Università degli Studi del Molise
Dipartimento di Economia

Economics and Statistics Discussion Paper
No. 099/25

Predicting Delays in Cohesion Infrastructure Projects

Giuseppe Coco
Gianluca Monturano
Giuliano Resce

(This page intentionally left blank.)

Predicting Delays in Cohesion Infrastructure Projects*

Giuseppe Coco[1], Gianluca Monturano [2,3], and Giuliano Resce[4]

[1]*Department of Economics and Finance, University of Bari*
[2]*Department of Economics, Management and Business Law, University of Bari*
[3]*Department of Economics, University of Modena and Reggio Emilia*
[4]*Department of Economics, University of Molise*

Abstract

Public investment in infrastructure is essential for economic growth, but delays in project implementation can undermine its benefits. This paper examines the determinants of such delays using data from cohesion projects in Italy. We predict which projects are likely to experience delays and identify the key contributing factors by means of machine learning (ML) techniques. To avoid endogeneity, we use only (lagged) features observed at the start of the project as predictors. Our findings show that socioeconomic factors and institutional weaknesses in various regions play a significant role in these delays. The discipline imposed by rules and strict implementation timing on EU funds seems to work, lending credibility to the hypothesis of the benefit of an external commitment. Results underscore the potential of ML in designing appropriate implementation policies, enhancing project management, and improving the outcomes of public investments.

Keywords: Territorial cohesion, Administrative efficiency, Machine learning, Project Delays

JEL Classification: H77, R58, O18, C55

*The Authors would like to thank Riccardo Crescenzi, Raffaele Lagravinese and partecipants in the presentation at RCEA International Conference in Economics, Econometrics and Finance, ERSA International Congress, Annual Conference of the Italian Economic Society, and AISRE, for useful comments. All responsibility for mistakes remains the authors'.

1 Introduction

There exists a widespread agreement among scholars that public investment, with a specific emphasis on infrastructure development, plays a crucial role in fostering economic growth and local development (Aschauer, 1989; Calderón et al., 2015; Leduc and Wilson, 2013; Ramey et al., 2021). In this sense, Chaurey and Le (2022) highlight that infrastructure development and maintenance programs, especially in vulnerable local contexts marked by socio-economic challenges, can increase employment and economic activity, encouraging the creation of new businesses. Acknowledging this, a substantial share of cohesion policies is committed to fostering infrastructure initiatives, as evidenced by the allocation of almost half of the European Union's cohesion policy investments during the 2014–2020 programming period towards infrastructure projects (European Commission, 2022). These projects, in the operational implementation phases, are often marked by delays and inefficiencies, due to absorptive capacity constraints, lack of skills, and management issues, that can undermine the benefits of public investment (Espinoza and Presbitero, 2022). Particularly in the cohesion context, local government may lack the capacity to select and manage projects efficiently, leading to cost inflation and delays (Surubaru, 2017; Incaltarau et al., 2020). It has been shown that delays negatively impact project outcomes, with a positive correlation between time and cost overruns (Espinoza and Presbitero, 2022).

This paper investigates the determinants of project delays by leveraging data on infrastructure cohesion projects implemented in Italy in the 2007-2013 wave. We use a Machine Learning (ML) model to predict projects that fail to be completed on schedule. Although traditional statistical and econometric tools have historically probed this issue in alignment with standard economic analysis, contemporary economic literature suggests a shift toward exploiting advancements in ML for addressing policy-related challenges, particularly with very large databases. This perspective contends that the potential of ML surpasses that of conventional econometric techniques for the prediction of policy problems, as argued by Kleinberg et al. (2015). ML techniques are gaining traction in addressing various challenges related to poverty targeting (Jean et al., 2016), evaluating the efficacy of public programs and spending (Andini et al., 2018; Christensen et al., 2024), identifying instances of corruption and political connections (De Blasio et al., 2022; Titl et al., 2024), and on the study of environmental and health problems (Fabra et al., 2022; Girma and Paton, 2024). In the specific context of Italy, recent studies have harnessed ML's potential to forecast the potential bankruptcy of local governments (Antulov-Fantulin et al., 2021), to predict vaccine hesitancy in municipalities (Carrieri et al., 2021), to estimate local mortality and inequality during the COVID-19 pandemic (Cerqua et al., 2021), to foresee Geographical Indications areas (Resce and Vaquero-Piñeiro, 2022), and to anticipate dropout rates in higher education (Delogu et al., 2024).

Predicting which infrastructure projects are likely to experience delays could become a valuable tool for proactively addressing potential project weaknesses and evaluating the strengths and weaknesses of different institutional and financial arrangements. Our results indicate that it is possible to forecast which projects will not be completed on time, and that territorial socioeconomic factors play a significant role.

Italy offers a unique opportunity for investigating delays in cohesion infrastructure projects. Cohesion budgets (125 bil euros in the cycle 2007-13) and the number of projects are both very large. Moreover, the Italian Cohesion Agency provides a unique publicly available database (Opencoesione), with complete records of programmed and actual dates of realization of all projects financed with cohesion funds. The record also includes financial and sectoral features for each project and geographic information (location). We use this database as a basis, adding information gathered from many different sources for demographic, administrative, economic, geographic, and quality of government indicators, based on the specific location of the project. Importantly, all these additional variables are lagged, reflecting conditions at the start of each project, to avoid issues of endogeneity and reverse causality that might arise if using contemporaneous or post-delay observations.

The remainder of the paper is structured as follows: Section 2 introduces the institutional context in which cohesion programs are inserted, discussing the structures and rules that regulate interactions between local authorities and supranational institutions, along with the evidence about its effectiveness; Section 3 presents the dataset used for the analysis, the variable selection criteria, and descriptive statistics that highlight the main characteristics of the data; Section 4 describes the empirical approach adopted for data analysis, including statistical models and methodologies for evaluating influential variables; Section 5 reports the results of the prediction estimates in terms of classification and regression; Section 6 shows the territorial heterogeneities that influence delays; Section 7 concludes the paper, summarizing the main findings, implications for policymakers and future directions for research in this area.

2 Institutional Setting and evidence

Cohesion policy is one of the pillars of the European Union (EU) since its inception. Roughly one-third of the EU Budget is devoted to Cohesion, and the intervention is for the most part distributed in two funds, the ERDF and the ESF. Since 2000, Programming periods for these funds have been extended for 7 years, and the allocation of resources to member states varies according to parameters that change in every programming period. Italy has always been a large beneficiary of the cohesion policy, due to the persistent problem of the *Mezzogiorno*, the largest less-developed area in Western Europe. Cohesion policies allocate funds on the basis of the level of development of different regions. All regions get some funds, but those with GDP per capita below 75% of the EU average get the largest share. Cohesion policies in each country are programmed with a Strategic National Framework (SNF) (Ministero per lo Sviluppo Economico, 2007). The SNF for the period 2007-13 approved for Italy by the EU Commission, was financed with 28.7 bil euro of EU funds (ERDF+ESF). The SNF required Italy to complement those funds with additional co-financing (31.6 bil. Euros). The Framework also envisaged the use of an additional national fund, a form of more flexible finance in which the member state had greater discretion about its use - the Underused Areas Fund ("Fondo Aree Sottoutilizzate", FAS))[1] endowed with 64.4 bil Euros. Thus, the overall endowment of cohesion policies in Italy for the seven-year programming period was 124.7 bil Euros. Of these funds, more than 100 bil Euros were allocated to the Southern less-developed regions. The Opencoesione database gathers information about all projects financed under this array of funds, even when partially financed by ordinary policies. The overall programming for 2007-13 was streamlined according to the ten priorities of the policy. The five priorities receiving the largest share of the funds were: Network and mobility, Competitiveness, Energy and Environment, Research and Innovation, and Human Resources. EU funds, National Cofinancing Funds, and FAS were allocated to Operative Plans for

[1]Later to be renamed definitively in Italian Fondo Sviluppo e Coesione (FSC).

Regions (POR))[2] and for Ministries (PON)[3] according to priorities. The overall allocation of funds was, thus, quite remarkable and complex. However, the policy ran into considerable difficulties in conjunction with the debt crisis that gripped the country. Nevertheless, it was mainly because of delays in the implementation of the programs that the government decided to ask the EU for a reformulation of the Framework. In particular, it asked for the possibility of reducing some national co-financing. Delays in delivering programs were particularly important at the regional level in the Mezzogiorno, and, considering the rules that required automatic de-financing of incomplete programs at the second year after the end of the programming period (the n+2 rule), the government decided to replan allowing for less national funding. This de-financing cut the national funding by half, canceling investments for 13.5 billion. At the same time, new regulations and programs of assistance were adopted for some administrations that were particularly late in delivering. Italy has never missed targets on final expenditure agreed with the EU Commission, but it is consistently among the countries in which delays in enactment of cohesion policy are greater and that frequently delivers most of the expenditure in the last two or three years. This was particularly true in the 2007-13 programming period when a large chunk of expenditure occurred in 2015, the last available year for using the funds. Several other elements point to a more general difficulty of the public administration in delivering public investment. Public investment as a share of GDP has been consistently diminishing since 2009, from an average of 3% of GDP in the first decade of the century to a minimal percentage of 1.9% in 2019.

Although general tests about the effectiveness of the EU cohesion policy signal a moderate positive effect as demonstrated in several studies by Becker et al. (2010, 2012, 2018), territorial heterogeneity seems to be large (Iammarino et al., 2019). Results about the effectiveness of the policy in Italy are rather dismal (Crescenzi and Giua, 2020). A large literature highlights scant results of individual policy measures (Accetturo and De Blasio, 2012) or for the overall policy (Barone et al., 2016; Albanese et al., 2022). Many studies point to the relevance of the quality of government (Rodríguez-Pose, 2013; Accetturo et al., 2014; Rodríguez-Pose and Garcilazo, 2015). Crescenzi and Giua (2020) discuss the composition of thematic expenditure and conclude that most of the positive effects of cohesion policy in Europe is concentrated in

[2] In Italian Piani Operativi Regionali.

