters removed via regular-expression cleaning.

Step 3—Organizational feature aggregation: Participant-level table aggregated at project level yielding: (1) num_organizations, (2) num_countries, (3) coordinator_country (proxy: most frequent participant country if no explicit coordinator).

Step 4—Missing value handling: Projects with missing/null financial data or missing coordinator country after proxy assignment were removed, reducing the dataset from 8536 to 5278 projects.

Step 5—Outlier removal: Projects above the 99th budget percentile were excluded (53 projects, ~1%). Sensitivity to 95th and 97th percentile thresholds produced comparable cluster structures. Final analytical dataset: 5225 projects.

The complete feature set used across all three analytical tasks is summarized in Table 3, which maps each feature to its notation in subsequent mathematical formulations.

Table 1. Data from CORDIS database.

Column	Data Type	Description
projectId	Integer	Unique project identifier (primary key)
acronym	String	Project acronym
title	String	Full title of the project
startDate	Date	Project start date
endDate	Date	Project end date
totalCost	Float	Total project budget (EUR)
ecMaxContribution	Float	Maximum EU financial contribution (EUR)
fundingScheme	String	Type of funding instrument
frameworkProgramme	String	Framework programme identifier (FP6)
topics	String	Thematic research topics/domains
status	String	Project status (CLOSED, TERMINATED, etc.)

Table 2. Participant-level table—attributes describing organizations involved in each project.

Column	Data Type	Description
projectId	Integer	Foreign key linking to project-level table
organizationId	Integer	Unique organization identifier
name	String	Name of the participating organization
country	String	Country of the organization (ISO code)
city	String	City of the organization
role	String	Role in project: Coordinator or Participant
ecContribution	Float	EU financial contribution to this organization (EUR)
activityType	String	Type of organization (HES, REC, PRC, PUB, OTH)
endOfParticipation	Date	Date organization left the project (if applicable)

Table 3. Complete feature set used for modeling (n = 5225 projects).

Feature Name	Symbol	Type	Used in	Description
num_organizations	x_1	Integer	Clustering, Regression, Classification	Number of participating organizations
num_countries	x_2	Integer	Clustering, Regression, Classification	Number of distinct countries
coordinator_country	x_3	Categorical	Clustering, Classification	Country of the lead organization
start_year	x_4	Integer	Regression, Classification	Project start year (2002–2008)
totalCost	x_5/y	Float (EUR)	Regression (target), Classification	Total project budget
ecMaxContribution	x_6	Float (EUR)	Regression, Classification	Maximum EU contribution
budget_per_org	$x_7 = x_5/x_1$	Float (EUR)	Classification	Budget per participating organization
internationalization	$x_8 = x_2/x_1$	Float	Clustering, Classification	Countries-to-organizations ratio
cluster_label	$c \in \{0, 1\}$	Binary	Classification (enriched feature)	Cluster assignment from unsupervised stage
funding_ratio	$R = ec/total$	Float [0, 1]	Target construction	Ratio of granted to requested funding
funding_outcome	$z \in \{0, 1\}$	Binary	Classification (target)	1 if $r \geq 0.70$, else 0

The three analytical tasks are connected through a sequential analytical pipeline designed to support empirical decision-making under uncertainty. For transparency, the

workflow can be represented as an approximation of the posterior probability of funding outcome z_i given observable project characteristics x_i :

$$P(z_i | x_i) = \int P(z_i | x_i, c_i, y_i) \cdot P(c_i | x_i) \cdot P(y_i | x_i) dc_i dy_i \quad (1)$$

where $c_i \in \{0, 1\}$ is the latent cluster assignment and y_i is the predicted budget. In operational terms, the workflow is factorized into three tractable steps: (1) unsupervised estimation of $P(c_i | x_i)$; (2) regression estimation of $E[y_i | x_i]$; and (3) supervised classification of $P(z_i | x_i, c_i, \hat{y}_i)$ using augmented features $[x_i, c_i, \hat{y}_i]$. This formulation is used to organize the empirical analysis and should be interpreted as an integration of existing methods, not as a new probabilistic algorithm.

The first stage identifies latent structural configurations among projects. Input features are $\{x_1, x_2, x_3, x_5, x_7, x_8\}$ —organizational scale, geographic scope, financial parameters, and internationalization ratio. Dimensionality reduction via PCA, t-SNE, and UMAP is applied for visualization.