[3] In Italian Piani Operativi Nazionali.

Germany and the UK. Expenditure in innovation is particularly important. Expenditure in Italy seems, on the contrary, concentrated on keeping alive low-productivity firms. Coco and Lagravinese (2021) also highlight the differences in the use of funds by identifying classes of expenditure that are more and less capital-intensive, and find that Italian regions display a comparatively less capital-intensive composition of expenditure. It is particularly interesting in the light of our main result, that Coppola et al. (2018) find that European structural funds allocation seems to be less sensitive to the local context (i.e. quality of government) relative to national funds, probably due to rigidity in the allocation and procedures. From this point of view, they argue that European funds are more effective in pursuing long-term development targets.

Delays in delivering investment projects are clearly correlated with the effectiveness and efficiency of expenditure. Problems arising from delays range from increased costs to obsolescence of projects to loss of potential growth. There is also a presumption that delays are (inversely) correlated with the quality of public investment as both depend on the quality of the public administration. Awareness of the problem posed by delays is acute also at the institutional level. In 2018, the Agency for Territorial Cohesion published a Report on the times of implementation of Public Works, a detailed statistical analysis of correlation with certain variables (Agenzia per la Coesione Territoriale, 2018). This Report did not present any causal inference, but it highlighted some obvious and some less obvious facts. Large projects were more likely to run late; a large fraction of the implementation length was due to so-called 'cross-through times', that is, times that could not be attributed to one phase or the next. A significant heterogeneity among regions emerged, in which some, but not all, southern regions displayed longer times. However, these analyses were simple univariate statistical correlations and could not be used for policy advice. Notwithstanding its crucial relevance, the academic literature on the topic of delays is quite scarce. D'Alpaos et al. (2013) show that delays in infrastructure public works at the realization stage can be due to opportunistic behavior on the part of firms and that the probability of being levied a penalty for overrun is the main determinant. They then proceed to show that indeed the differential efficiency of courts can partly explain differentials in time overrun for public works among regions. At the international level, Espinoza and Presbitero (2022) show that country characteristics significantly impact delays, with projects in countries with weaker institutions and during periods of increased public investment facing longer delays.

Wong et al. (2017) underline how a more careful and inclusive planning and management of infrastructure projects, supported by participatory governance and the active involvement of local communities in decisions, can improve the quality of projects and ensure lasting economic benefits. Coviello and Mariniello (2014) show that the cost efficiency of a procurement procedure can be enhanced by greater publicity of the auction, and that is not associated with more delays. A small strand of literature focuses on the best mechanism to prevent delays in the phase of contract award (Reeves et al., 2017), particularly in the case of public-private partnership (PPP) contracts (Palcic et al., 2022). More recently, He et al. (2023) found that negotiated procedures for contract awards might be more timely and efficient and, based on an analysis of UK data, identified the organizational features that made the award procedure faster.

3 Dataset and descriptive statistics

3.1 The dataset

This work intends to study the most relevant features of the inefficiencies of cohesion projects and, in particular, to evaluate the features and impacts of these on project delays. To this end, we exploit the information heterogeneity of a very rich database for public use, available on the government portal Opencoesione.it[4], which reports the details of all the territorial cohesion projects implemented in Italy relating to the European programming cycles, 2007-2013; 2014-2020; 2021-2027[5]. Specifically, for each individual project, the database reports a large amount of procedural information (for example the type of tender, the implementation sector, the nature and operations of the bodies involved in the various procedural phases, presented in Figure 2), financial information (quantity of resources assigned to each project, type and nature of financing, etc.), as well as on the temporal dimension (with the expected and actual start and end dates of the projects and of the multiple intermediate phases). Furthermore, it provides details on the territorial scale of reference, with information on the geolocation of the projects, up to the granular municipal level. Consequently, each Italian municipality can be involved in multiple cohesion projects during the annual programming cycles. These interventions can extend on a larger scale than a single

[4] The database can be freely reached at this link: https://opencoesione.gov.it/it/

[5] At the time of writing (2024), the European programming cycle 2021-2027 was still being defined.

administrative unit (for example the construction or modernization of physical infrastructures), affecting multiple regions, provinces and municipalities (multi-municipal interventions).

Operationally, for the predictive estimates, following our empirical design illustrated in detail in Section 4, we proceed to select only the concluded cohesion projects[6] for the 2007-2013 programming cycle. We therefore exclude projects that are not entirely completed in 2024 and consider only the infrastructural projects that concern the theme of "construction of public works (interventions and installations)", as these are the projects that, as widely shown in the literature, display the greatest delays and inefficiencies (for example Gondia et al., 2020; Egwim et al., 2021; He et al., 2023).

Starting from this battery of data, we created our database, which is made up of almost 19 thousand observations at the project scale and around 200 features. We collect endogenous information relating to individual cohesion projects (project-level variables), linked directly or indirectly to the information available with the database, and variables at the municipal level, coming from multiple official statistical sources[7], of an economic, socio-demographic, environmental, administrative and political nature. The data on institutional quality are at the provincial level. We control also at the territorial level, with macro-regional and regional dummies. Table A2 in the Appendix provides details of all the characteristics used in our estimates.

Our main outcome variable is the status of the project, represented by a dummy variable that takes value 0 if a project is completed on time, and value 1 if the project had any delay compared to the deadline declared by the implementing bodies. This variable derives from calculating the days of delay of each project, which is obtained, following Eq. 1, as the difference between the actual end date and the expected end

[6] A cohesion project is defined as concluded if it presents a liquidated financial progress (higher at 95%), as well as an execution phase completed within the monitoring times.

[7] The data on municipal incomes come from the public database of the Ministry of Economy and Finance on tax declarations. The geographical and demographic data are published by the Italian National Institute of Statistics. The data on local administrations are from the Ministry of the Interior. The variables on the municipal elections are obtained by manipulating the data of the Ministry of the Interior, made public on the Eligendo portal. The data on public spending by municipalities for technical offices comes from the data of the Open Municipalities Portal. The institutional quality indicators are those created by (Nifo and Vecchione, 2014).

date of each project:

$$d_j = e_j - p_j \tag{1}$$

where:

- d_j represents the delay in days for the j-th project;
- e_j is the actual end date of the j-th project;
- p_j is the expected end date of the j-th.

In addition to the estimates conducted with the dummy variable, as a second outcome variable we exploit the heterogeneity available from the calculation of the days of delay of each project (see Eq.1). This continuous variable takes on a positive value when the delay is at least one day and a negative one when there is an advance in the closing of the project. This additional estimate allows us to conduct further analyses that also take into account the specific differences in terms of delays between individual infrastructure interventions, thus providing a deeper understanding of the characteristics that condition these temporal differences.

3.2 Descriptive statistics

The tables 1, 2, 3 and Figure 1 show descriptive statistics for the distribution of infrastructure cohesion projects 2007-2013, focusing on delays for each Italian region and province and with respect to the type of implementing body.

Overall, these statistics suggest heterogeneous territorial effectiveness in the management of cohesion projects, as they show significant regional and provincial disparities in the rate of delayed projects.

Interestingly, Table 1 and Figure 1b, show that the share of late projects out of the total projects in each territorial area does not reflect the Italian regional disparities between North and South (Lagravinese, 2015; Guzzardi et al., 2023). For example, in Friuli-Venezia Giulia, which is a Northern Region, 80.12% of projects are delayed.[8] On the contrary, some southern regions such as Abruzzo and Molise have

[8] The small total number of projects (308) should be considered in this. In Sicily, however, the total number of projects is much higher, as is the number of late projects (equal to 56.14%).

much lower percentages of late projects: 6.92% and 8.31% respectively. The absence of a regional pattern that follows the dualism in Italian development is confirmed at a macro-regional and national level. The South and Centre have the lowest number of delayed projects out of the total in that area (less than 35%). The Islands display the larger percentages of projects with delays (nearly 55%). The North-West and the North-East (with 42% and 39%, respectively) show delays similar to the Italian average (38.18%). In general, such high percentages on a macro-territorial and national scale raise important questions about the efficiency of the Italian administrative structures and the project management methodologies adopted (Accetturo et al., 2014; Rodríguez-Pose and Garcilazo, 2015; Iammarino et al., 2019; Espinoza and Presbitero, 2022).

The Italian dualism reappears importantly in Figure 1a, when we consider the number of projects in each region and province, compared to the total number of Italian projects. In this case, the southern regions have more late projects than the northern regions (Baltrunaite et al., 2023).

Table 1: Regional distribution of infrastructure cohesion projects

Region	**Total Coeshion Project**	**Coeshion Projects delayed**	**Projects delayed on Total Project (%)**
Abruzzo	751	52	6.92
Basilicata	1160	396	34.13
Calabria	2053	663	32.29
Campania	5947	2044	34.37
Emilia-Romagna	263	78	29.65
Friuli-Venezia Giulia	312	250	80.12
Lazio	308	88	28.57
Liguria	879	377	42.88
Lombardy	564	123	21.8
Marche	566	187	33.03
Molise	397	33	8.31
Piedmont	880	494	56.14
Apulia	1985	850	42.82
Sardinia	762	266	34.9
Sicily	3216	1908	59.32
Tuscany	697	249	35.72
Trentino-Alto Adige	376	31	8.24
Umbria	589	175	29.71
Aosta Valley	29	17	58.62
Veneto	428	181	42.29
North-east	1379	540	39.16
North-west	2352	1011	42.98
Center	2160	699	32.36
South	12293	4038	32.85
Islands	3978	2174	54.65
N. Observation	**22162**	**8462**	**38.18**

Source: Authors' elaboration on Opencoesione.it

Table 2 and Figures 1c, 1d, which explore the total days of advance and delay accumulated in cohesion projects, offer a quantitative and more impactful picture and highlight the regional disparities between North and South (Baltrunaite et al., 2023). Many southern and island regions, such as Apulia, Calabria, Campania, and Sicily, display extremely high net negative balances of days of average delay (over 160 days of average delay in the South and more than 340 in the Islands). These findings point to potential systemic and structural problems in project management, which may include bureaucratic inefficiencies, poor planning and coordination, as well as possible deficiencies in transparency and accountability. On the other hand, some provinces in the North-East and the Center show a more balanced or even positive balance, such as Trentino-Alto Adige with an average of 13 days early, which suggests that the projects were completed faster than expected. Such outcomes may be due to better project phase management, effective resource allocation, and solid support infrastructure, which help prevent delays and promote efficient implementation.

Table 3 distinguishes delays in infrastructure cohesion projects among different public Authorities: Municipalities, Provinces, Regions, and State. A large number of late projects emerges at municipal and regional levels, especially in the South and on the Islands. This result confirms the data above, further indicating that the most significant challenges in managing cohesion projects are found at the levels of government closest to citizens, where the administrative capacity and resources involved may be limited (Rodríguez-Pose, 2013; Rodríguez-Pose and Garcilazo, 2015; Albanese et al., 2024). The inefficiency of local institutions in completing cohesion projects within the expected timescales represents a very relevant point for territorial development, as these institutions, during the 2007-2013 European programming cycle, appear to be those most involved in the phase of project implementation, as can be seen from Figure 2. Furthermore, these results suggest that some specific regional systems are more resilient and effective than others, probably due to better coordination, less bureaucracy or a greater availability of resources for timely implementation of projects.

Further descriptive analyses of the delays of infrastructure cohesion projects, implementing and programming bodies, intervention sectors and on the municipal, provincial and territorial characteristics, are reported in the Appendix 7.

Overall, the descriptive statistics presented provide a complex picture of delays in infrastructure cohesion projects in Italy, highlighting the need for a predictive in-depth study approach, to accurately study the characteristics that influence these inefficiencies. A more thorough understanding of the causes of delays and effective strategies to mitigate them are essential to improve the implementation of cohesion projects, fundamental for the development and balanced growth of the country.

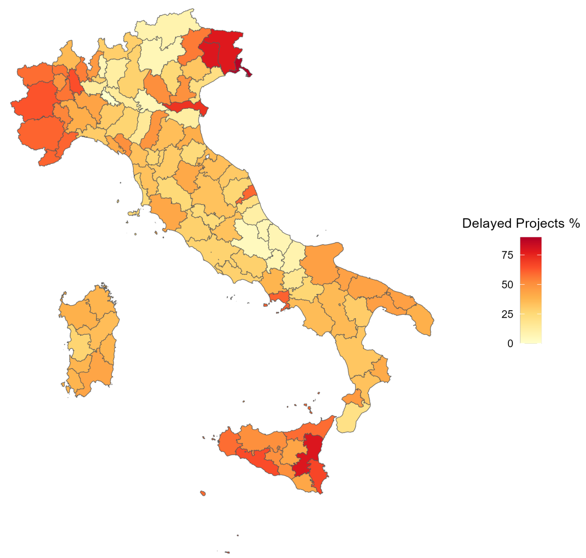
Table 2: Days of Projects Delay

Region	**Total Days Early**	**Total Days of Delay**	**Difference between Early and Late**	**Total Number in Coeshion Project**	**Average Differential Days by Region**
Abruzzo	6488	6253	-235	751	-0.31
Basilicata	18546	161550	143004	1160	123.28
Calabria	27578	595723	568145	2053	276.74
Campania	97751	965901	868150	5947	145.98
Emilia-Romagna	2669	11847	9178	263	34.90
Friuli-Venezia Giulia	7285	69745	62460	312	200.19
Lazio	1458	21444	19986	308	64.89
Liguria	37961	97100	59139	879	67.28
Lombardy	3956	38257	34301	564	60.82
Marche	30497	40473	9976	566	17.63
Molise	3153	7823	4670	397	11.76
Piedmont	9481	177914	168433	880	191.40
Apulia	66553	516917	450364	1985	226.88
Sardinia	33245	56685	23440	762	30.76
Sicily	28653	1368741	1340088	3216	416.69
Tuscany	12666	39011	26345	697	37.80
Trentino-Alto Adige	15385	10497	-4888	376	-13.00
Umbria	38292	44311	6019	589	10.22
Aosta Valley	76	3404	3328	29	114.76
Veneto	10780	27593	16813	428	39.28
North-east	36119	119682	83563	1379	60.60
North-west	51474	316675	265201	2352	112.76
Center	82913	145239	62326	2160	28.85
South	220069	2254167	2034098	12293	165.47
Islands	61898	1425426	1363528	3978	342.77
Total	**452473**	**4261189**	**3808716**	**22437**	**171.85**

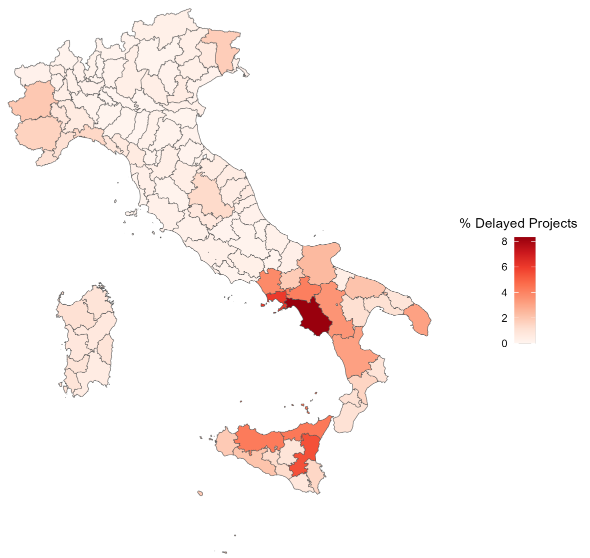
Source: Authors' elaboration on Opencoesione.it

Figure 1: Provincial distribution of delayed infrastructure cohesion projects and average days of delay at provincial level

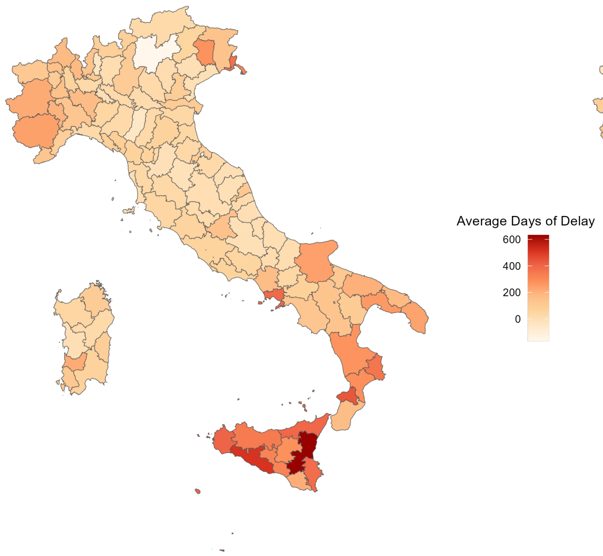
(a) Percentage of delayed cohesion projects per province on to the total cohesion projects in that province



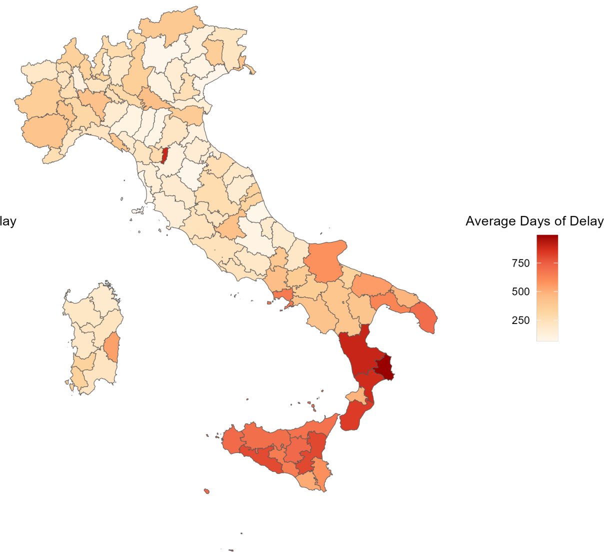
(b) Percentage of delayed cohesion projects per province on to the total cohesion projects in Italy



(c) Average days of delay of cohesion projects per province on the total cohesion projects in that province



(d) Average days of delay of cohesion projects per province calculated on the delayed cohesion projects in that province