Hierarchical agglomerative clustering (HAC) with Ward's linkage minimizes total within-cluster variance at each merge step:

$$d(C_\alpha, C_\beta) = (n_\alpha \cdot n_\beta) / (n_\alpha + n_\beta) \cdot \|\mu_\alpha - \mu_\beta\|^2 \quad (2)$$

where C_α and C_β are clusters, n_α and n_β are their sizes, and μ_α, μ_β are centroids. The cluster label $c_i \in \{0, 1\}$ is assigned to each project and passed to the classification stage as an enriched feature. Anomaly detection flags projects whose distance to the nearest centroid exceeds the 99th percentile.

Figure 1 illustrates the relationship between predicted and actual project budgets for the linear regression model. The distribution of points around the ideal prediction line demonstrates that the model provides accurate estimates for small and medium-scale projects, which constitute the majority of the dataset. For larger projects, the dispersion of prediction errors increases, indicating greater uncertainty and reduced model precision at higher budget levels. This pattern reflects the presence of unobserved factors influencing large-scale projects, such as thematic priorities or strategic policy considerations, which are not captured by the available features.

Feature importance analysis further highlights the dominant role of consortium size in budget prediction. The number of participating organizations accounts for more than 60% of the explanatory power in the ensemble models, followed by measures of international collaboration intensity. The country of the coordinating organization and the project start year contribute comparatively less to budget variation.

Building on the descriptive statistics and regression-based budget analysis presented above, further insights into the heterogeneity of FP6 projects were obtained through unsupervised learning techniques. These methods were employed to examine whether the observed variation in budgets and consortium characteristics corresponds to stable structural patterns within the dataset, rather than representing purely continuous variability.

The analysis confirms that FP6 projects are not distributed uniformly in the feature space. Instead, distinct structural configurations emerge when organizational scale, international composition, and financial parameters are considered jointly. Dimensionality reduction results indicate that a substantial share of variance can be explained by a limited number of latent dimensions, while nonlinear projections reveal clearer separation between groups of projects with different structural profiles. This suggests that project characteristics cluster around a small number of recurring configurations rather than forming a single homogeneous population. Figure 2 provides a comparative visualization of linear and nonlinear dimensionality reduction techniques applied to the FP6 project dataset.

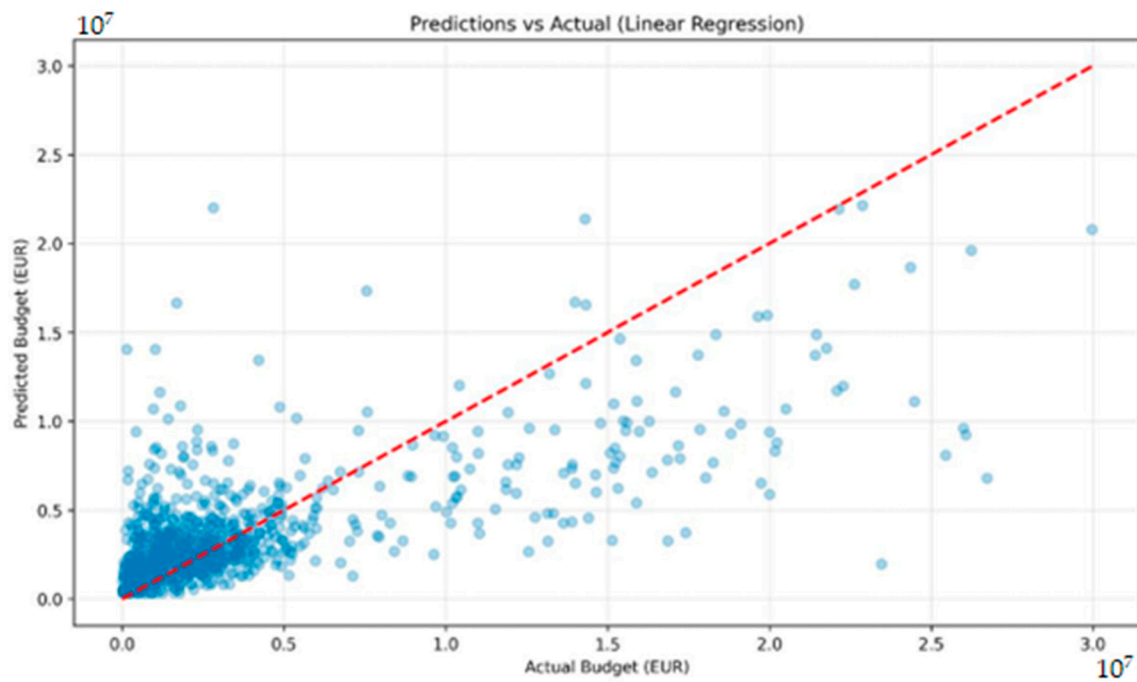


Figure 1. Predicted vs. actual project budgets using the linear regression model.