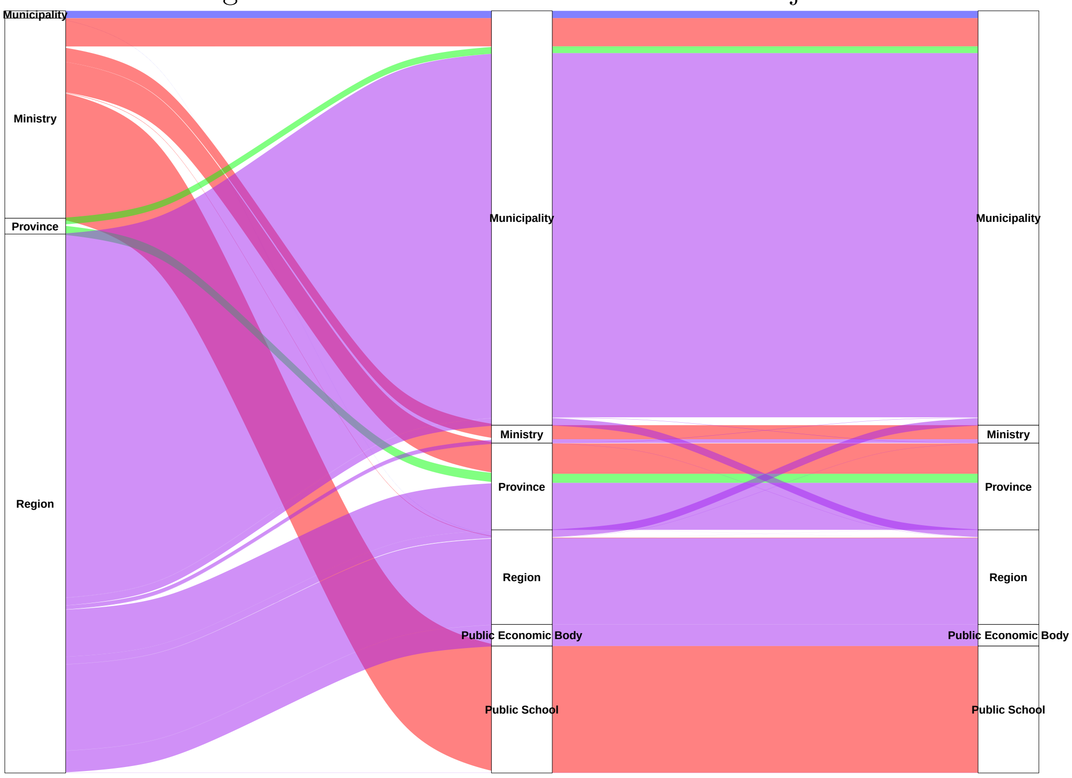
Source: Authors' elaboration on Opencoesione.it

Table 3: Type of Entities per Macro Area

Entities	Status	Marco Area					Total Status by Entities
		Northeast	North-west	Center	South	Islands	
Municipalies	**0**	262	763	833	3522	581	**5961**
	1	244	507	379	1046	703	**2879**
	Total	**506**	**1270**	**1212**	**4568**	**1284**	**8840**
Province	**0**	197	151	119	461	278	**1206**
	1	147	200	164	158	76	**745**
	Total	**344**	**351**	**283**	**619**	**354**	**1951**
Region	**0**	110	2	14	442	704	**1272**
	1	47	14	4	50	402	**517**
	Total	**157**	**16**	**18**	**492**	**1106**	**1789**
National	**0**	42	123	29	250	34	**478**
	1	3	10	13	171	53	**250**
	Total	**45**	**133**	**42**	**421**	**87**	**728**
Total Status by NUTS	**0**	**611**	**1039**	**995**	**4675**	**1597**	**8917**
	1	**441**	**731**	**560**	**1425**	**1234**	**4391**
	Total	**1052**	**1770**	**1555**	**6100**	**2831**	**13308**

Source: Authors' elaboration on Opencoesione.it

Figure 2: Flows of Entities in Cohesion Project



Source: Authors' elaboration on Opencoesione.it

Note: The first column refers to the programming bodies. The second to the implementing authorities. The third to the beneficiary bodies.

4 Empirical approach

In our work, the Prediction task is defined following the Eq. 2.

$$\{Features_i\} \xrightarrow{f(\cdot)} Prediction_i \tag{2}$$

In particular, for each infrastructural cohesion project i started at the time t, we want to obtain a function $f(\cdot)$ represented by a machine learning model that employs a series of features/predictors known at the time t, able to estimate the probability that an infrastructural cohesion project will have a delay at the end (T):

$$\{Pro, SocioEco, Demo, Ter, Geo, IQI, Pol, Tech\}_{it} \xrightarrow{f(\cdot)} Delay_{iT} \tag{3}$$

where:

- *Pro*: Project-level features of cohesion projects;
- *SocioEco*: Municipal Socioeconomic features;
- *Demo*: Municipal Demographic features;
- *Ter*: Regional and Macro Area features;
- *Geo*: Municipal Geographical features;
- *IQI*: Provincial Institutional Quality and Government Effectiveness features;
- *Pol*: Municipal Local Policy and Administrative features;
- *Tech*: Municipal Features of Public Spending for Technical Offices;
- *Delay*: Status of delay.

To have a predictive tool useful for policy, we only used features observed at the start date of the project (t) to predict the delay at T, so with $t < T$. This approach not only ensures a forward-looking predictive model but also avoids issues of endogeneity and reverse causality. In line with standard practices in prediction analyses with supervised machine learning, classification, and regression algorithms, conducted in the field of social sciences (see for example Antulov-Fantulin et al., 2021; Bloise et al., 2021; Carrieri et al., 2021), the dataset is randomly divided between training set and test set. In this study, we assign 70% of the data to the training set, while the remaining 30% constitutes the testing set.

As described in Section 3, to conduct our Machine Learning analyses, we use the dummy of late projects as a dependent variable. Furthermore, to further corroborate our estimates, we also implement ML algorithms with the continuous variable. For these analyses, the outcome variable is represented by the days of delay (see Eq. 1), which allows us to overcome data interpretation problems, in consideration of the high heterogeneity present in the days of delay between individual projects. As a robustness analysis, in consideration of the marked territorial and regional divisions present in Italy (Lagravinese, 2015; Guzzardi et al., 2023), especially in reference to the effectiveness of cohesion policies (Cerqua and Pellegrini, 2018; Crescenzi and Giua, 2020) and the projects implemented, we repeat the main ML estimates, with the binary variable of the projects delay, by splitting our dataset on a macro-regional scale, between North and South[9]. This is in order to detect socio-spatial differences regarding the most important features in predicting the delay of infrastructural cohesion policies.

The optimization of model parameters is derived from a repeated cross-validation process exclusively on the training set (k-fold = 10, with 5 repetitions).

The accuracy of the model in predicting delays in classification estimates is quantified using the Receiver Operating Characteristics (ROC) curve (Bradley, 1997; Fawcett, 2006), with the test set as the evaluation context. This binary classification delineates late projects as a positive class and early projects as a negative class. In particular, the ROC curve clarifies the competence of the classifier, relating the true positive rate (TPR) to the false positive rate (FPR) at various discrimination thresholds. An uninformative classifier results in an ROC curve that mirrors the line of randomness, producing an area under the curve (AUC) of 0.5, while a flawless classifier boasts an AUC of 1.0. The AUC therefore serves as a benchmark for the effectiveness of the classifier, with higher values indicating superior discriminatory power.

The ROC curve is defined as:

[9]The North with Emilia Romagna, Liguria, Lombardy, Marche, Piedmont, Tuscany, Umbria, Veneto. Friuli Venezia Giulia, Trentino Alto Adige, and Aosta Valley do not present any infrastructure cohesion projects, after the data manipulation operation, for the construction of the dataset. The South with Abruzzo, Apulia, Basilicata, Calabria, Campania, Lazio, Molise, Sicily, Sardinia.

$$TPR(y) = \frac{TP}{TP + FN} \tag{4}$$

where:

- TP: is the number of true positives;
- FN: is the number of false negatives.

$$FPR(x) = \frac{FP}{FP + TN} \tag{5}$$

where:

- FP: is the number of false positives;
- TN: is the number of true negatives.

In the context of regression, where continuous outcomes such as project delay days are expected, we adopt several performance metrics to evaluate the effectiveness of Machine Learning models. We use metrics, such as the Root Mean Square Error (RMSE), the Mean Squared Error (MSE), and the Mean Absolute Error (MAE), which provide a complete view of both the accuracy of the predictions and the fit of the model to the data. Thanks to these indicators, we can effectively compare the performance of the various models and select the most appropriate algorithm for the specific context of use (Chicco et al., 2021; Botchkarev, 2018).

For our estimates we exploit a battery of eight ML algorithms for classification and regression, as well as classic regression models for comparisons (logistic model and linear model)[10]:

- **Elastic Net (EN)**: this is a statistical regression that combines the penalties of Lasso (L1) and Ridge regression (L2), in order to optimize the selection and regularization of variables, in case of high dimensionality. This algorithm is, therefore, particularly effective in contexts where the variables present multicollinearity, offering a robust compromise between the variable selection capacity of the Lasso and the regularization of the Ridge, thus improving the stability and accuracy of the model (Zou and Hastie, 2005; De Mol et al., 2009; Cerulli, 2023);

[10]All ML algorithms and optimization processes were implemented using R software via the `caret` package (Kuhn, 2015; Kuhn et al., 2020).

- **Random Forest (RF)**: is a machine learning model based on decision trees. Each tree is built using random subsets of the features. It aims to improve the accuracy of predictions by reducing the risk of overfitting in the data. This technique, known as bootstrap aggregation, allows for greater generalization on unseen data by aggregating predictions from all trees to obtain the final result. Thanks to this mode of random selection of variables and the combination of different trees, a robust and effective solution is provided for complex classification and regression problems, especially with large datasets (Breiman, 2001; Liu et al., 2012; Cerulli, 2023);

- **Gradient Boosting Machines (GBM)**: this is an algorithm that iteratively improves its predictions, correcting the errors of previous models. This approach is structured as a basic decision tree model which increases its accuracy by adding new models sequentially, resolving the residual errors of the aggregate model up to that point. Each new model in the sequence is then trained to be as effective as possible at correcting the errors left by its predecessors. The algorithm uses techniques such as learning rate control and tree depth tuning to prevent overfitting, maintaining a balance between adapting to training data and generalizing to new data (Friedman, 2001; Natekin and Knoll, 2013);

- **Neural networks (NN)**: these are modelled on the functioning of the human brain. They are composed of nodes - or neurons - interconnected and organized in layers. Each node receives input from previous nodes on which it performs a weighted sum which it transmits as output to subsequent nodes. Neural networks adapt the weights of connections during the learning process, through algorithms such as gradient descent, which allow the error to be minimized. Their uses are very widespread in classification and regression studies in the presence of complex features, since they are characterized by a high learning capacity (Sutskever et al., 2014; Schmidhuber, 2015);

- **Support Vector Machine (SVM)**: its main objective is to find a hyperplane that optimally separates the different classes of data points. To facilitate this separation, this model transforms the data into a higher dimensional space. This transformation is particularly useful in scenarios where the number of features, and therefore data dimensions, exceeds the number of observations and samples. In these cases, the SVM allows for the effective management of complexity while avoiding overfitting. It is suitable for applications that require good generalization ability, even in the presence of a limited number of samples (Hearst et al., 1998; Steinwart and Christmann, 2008);

- **Decision Tree (DT)**: the decision tree is structured in a manner whereby each node is based on a specific characteristic and each branch represents the answer to that question. Trees are built by dividing the dataset into increasingly smaller subsets, based on statistical criteria that maximize the homogeneity of subsequent nodes. However, the DT is prone to overfitting, especially when the tree becomes very deep. Consequently, the choice of splitting criteria and tree depth parameters are crucial to balance accuracy and generalization (Charbuty and Abdulazeez, 2021; Cerulli, 2023);

- **Naive Bayes (NB)**: this is effective only in classification problems, when the features used are independent of each other. It estimates the probability that an element belongs to a certain class, through the study of the frequency and correlation between the observed characteristics. This approach is especially advantageous when working with large amounts of data, as it allows for quick and efficient estimates of the probability of class membership (Rish et al., 2001; Murphy et al., 2006);

- **K-Nearest Neighbors (KNN)** this is based on the principle of closeness in similarity, meaning that similar elements tend to be close to each other. Practically it identifies the distance between the k closest points in the training set and assigns to the object, among its neighbors, the most common class, or the average value in the case of regression. Performance is influenced by the choice of the parameter k which identifies the number of neighbors and by the distance metric used (Peterson, 2009; Kramer and Kramer, 2013).

5 Results

In this Section we set out our predictive results, i.e. model estimates, focusing on cohesion project delays. We focus on the ability to predict the dependent variables and the key role of the independent variables (features) used in the predictions, described in Section 3.1.

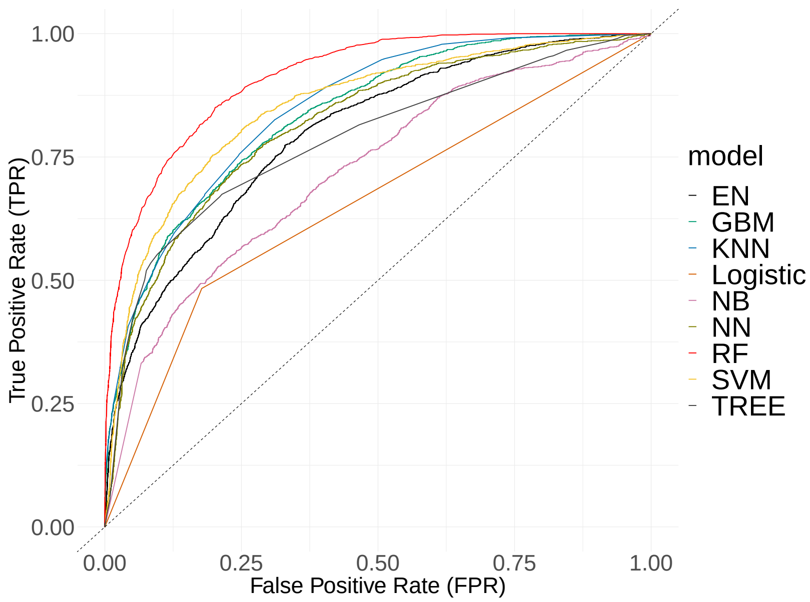
5.1 Classification task

5.1.1 Prediction performance

In this section, we evaluate the performance of the various machine learning models on the main outcome variable of our estimates (out-of-sample), represented by the dummy of delayed projects, and subsequently, we evaluate the optimization of the hyperparameters based on the training data.[11].

The AUC analysis, illustrated in the ROC curve graph (see Figure 3), highlights that the Random Forest (RF) model stands out as the best model for performance, compared to other machine learning algorithms, including regression logistics (Bradley, 1997; Fawcett, 2006). In particular, the AUC shows from a comparative perspective, that RF significantly outperforms all other machine learning models. In practice, this reveals the better performance of Random Forest compared to SVM, GBM, KNN, NB, NN, EN and Logistic Regression, regarding the prediction of delays in infrastructure cohesion projects in Italy. Furthermore, Figure 3 shows the heterogeneity among the various models.

Figure 3: ROC Curve



[11]For the models analyzed on the entire set of Italian infrastructure projects, the best hyperparameters identified are shown in Appendix A1

The results of the DeLong statistical test (DeLong et al., 1988) show that, for the RF model, the differences between the ROC curve and the remaining models are statistically significant (with p value < 0.05), in all comparisons between the algorithms. These results are in line with previous empirical studies (for example, Antulov-Fantulin et al., 2021; Carrieri et al., 2021), which reveal that Tree models are the most predictive algorithms on structured tasks, especially in classification contexts with binary response variables.

Table 4: DeLong's test

	Z	p-value
RF vs SVM	15.02	0.00
RF vs KNN	16.43	0.00
RF vs GBM	18.32	0.00
RF vs NN	19.31	0.00
RF vs EN	21.64	0.00
RF vs TREE	21.88	0.00
RF vs NB	27.63	0.00
RF vs Logistic	39.83	0.00

Note: Averages calculated across repetitions for 10 different random splits.

In general, the metrics reported in Table 5 show high performances of all the estimated algorithms. In particular, the high performance of these models in predicting late projects is confirmed by a No Information Rate lower than the accuracy level (Accuracy > Accuracy Null) and by the other reported metrics (Hand and Till, 2001; Tharwat, 2021). Furthermore, Cohen's Kappa coefficients show that the Random Forest emerges as the model with the best value (0.615), a figure which is in the range of 'good' strength of agreement about the prediction reliability (Altman, 1990; Landis and Koch, 1977).

Overall, the results confirm that, in the presence of a range of different predictive models, it is possible to make efficiency estimates, associated with the prediction of project delays, despite the characteristics and complexity of the data. The accuracy of the models, therefore, highlights their effectiveness and reliability in predicting delays in Italian infrastructure cohesion projects. In particular, the RF represents a highly efficient performing algorithm in predicting project delays and, therefore, a valuable resource for policy makers in the field of territorial cohesion and projects.

In general, as already explained in several scientific works and for different sectors, for example, Kleinberg et al. (2015), Antulov-Fantulin et al. (2021), and Sansone and Zhu (2023), a machine learning approach can have some potential as an Early Warning System (EWS). In our case, these results demonstrate how it is indeed possible to effectively predict delays in Italian infrastructure cohesion projects, thus answering our main research question affirmatively.

Table 5: Models' performances

Model	RF	SVM	TREE	KNN	NN	GBM	EN	NB	Logit
Accuracy	0.829	0.795	0.779	0.773	0.767	0.767	0.744	0.700	0.700
Kappa	0.615	0.541	0.485	0.488	0.474	0.451	0.398	0.326	0.320
AccuracyLower	0.819	0.784	0.768	0.762	0.756	0.756	0.732	0.685	0.688
AccuracyUpper	0.839	0.806	0.790	0.784	0.778	0.779	0.755	0.710	0.713
AccuracyPValue	0.000	0.000	0.000	0.000	0.00	0.000	0.000	0.000	0.000
Sensitivity	0.675	0.645	0.535	0.597	0.865	0.496	0.473	0.524	0.483
Specificity	0.916	0.879	0.916	0.872	0.591	0.919	0.896	0.794	0.822
PosPredValue	0.818	0.749	0.781	0.723	0.791	0.775	0.717	0.588	0.604
NegPredValue	0.834	0.816	0.779	0.795	0.710	0.765	0.752	0.749	0.740
Precision	0.675	0.645	0.535	0.597	0.865	0.496	0.473	0.524	0.483
Recall	0.818	0.749	0.781	0.723	0.791	0.775	0.717	0.588	0.603
F1	0.740	0.693	0.635	0.654	0.826	0.605	0.570	0.554	0.537
Prevalence	0.359	0.359	0.359	0.359	0.641	0.359	0.359	0.359	0.358
DetectionRate	0.242	0.232	0.192	0.214	0.554	0.178	0.170	0.188	0.173
DetectionPrevalence	0.296	0.309	0.246	0.296	0.701	0.230	0.236	0.320	0.287
BalancedAccuracy	0.795	0.762	0.725	0.735	0.728	0.708	0.684	0.659	0.653
NoInformationRate	0.641	0.641	0.641	0.641	0.641	0.641	0.641	0.641	0.641

Note: These values are estimated on the confusion matrix. which shows a cross-tabulation of the observed and predicted classes, generating the predicted classes based on the typical 50% cutoff for the probabilities (Kuhn et al., 2020). Averages were calculated across repetitions for 10 different random splits.

5.1.2 Features' Importance

In this Section, the most important predictors of the delays of the infrastructure cohesion projects of the 2007-2013 cycle are presented (Figure 4). In this analysis we only use the RF which, as illustrated in the previous Section (5.1.1), is the algorithm that predicts delays of infrastructural cohesion projects best of all, in terms of distance of the bisector of the ROC curves and compared to the DeLong test results (see Figure 3 and Table 4).

To make the most of the data available for this analysis, following established literature (e.g. Cawley and Talbot, 2010; Krstajic et al., 2014), we trained the Random Forest model on the entire dataset. This methodological approach ensured that every aspect of our database was exploited effectively for estimating what matters most in predicting delays.

In terms of the importance of features (see Figure 4), schools, both considered as implementing bodies and beneficiaries of cohesion projects, emerged as the most relevant features of the delay. This reflects the relevance of the management and internal organization of these educational institutions, which affect the level of institution and training of the youngest (Bryk and Schneider, 2002). The efficiency with which these entities organize and manage cohesion funds and, therefore, coordinate project execution activities, can determine a significant acceleration or slowing down of the overall progress of the projects. An administrative support system with a task force at the Ministerial level may help.