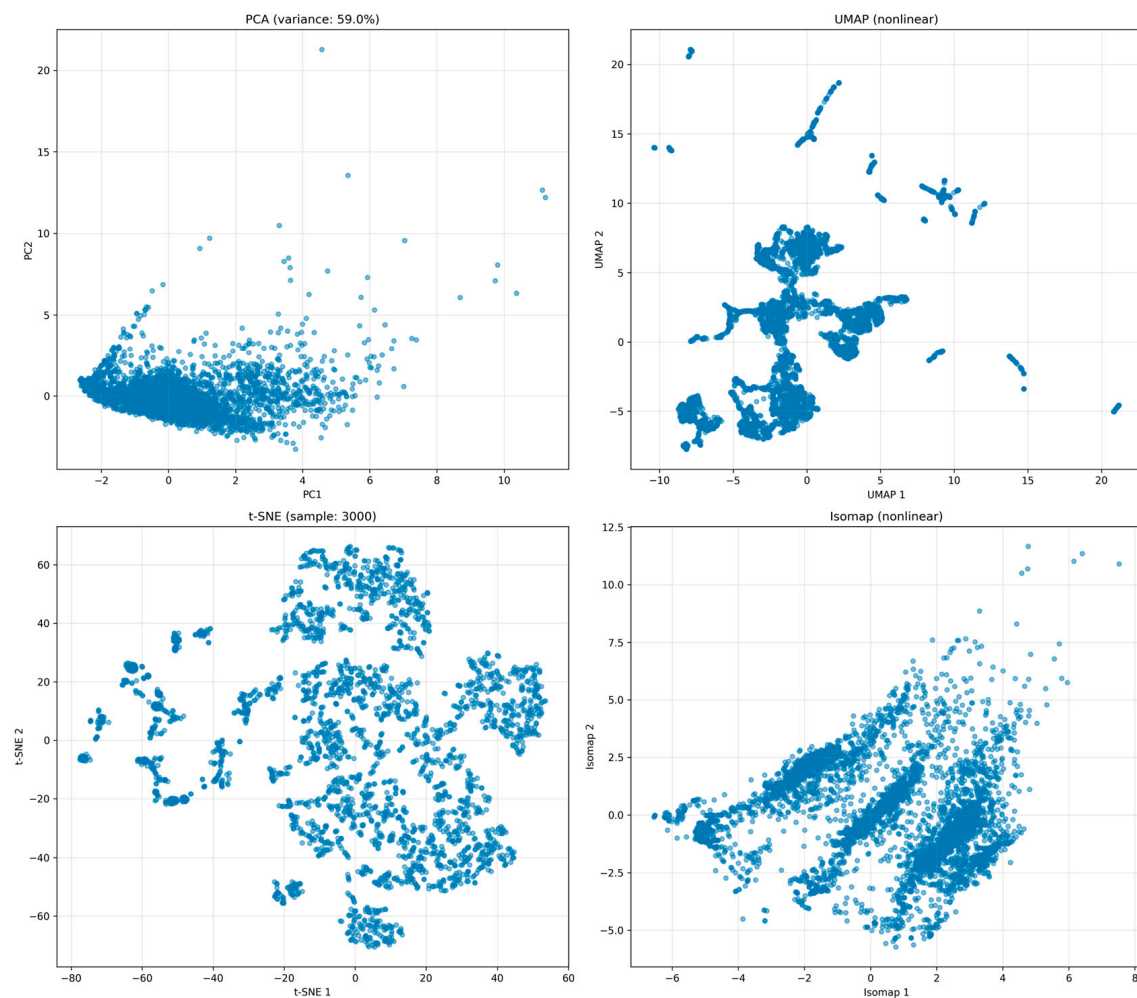


Figure 2. Dimensionality reduction comparison (FP6 dataset).

The comparison of dimensionality reduction techniques illustrates that while linear projection captures dominant variance directions, nonlinear methods provide a more informative representation of project heterogeneity. UMAP and t-SNE reveal compact groups and isolated observations, supporting the existence of distinct project configurations rather than a continuous distribution of characteristics.