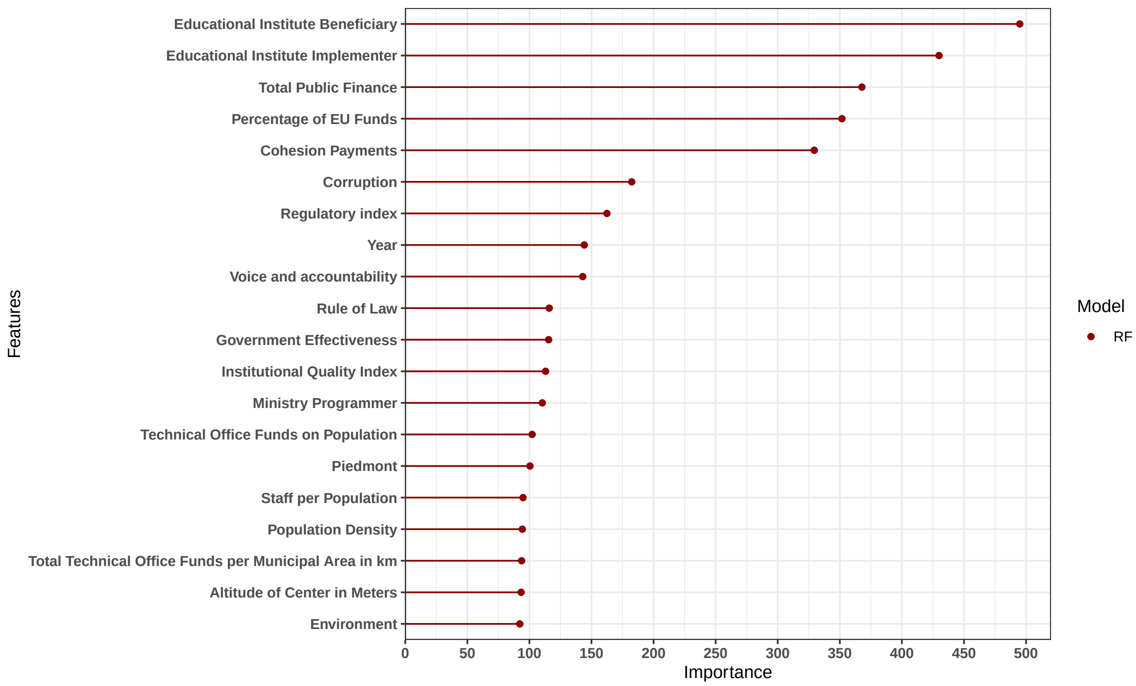
The resources assigned to cohesion projects, such as the amount of public funding, the percentage of EU Funds, and the total cohesion payments obtained, are fundamental features that influence a continuous project workflow. In this sense, delays in payments or mismanagement of finances can create significant bottlenecks, highlighting the need for robust and transparent financial systems to ensure that funds are available and used promptly (Tosun, 2014; Bachtler et al., 2017) .

The Institutional Quality Indicator (IQI) and the composite sub-indicators estimated by Nifo and Vecchione (2014), such as Corruption, Voice and Accountability, Regulatory Index and Government Effectiveness, have a significant impact on the efficiency and timing of infrastructure cohesion projects, albeit at lower levels of intensity. For example, corruption can divert resources and slow down projects (D'Alpaos et al., 2013). Likewise, the strength of regulatory policies and the integrity of government practices are critical for minimizing delays and improving project effectiveness (Espinoza and Presbitero, 2022).

Public spending on technical offices also appears to be a factor that has a significant impact on the delay, which is also related to the geographical and territorial characteristics of the municipalities, particularly with reference to the altitude and surface area and, therefore, to urbanization (Albanese et al., 2024). In particular, the relevance of the altitude of the municipality highlights that certain territorial contexts can represent obstacles that complicate the management of projects, or that can facilitate their execution. These factors are particularly problematic in mountain contexts, where the complexity of logistics systems can significantly influence implementation times (Davoudi and Strange, 2009).

Excluding the dummy relating to projects located in the Piedmont Region, the spatial variables (regional dummies and macro regional project dummies) in this analysis of projects delays, are not among the most important features, as do the political economy variables relating to administrative elections and the characteristics of local administrators.

Figure 4: Importance of features with the best model (RF) in case of classification



Note: The RF is trained on the entire database.

Overall, these features' importances confirm what has been demonstrated by the growing literature on cohesion policies, which highlights how the quality and characteristics of local institutions can undermine or improve the effectiveness and absorption of cohesion funds, for example, Becker et al. (2013), Rodríguez-Pose and Garcilazo (2015), Albanese et al. (2024).

In summary, the interaction of these variables creates a complex network that requires careful management of cohesion policies in order to optimize project implementation times. Adopting targeted strategies to address each of these variables could not only reduce delays but also increase transparency and effectiveness in the use of cohesion funds, completing projects on time.

5.1.3 Partial Dependence

In this paragraph, we study the direction and intensity of the most important characteristics regarding project delays. To highlight the signs of the individual features, we use partial dependence plots (PDP), which allow us to study the relationship between variables and delays in projects, also allowing us to detect varying dynamics and time trends.

The PDP relating to the characteristics of educational institutions as implementing bodies of Italian infrastructural cohesion projects (in Figure 5a) shows a positive relationship with delay. Specifically, we can observe how the probability of delay in cohesion projects increases with the involvement of educational institutions as implementing bodies. The direction and trend of this relationship is linked to the very nature of this feature (dummy variable which is equal to 1 if the project is implemented by an educational institution or 0 if it is not). As a result, the efficiency and ability to manage complex projects worsens when educational institutions are directly involved in project implementation, contributing to increased delays. This evidence can be attributed to a poor level of efficiency of local institutions in managing projects not connected to their educational purposes (Accetturo et al., 2014; Incaltarau et al., 2020).

Figure 5b highlights, however, the relationship between total public funding for cohesion projects and their delays, revealing a relationship connected to the amount of resources assigned. Initially, in fact, increases in funding lead to a rapid increase in project delays, up to a certain resource threshold. With further increases in funding, the delay in projects remains essentially constant. This trend shows that a greater allocation of financial resources beyond that threshold does not produce significant changes in the project completion times.

When analyzing the percentage of cohesion on total funding, an opposite behavior pattern is observed compared to total funding. In fact, Figure 5c highlights a positive initial trend, where an increase in the percentage of cohesion funding corresponds to a decrease in project delays. However, as in the case of total funding, beyond a certain threshold, further increases in cohesion funding do not lead to significant improvements, suggesting that delays remain largely unchanged. This shows again that, although the presence of EU cohesion funds can increase the speed of execution, there is a limit beyond which increases in cohesion funding are no longer effective in further reducing delays.

The latter evidence can be determined by potential saturations in the capacity to manage funds or by other critical issues linked to the inefficiencies of cohesion projects, as highlighted in the works of Incaltarau et al. (2020), Espinoza and Presbitero (2022), which suggests that an overall review of the allocation of funds is required, to place greater focus on the optimization of processes and management.

Other local characteristics, such as institutional quality indicators, present non-linear trends. In particular, the corruption indicator at the provincial level (see the PDP in Figure 5d) shows a positive relationship with delays[12]. In fact, a decrease in corruption leads to a significant reduction in delays D'Alpaos et al. (2013). Similar and more linear trends, albeit with an opposite direction in terms of the sign of the indicators (when the index is low the institutional quality is low), are also recorded in other sub-indicators of institutional quality, such as Voice and Accountability (Figure 6a) and Rule of Law (Figure 6b). In fact, they show a significant reduction in delays, with the progressive increase in institutional quality, up to a certain threshold, beyond which the delay begins to increase, although remaining at values lower than the initial ones.

In general, these results show how several features of institutional quality are very

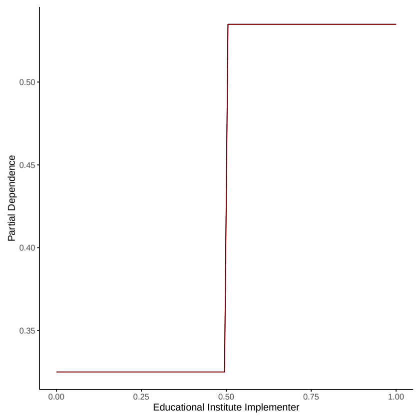
[12]High levels of corruption correspond to low values of the indicator.

relevant factors in determining the delay of cohesion projects, especially in the low spectrum of institutional quality, confirming the studies of Accetturo et al. (2014) and Espinoza and Presbitero (2022).

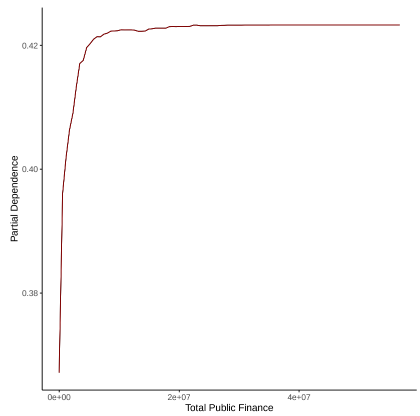
At a geographical level, the municipal altitude is a very relevant feature indicative of the delay (see Figure 6c). Specifically, a positive relationship is found between delays and altitude in meters. In fact, municipal altitude determines an almost linear increase in the delay of cohesion projects, up to 100 meters (maximum delay), and then reduces slightly to remain stable in the more mountainous municipalities (Davoudi and Strange, 2009; Albanese et al., 2024).