To further examine whether the structural patterns revealed by dimensionality reduction correspond to stable and interpretable groupings of projects, hierarchical clustering was applied to the reduced feature space. While dimensionality reduction techniques demonstrated that FP6 projects occupy a heterogeneous and non-uniform space, clustering analysis enables the identification of discrete project groupings characterized by similar organizational and financial configurations.

The second stage estimates total project cost y_i from features $\{x_1, x_2, x_3, x_4, x_6\}$:

$$y_i = f(x_i) + \varepsilon_i \quad (3)$$

where y_i is the total budget (EUR), $x_i \in \mathbb{R}^m$ ($m = 5$) is the input feature vector, and ε_i is a stochastic error term capturing unobserved factors.

For the linear baseline:

$$y_i = \beta_0 + \sum_{j=1}^m \beta_j x_{ij} + \varepsilon_i \quad (4)$$

Coefficients β_j are estimated by minimizing the residual sum of squares:

$$\min_{\beta} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

Ensemble models approximate f as a weighted combination of K decision trees $h_k(x)$:

$$f(x) = \sum_{k=1}^K w_k h_k(x) \quad (6)$$

Overfitting is diagnosed when training error is substantially lower than test error:

$$E_{\text{train}}[(y_i - \hat{y}_i)^2] \ll E_{\text{test}}[(y_i - \hat{y}_i)^2] \quad (7)$$

Table 4 presents the main characteristics of the two identified clusters. The first cluster comprises approximately 72% of projects and is characterized by relatively small budgets, compact consortia, and higher internationalization ratios. The second cluster, accounting for roughly 28% of projects, corresponds to large-scale initiatives with significantly higher budgets, substantially larger consortia, and a lower relative level of internationalization per participating organization. The magnitude of the differences across clusters indicates that the identified grouping captures meaningful distinctions in project scale and organizational structure rather than marginal variation.

Table 4. Cluster results.

Characteristic	Cluster 0 (Small Projects)	Cluster 1 (Large Projects)	Relationship
Number of projects	~3700 (72%)	~1400 (28%)	2.6:1
Average budget	1.62 M EUR	10.41 M EUR	6.4×
Median budget	1.25 M EUR	8.74 M EUR	7.0×
Average number of organizations	6.75	21.44	3.2×
The median of organizations	7	20	2.9×
Average number of countries	4.48	10.13	2.3×
Median of countries	5	10	2.0×
The level of internationalization	0.75	0.49	0.7×
The budget for the organization	260 K EUR	539 K EUR	2.1×

The target variable is constructed from the funding ratio $r_i = \text{ecMaxContribution}_i / \text{totalCost}_i$:

$$z_i = 1 \text{ if } r_i \geq 0.70, \text{ else } 0 \quad (8)$$

The threshold $\tau = 0.70$ was selected based on the empirical distribution of r_i in the FP6 dataset, where approximately 65% of projects receive coverage ≥ 0.70 , creating a meaningful and non-degenerate binary split. Alternative values $\tau \in \{0.60, 0.65, 0.75, 0.80\}$ were also evaluated; all produced qualitatively similar results with accuracy differences < 2 percentage points. Since CORDIS contains only approved projects, $z = 0$ indicates a budget adjustment rather than project rejection—explicitly acknowledged as a limitation.

The classification feature vector is the augmented set $[x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, c]$. Six algorithms are evaluated: logistic regression, decision tree, random forest, gradient boosting, XGBoost, MLP, and k-NN. All models are trained using stratified 5-fold cross-validation.

Model performance is assessed using accuracy, F1-score, and AUC:

$$\text{AUC} = \int_0^1 \text{TPR}(t) d(\text{FPR}(t)) \quad (9)$$

where higher AUC values indicate better classification performance.

Figure 3 illustrates the ROC curves for all evaluated models. The results demonstrate consistently high predictive performance, with most models achieving AUC values above 0.97, indicating strong separability between fully and partially funded projects.

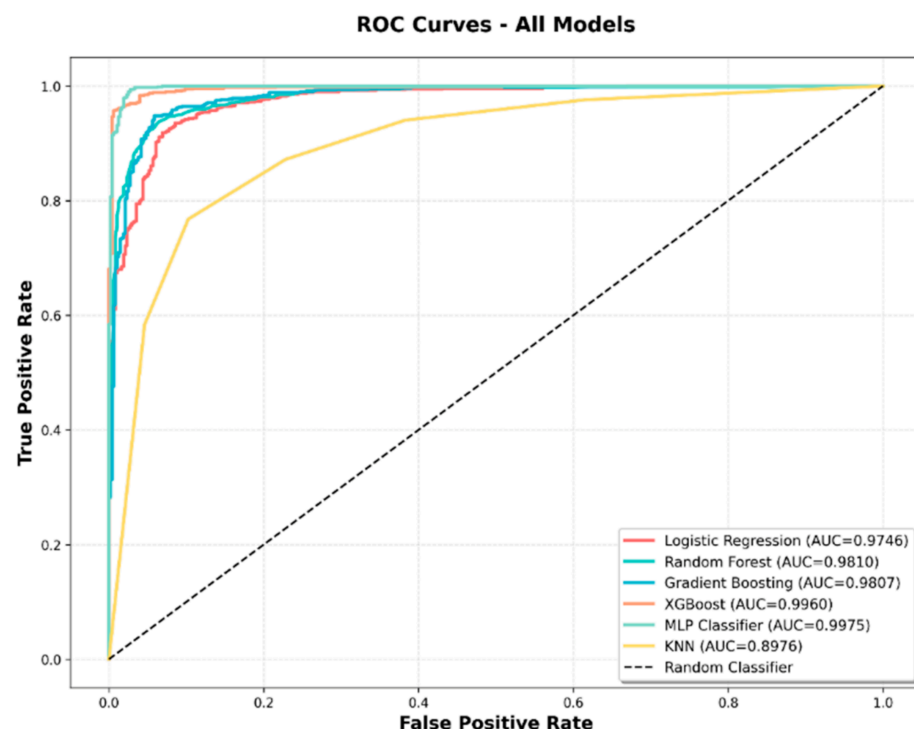


Figure 3. ROC curves for all evaluated classification models, demonstrating high predictive performance and strong separability between fully and partially funded projects.

Across all evaluated models, classification results demonstrate consistently high predictive performance, indicating that funding-completeness categories are strongly associated with observable project characteristics in the FP6 dataset. All six models achieved accuracy values above 83%, confirming systematic relationships between consortium structure, financial parameters, and the constructed funding-completeness target. The strongest

performance was obtained using nonlinear models, particularly the multilayer perceptron and gradient boosting-based approaches; however, these results are interpreted as evidence of structural regularity in the data rather than as proof of a new methodological advance.

Association rule mining is applied to the discretized feature space. Features x_1 , x_2 , x_5 , x_6 are binned into low/medium/high categories (tertile thresholds). The Apriori algorithm is run with minimum support = 0.05 and minimum confidence = 0.60. Rules are ranked by lift:

$$\text{lift}(A \rightarrow B) = P(A \cup B) / (P(A) \cdot P(B)) \quad (10)$$

The multilayer perceptron achieved the highest overall performance, with an accuracy of 98.50%, an F1-score of 98.39%, and a ROC-AUC of 0.9975. This model exhibited a very low error rate, with only two false negatives in the test set, indicating a high level of sensitivity in identifying fully funded projects. XGBoost demonstrated similarly strong performance, achieving an accuracy of 96.84% and the most stable cross-validation results, suggesting robustness to variations in the training data. In contrast, linear logistic regression achieved lower but still substantial accuracy (90.92%), indicating that a significant portion of the decision boundary is linearly separable, while nonlinear interactions further improve classification performance.

Analysis of ROC curves shows that all models, except k-nearest neighbors, exhibit strong discriminatory power, with ROC-AUC values exceeding 0.97. The limited performance of k-nearest neighbors reflects the heterogeneous and unevenly distributed nature of projects in the feature space, which was previously observed in the unsupervised analysis. Importantly, misclassification errors are concentrated among projects with intermediate budgets and consortium sizes, while projects at the extremes of the structural spectrum are classified with substantially higher accuracy.

To complement predictive results with interpretable insights, associative rule mining was applied to identify frequent patterns linked to funding outcomes. Rules with the highest lift values reveal that projects involving many participating countries are substantially more likely to receive full funding, while excessively large consortia combined with high requested budgets are strongly associated with partial funding outcomes. The strongest rule for partial funding exhibits a lift of 6.30, indicating a markedly increased likelihood of reduced funding for overly ambitious project configurations.

Finally, feature importance analysis across multiple models confirms the dominance of financial variables in funding-completeness prediction. Total project cost, maximum EU contribution, and budget per organization jointly account for approximately 69% of explanatory importance, while consortium size and thematic domain provide additional discriminatory power. The convergence of results across classification models, association rules, and feature importance measures reinforces the empirical reliability of the identified patterns and highlights the central role of structural and financial configuration in determining funding completeness.