Figure 5: PDP: Overview of Features Importance 1)

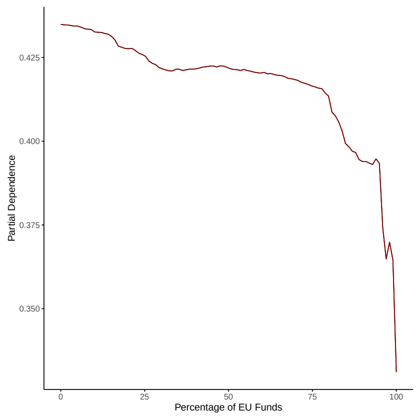
(a) Educational Institution Implementing Body



(b) Total Public Finance



(c) Percentages of EU Funds



(d) Provincial Corruption Indicator

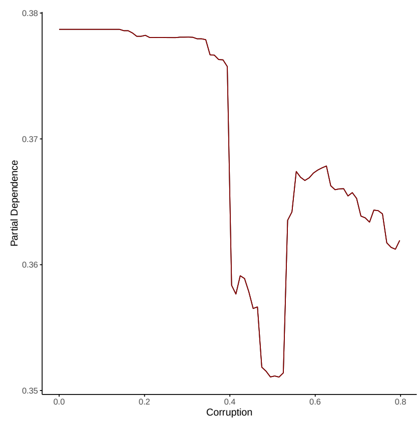
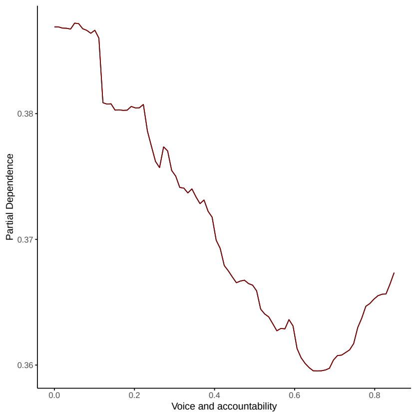
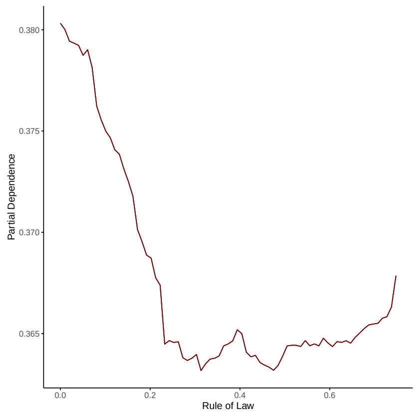


Figure 6: PDP: Overview of Features Importance 2)

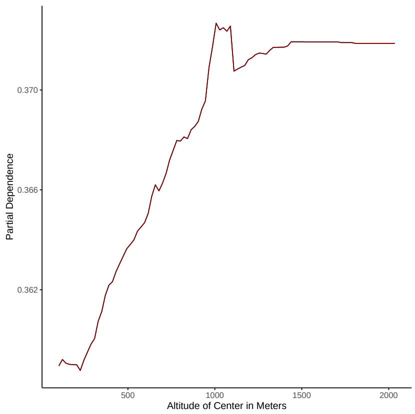
(a) Provincial Voice and Accountability Indicator



(b) Provincial Rule of Law Indicator



(c) Altitude of the municipalities



5.2 Regression task

In this Section, we present our second predictive regression task, conducted on the continuous dependent variable represented by the number of days of delay of each project.

Also in this case, Random Forest is the best algorithm for predicting delays of infrastructure projects since it has excellent performance in terms of metrics of prediction (see the values of the metrics expressed in Table 6.). Indeed, the RF displays the lowest mean absolute errors MAE[13] (on average 65.15 days of delay of the infrastructure cohesion projects), compared to the other algorithms. The predictive superiority of the RF is also supported by the lowest values marked by the MSE and RSME (with an RSME of 146.54, in contrast to the values of the other algorithms between 220 and 400). Furthermore, the NN shows limited performances, indicating large errors in the prediction, making it unsuitable for regression estimates.

Table 6: Performance of Regression Models in Predicting Project Days of Delay

Model	**RMSE**	**MSE**	**MAE**
RF	146.54	21474.31	65.15
KNN	221.44	49036.99	121.26
GBM	237.1	56215.46	130.43
SVM	239.98	57590.39	123.83
TREE	250.92	62961.77	138.98
OLS	251.34	63171.90	142.62
EN	254.39	64714.77	139.74
NN	401.34	161076.87	190.74

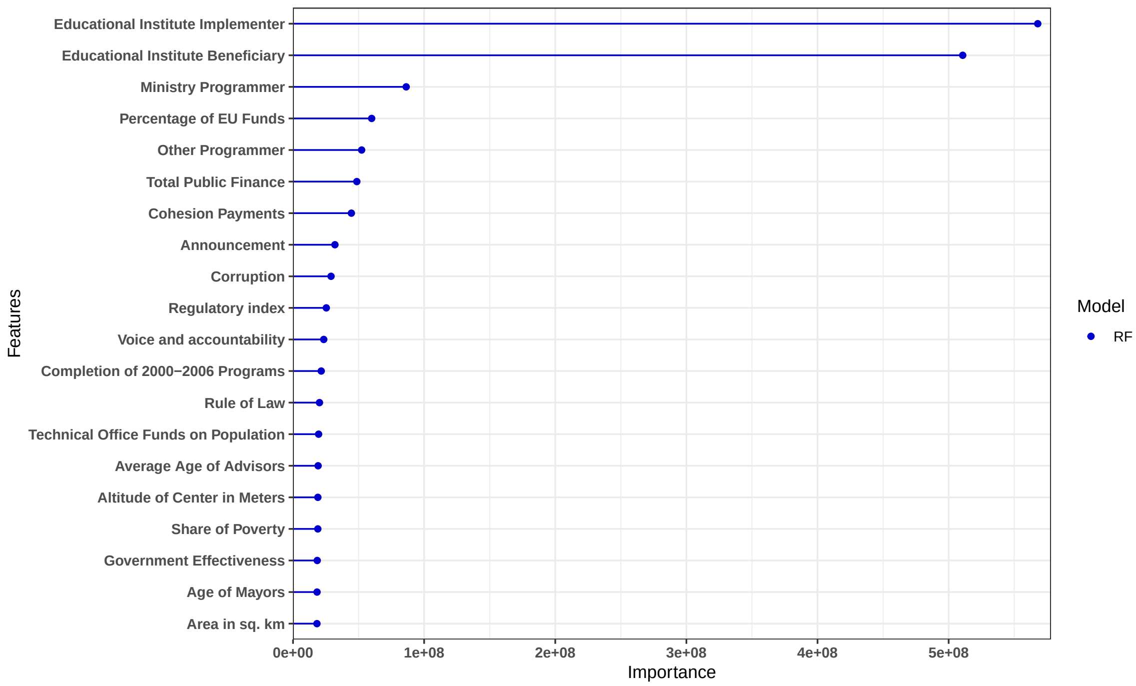
Note: RMSE stands for Root Square Mean Error; MSE stands for Mean squared error; MAE stands for Mean Absolute Error.

Figure 7 reports the features' importance in the case of regression. The importance of the features with a continuous variable follows that recorded in the classification analysis with a binary variable (see Figure 4), with a correlation between the features' importance of these two analyses that is very positive and statistically significant (of

[13]Which is a measure of accuracy used to evaluate regression models when the model predictions deviate on average from the actual values.

0.734, as can be observed from the coefficients described in the note[14]). However, in this case a significant heterogeneity among the variables is highlighted. In particular, even in this second analysis, schools appear as the implementing and beneficiary bodies that have the greatest and largest impact on the delays. Although of lesser importance, financial factors, such as the total public funding of each project, the percentage of funding from cohesion funds, institutional and local administration quality variables, public spending per capita for technical offices and geographical variations, are the features that most impact on project timing.

Figure 7: Importance of features with the best model (RF) in case of Regression



Note: The RF model is trained on the entire database, which presents as an outcome variable the continuous variable relating to the days of delay of each cohesion project considered.

The Partial Dependence Plots relating to the main features' importance for the predictive analysis conducted using the continuous variable of the days of delay (pre-

[14]The correlation between feature importances obtained from regression and classification models is 0.734, indicating a strong positive association between the two analyses. The 95% confidence interval for this correlation is between 0.6610 and 0.7925, confirming the robustness of the result. The associated p-value is extremely low, 4.87×10^{-34}, suggesting that the observed correlation is statistically significant and not causal.

sented in the Appendix 7, in Figures A2a, A2b and A2c) show very similar trends to those observed in the classification analysis (reported in Section 5.1.3).

In general, this second predictive analysis with ML algorithms, conducted on the variable relating to the number of days late for each project, shows very similar trends to the main analysis conducted on the late projects, confirming the validity of the estimates developed. The predictive differences obtained derive, in fact, mainly from the nature of the continuous variable, which displays a significant heterogeneity in terms of days of delay.

6 Territorial heterogeneity

In this Section, we present a territorial analysis, conducted on a macro-regional scale. We do this in consideration of the marked Italian regional inequalities in socio-economic development, capable of impacting the delays of cohesion projects. Specifically, we divided our dataset between projects carried out in the municipalities of the Centre-North and the South.

The graphs in Figure 8 and 9 depict the AUC of the ROC for the municipalities for the two macro-regions. Differences are found in the out-of-simple performances between the two spatial analyses, compared to the various machine algorithms learning and the identified hyperparameters.

Specifically, the performances of the trained models, considering only the infrastructure projects implemented in the Southern regions, are similar to those obtained for the total Italian dataset. Less performing results are obtained when considering only the projects implemented in the Centre-North.[15] However, in both these analyses, the Random Forest (RF) model stands out as the best Machine Learning algorithm, able to prevent delays in infrastructure cohesion projects in Italy, confirming the predictive goodness of the main estimate.

Territorial heterogeneity in delays between Northern and Southern Italy also emerges in the importance of features. Indeed, Figures 10 and 11 highlight significant

[15] Consider that there is a lower concentration of cohesion projects in these regions.