4. Results

Dimensionality reduction confirms that FP6 projects are not uniformly distributed in the feature space. PCA reveals that the first two principal components explain approximately 68% of total variance. Figure 1 shows the PCA projection, while Figure 2 presents the UMAP projection. The cluster profile comparison of key characteristics is presented in Table 5.

Table 5. Cluster profile comparison of key characteristics.

Characteristic	Cluster 0 (Small Projects)	Cluster 1 (Large Projects)	Ratio (C1/C0)
Number of projects	~3700 (72%)	~1525 (28%)	—
Mean totalCost (x_5)	1.62 M EUR	10.41 M EUR	6.4×
Median totalCost (x_5)	1.25 M EUR	8.74 M EUR	7.0×
Mean num_organizations (x_1)	6.75	21.44	3.2×
Mean num_countries (x_2)	4.48	10.13	2.3×
Mean internationalization (x_8)	0.75	0.49	0.65×
Mean budget_per_org (x_7)	260 K EUR	539 K EUR	2.1×

Cluster 0 (72%): compact consortia, higher internationalization, modest budgets. Cluster 1 (28%): large-scale initiatives with substantially larger consortia and budgets. Differences are statistically significant (Mann–Whitney U, $p < 0.001$). Anomaly detection identified 41 atypical projects; none were misclassified in the supervised stage.

Table 6 summarizes regression model performance on the held-out test set ($n = 1045$ projects). Confidence intervals computed via bootstrap resampling (1000 iterations).

Table 6. Regression model performance on the test set.

Model	R ² (Test)	95% CI	RMSE (EUR)	Train R ²	Overfitting?
Linear Regression	0.499	[0.471, 0.527]	2,140,000	0.502	No
Ridge Regression ($\lambda = 1.0$)	0.501	[0.473, 0.529]	2,135,000	0.503	No
Random Forest ($K = 100$)	0.503	[0.475, 0.531]	2,128,000	0.741	Yes
Gradient Boosting	0.488	[0.461, 0.516]	2,167,000	0.782	Yes

All models achieve $R^2 \approx 0.50$ on the test set. This result warrants critical interpretation: the unexplained variance ($\approx 50\%$) reflects latent factors not captured in administrative data—thematic priorities, geopolitical considerations, and qualitative aspects of the research agenda. Linear and ridge models generalize well (train $\approx$ test R^2), confirming an approximately linear budget–organization relationship. Ensemble methods overfit without improving test performance.

All models achieve accuracy above 83%, with nonlinear models reaching 96–98%. The high AUC values (>0.97) raise a legitimate concern about potential data leakage and dependence between predictors and the constructed target variable. Models retrained excluding x_6 (ecMaxContribution) yielded AUC = 0.91–0.95, which substantially mitigates, but does not fully eliminate, this concern. The remaining high performance is therefore interpreted cautiously as evidence of structural regularity within the FP6 programme rather than as a universally generalizable result. Cross-validation results confirm stability across folds (CV Std < 0.01 for best models).

Association rule mining identified 127 rules meeting minimum thresholds. Table 7 presents the eight highest-lift rules; the full rule set.

Table 7. Top-eight association rules by lift value.

Antecedent	Consequent	Support	Confidence	Lift
high totalCost $\wedge$ high num_org	$z = 0$ (partial)	0.08	0.81	6.30
high num_countries $\wedge$ low budget_per_org	$z = 1$ (full)	0.12	0.87	4.92
cluster = 1 $\wedge$ high totalCost	$z = 0$ (partial)	0.10	0.79	4.51
low num_org $\wedge$ high internationalization	$z = 1$ (full)	0.14	0.83	4.38
medium totalCost $\wedge$ medium num_org	$z = 1$ (full)	0.18	0.78	3.97
high ecMaxContribution $\wedge$ cluster = 1	$z = 0$ (partial)	0.07	0.75	3.84
low totalCost $\wedge$ high num_countries	$z = 1$ (full)	0.09	0.80	3.72
high num_org $\wedge$ low internationalization	$z = 0$ (partial)	0.06	0.72	3.61

5. Discussion and Limitations

The central contribution of the study is not the introduction of new machine learning algorithms, but the systematic integration of mature analytical tools for project-level research funding analysis. The stable two-cluster structure separating small-scale and large-scale projects is robust across the analytical stages. The dominance of consortium size (x_1) as the primary budget predictor (61% of feature importance) indicates that financial requirements in FP6 are driven by coordination complexity, consistent with prior research on collaboration structures in science [12,13].

The $R^2 \approx 0.50$ for budget regression deserves explicit discussion. The unexplained half of budget variance likely reflects thematic and strategic factors absent from administrative records: research priority areas, geopolitical considerations, and qualitative project characteristics that influence funding decisions but leave no trace in the CORDIS data. This finding itself carries a governance implication—richer data sources (topic models applied to project abstracts, reviewer scores) could substantially improve explanatory power.

The high AUC values (>0.97) are unusual and require careful interpretation. As discussed above, removing `ecMaxContribution` reduces AUC to 0.91–0.95, indicating partial but not complete dependence on a target-correlated feature. The residual high performance appears to reflect the structural regularity of the FP6 programme, but this result may not generalize to programmes with different funding logics. External validation is therefore essential before applying the workflow as an operational decision-support tool.

First, the analysis is based on a single funding programme (FP6); findings may not transfer directly to FP7, Horizon 2020, or Horizon Europe. External validation was not performed and constitutes a priority for future work.

Second, the operationalization of funding success as $r \geq 0.70$ conflates budget adjustment with funding completeness. Within a dataset of only approved projects, $z = 0$ indicates a negotiated budget reduction, not project failure.

Third, the feature set is restricted to administrative and organizational descriptors. Thematic content, scientific quality signals, and applicant track records are unavailable in CORDIS open data but are known to influence outcomes.

Fourth, comparison with non-ML baseline methods was not performed. Such comparison would be necessary to empirically demonstrate the added value of the integrated analytical pipeline relative to descriptive statistics or expert-only assessment procedures.

Overall, the primary contribution lies in methodological integration and empirical insights: the study shows how structure identification, budget prediction, funding-completeness classification, and association rule mining can be combined within one reproducible workflow for research management. Future work should pursue: (1) external validation on FP7/Horizon 2020; (2) enrichment with topic models; (3) comparison with expert evaluation and non-ML baselines; (4) development of a fully Bayesian probabilistic formulation; and (5) fairness metric analysis for geographic and organizational biases.

6. Conclusions

This paper presented an integrated machine learning-based analysis applied to 5225 projects from the CORDIS FP6 database. The study does not claim to develop new algorithms; instead, it demonstrates how established methods can be combined to generate empirical insights for scientific project governance. Three key conclusions emerge.

First, FP6 projects form two structurally distinct configurations. Consortium size is the dominant driver of both budget scale and cluster membership, accounting for over 60% of predictive importance in regression models.

Second, budget prediction is achievable with $R^2 \approx 0.50$ using organizational and collaboration features alone. The 50% unexplained variance reflects genuine latent fac-

tors absent from administrative data, not a modeling failure. Ensemble methods do not outperform linear models on the test set.

Third, funding completeness is predictable with high accuracy (AUC > 0.97 for most models), but this result should be interpreted cautiously because the target variable is constructed from financial variables available in the dataset. Feature ablation experiments reduce, but do not entirely remove, this concern. Association rule mining identifies interpretable patterns: high country diversity with low budget-per-organization predicts full funding; large consortia with high total costs are associated with partial funding.

These findings demonstrate that integrated machine learning analysis can support evidence-based decision-making in research funding governance when used as a complementary tool alongside expert assessment. The value of the work lies in the systematic integration of analytical tasks and the domain-specific insights obtained from FP6 data, subject to the limitations discussed and the external validation that future work should provide.

Author Contributions: Conceptualization, A.M. (Ailym Medetbek), A.M. (Aidos Mukhatayev) and A.B.; methodology, A.M. (Ailym Medetbek) and S.B.; software, D.B.; validation, A.B., S.B. and A.M. (Aidos Mukhatayev); formal analysis, A.B.; investigation, A.M. (Ailym Medetbek); resources, A.M. (Ailym Medetbek); data curation, D.B.; writing—original draft preparation, A.M. (Ailym Medetbek); writing—review and editing, A.M. (Aidos Mukhatayev) and S.B.; visualization, D.B.; supervision, A.B. and A.M. (Aidos Mukhatayev). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in European data at <https://data.europa.eu/data/datasets/cordisfp6projects?locale=en> (accessed on 4 April 2026).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. European Union Open Data Portal. CORDIS FP6 Projects Dataset. Available online: <https://data.europa.eu/data/datasets/cordisfp6projects> (accessed on 4 April 2026).
2. European Commission. Sixth Framework Programme (FP6), 2002–2006. Available online: <https://cordis.europa.eu/programme/id/FP6> (accessed on 4 April 2026).
3. Biloshchytskyi, A.; Kuchansky, A.; Andrashko, Y.; Biloshchytska, S.; Kuzka, O.; Shabala, Y.; Lyashchenko, T. A method for the identification of scientists' research areas based on a cluster analysis of scientific publications. *East.-Eur. J. Enterp. Technol.* **2017**, *5*, 4–11. [CrossRef]
4. Bornmann, L. Measuring impact in research evaluations: A thorough discussion of methods for, effects of, and problems with impact measurements. *High Educ.* **2017**, *73*, 775–787. [CrossRef]
5. Zhang, T.; Qin, S. Structural Deformation Prediction and Uncertainty Quantification via Physics-Informed Data-Driven Learning. *Appl. Sci.* **2026**, *16*, 3194. [CrossRef]
6. Biloshchytskyi, A.; Omirbayev, S.; Mukhatayev, A.; Kuchanskyi, O.; Biloshchytska, S.; Andrashko, Y.; Toxanov, S.; Faizullin, A. A structural model for building a system for the development of methodological competence and methods for evaluating its effectiveness. *East.-Eur. J. Enterp. Technol.* **2023**, *5*, 6–22. [CrossRef]
7. Reis, V.B.; Baptista, J.S.; Duarte, J. Machine Learning in Surface Mining—A Systematic Review. *Appl. Sci.* **2026**, *16*, 3246. [CrossRef]
8. Biloshchytskyi, A.; Omirbayev, S.; Mukhatayev, A.; Faizullin, A.; Toxanov, S.; Kassenov, K. Research on the formation level of methodological competence of it disciplines teachers. In *Proceedings of the 2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT), Kyiv, Ukraine, 25–27 November 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 242–245. [CrossRef]
9. Liu, Z.; Fan, L.; Chen, Q.; Li, H.; Wei, A. Performance Evaluation of Publicly Funded Agricultural Research Projects with Light-TabNet. *Appl. Sci.* **2026**, *16*, 3230. [CrossRef]

10. Liu, J.; Gratz, J.; Amour, C.; Nshama, R.; Walongo, T.; Maro, A.; Mduma, E.; Platts-Mills, J.; Boisen, N.; Nataro, J.; et al. Optimization of Quantitative PCR Methods for Enteropathogen Detection. *PLoS ONE* **2016**, *11*, e0158199. [[CrossRef](#)] [[PubMed](#)]
11. Biloshchytskyi, A.; Kuchanskyi, O.; Mukhatayev, A.; Andrashko, Y.; Toxanov, S.; Faizullin, A.; Kassenov, K. Application of Time-Weighted PageRank Method with Citation Intensity for Assessing the Recent Publication Productivity and Partners Selection in R&D Collaboration. *Publications* **2024**, *12*, 48. [[CrossRef](#)]
12. Bertsimas, D.; Kallus, N. From predictive to prescriptive analytics. *Manag. Sci.* **2020**, *66*, 1025–1044. [[CrossRef](#)]
13. Biloshchytskyi, A.; Omirbayev, S.; Mukhatayev, A.; Biloshchytska, S.; Toxanov, S.; Faizullin, A. The concept of the Internet of Things in the development of information and analytical systems based on the method of constructing a scalar assessment of the results of research activities of scientists. *Procedia Comput. Sci.* **2024**, *231*, 684–690. [[CrossRef](#)]
14. He, C.; Duan, Y.; Mamat, T.; Zhu, X.; Dridi, M.; Mualla, Y.; Abbas-Turki, A. Research on a Prediction Method for Maintenance Decision of Expressway Asphalt Pavement Based on Random Forest. *Appl. Sci.* **2026**, *16*, 3343. [[CrossRef](#)]
15. Ashraf, R.; Khan, M.A.; Khuhro, R.A.; Bhatti, Z.A. Knowledge creation dynamics of technological forecasting and social change special issues. *Technol. Forecast. Soc. Change* **2022**, *180*, 121663. [[CrossRef](#)]
16. OECD. Artificial Intelligence, Machine Learning and Big Data in Public Policy. 2021. Available online: <https://www.oecd.org/gov/digital-government/ai-ml-big-data-public-policy.htm> (accessed on 5 April 2026).
17. Okamura, K. Scientometric engineering: Exploring citation dynamics via arXiv eprints. *Quant. Sci. Stud.* **2022**, *3*, 122–146. [[CrossRef](#)]
18. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI. *Nat. Mach. Intell.* **2020**, *2*, 252–259. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.