Figure 8: ROC Curve for the South

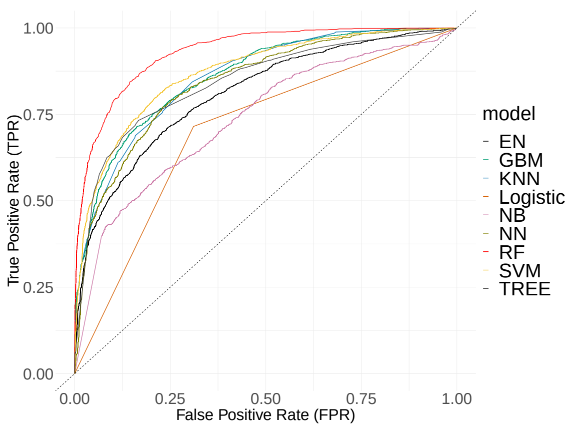
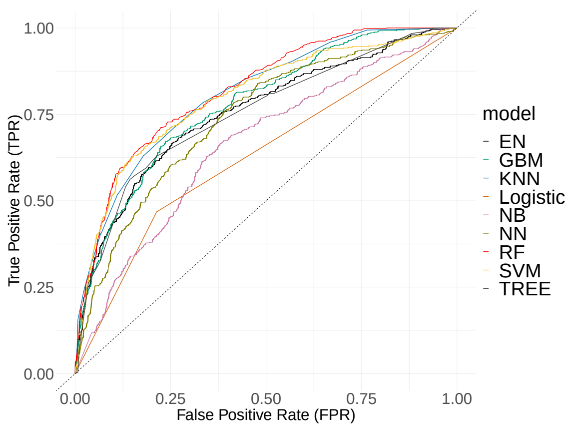


Figure 9: ROC Curve for the North



differences both in the types of features identified and in their relevance. In the analysis relating to projects in the Centre-North, the importance of 5 individual features is comparable. On the contrary, in the South a greater variation in the importance attributed to the different predictors is observed, with education institutions emerging overwhelmingly as the main predictor.

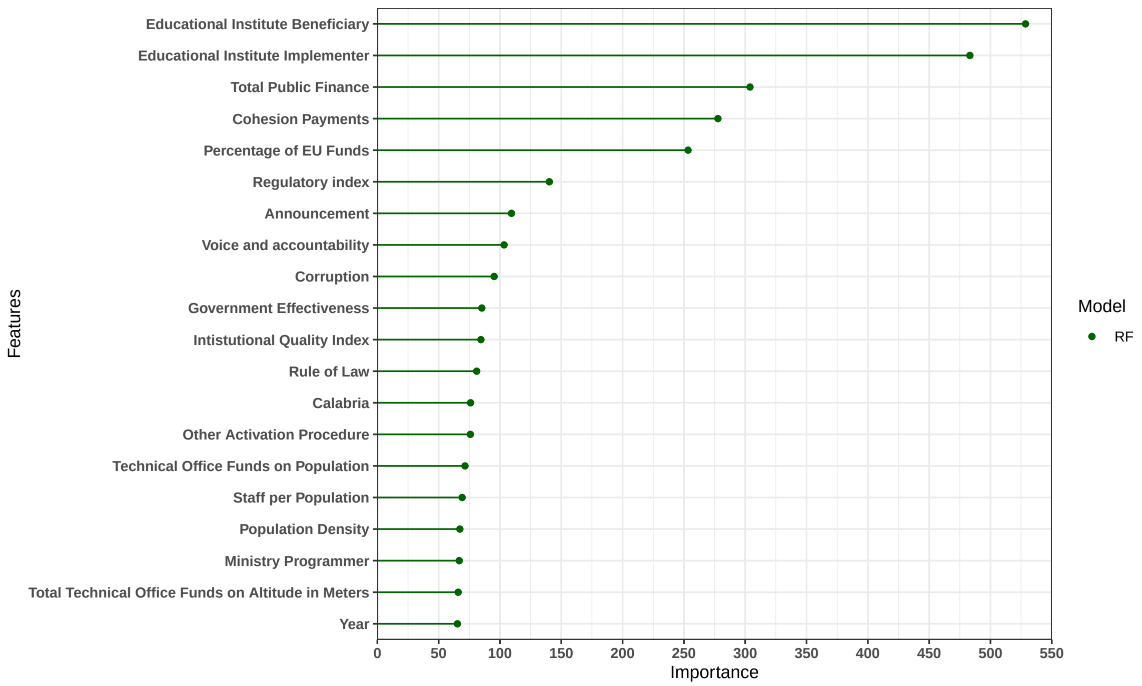
However, despite the differences that emerged in the importance of the predictors, the correlation between the feature importance for the estimates between projects implemented in the South and the North presents a high correlation coefficient (0.797)[16].

In general, the features importance in the South follows closely the results obtained in the main analysis in Section 5.1, with schools identified as the most important features for the completion of cohesion projects, followed by financial and management characteristics of projects and by provincial and municipal socio-institutional factors. Schools are less relevant for the conclusion of cohesion projects in the Centre-North. However, the importance of the educational sector in the timing of project implementation also emerges in projects implemented in the Center-North, where the characteristics relating to the educational institution are among the most important, together with those of a financial nature.

[16]The correlation between feature importance is very high 0.797 is statistically significant (p-value: 1.79×10^{-36} with a 95% confidence interval ranging from 0.733 to 0.848.

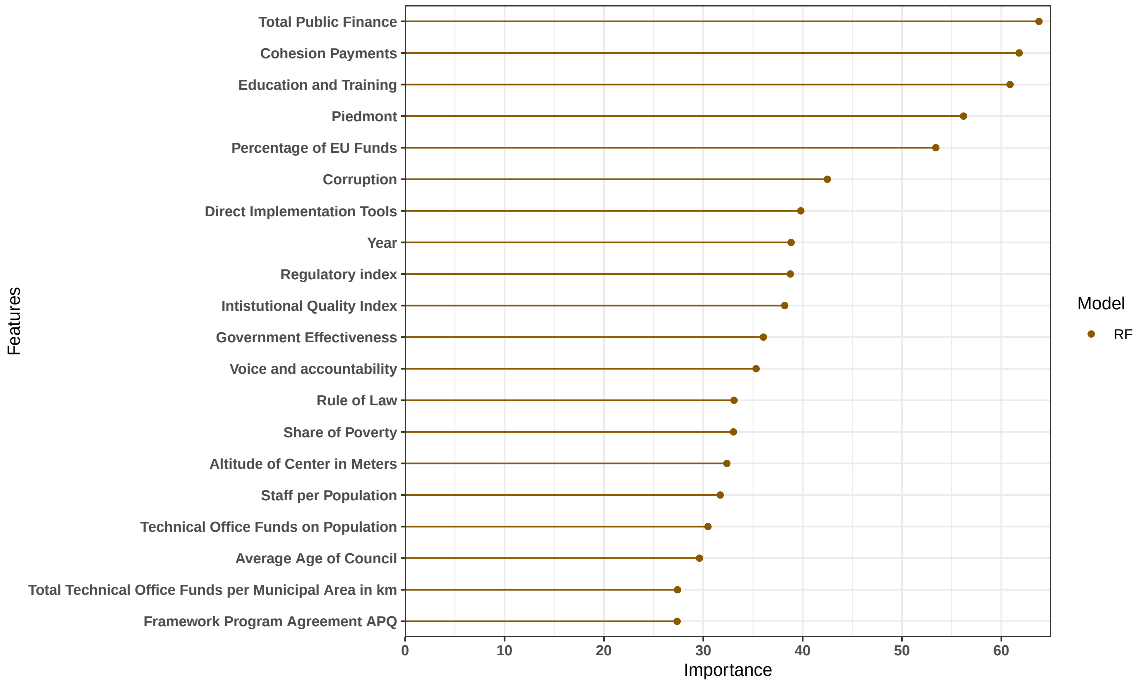
These results suggest that the huge regional variations in educational attainments (Coco et al., 2020) may be linked to the management, and standardization frameworks of infrastructural cohesion projects in Southern regions. Indeed, delays and the correlated inefficiency may be the main explanation for the clear disadvantage of southern schools in terms of infrastructures (Bucci et al., 2021). Both attainments and infrastructure in turn may be correlated to the managing quality of institutions.

Figure 10: Importance of features with the best model (RF) for the South



Note: The RF model is trained on the entire database, which presents as an outcome variable the continuous variable relating to the days of delay of each cohesion project considered. .

Figure 11: Importance of features with the best model (RF) for the North



Note: The RF model is trained on the entire database, which presents as an outcome variable the continuous variable relating to the days of delay of each cohesion project considered.

7 Conclusions

This study presents an analysis of cohesion project delays, using an advanced approach based on machine learning techniques to identify and interpret the key factors for prediction. Through the use of eight different algorithms, applied to a uniquely rich database at the territorial level, we were able to isolate the most influential variables, offering a more detailed view than traditional analytical approaches. The exclusive use of features observed at the start of each project ensures that the model avoids endogeneity, enhancing its robustness and making the findings particularly relevant for designing effective policies to address project delays.

The results demonstrate that project delays can be predicted using a set of features observed at the start date of the project. Both economic factors and institutional features play significant roles in these predictions. The most important predictor is the involvement as an implementer or beneficiary institution of a school, particularly in the less developed South. Cohesion funds are particularly concentrated in the south, making this result of tantamount importance for policy. It also explains why infrastructure endowments in this sector are heterogeneous.

The use of European funds (as opposed to national co-financing) is a good predictor of the timely execution of projects. This suggests an 'advantage of tying one's hand' and offers an argument in favor of keeping cohesion policy at the EU level. It also has important consequences on the most appropriate way to implement the next budget cycle of EU funds. The importance of the EU Cohesion funds should remain intact and the discipline element, both financial and thematic, should be reinforced.

Institutional quality, in several dimensions including corruption, also emerges as a crucial determinant, suggesting that sometimes even small improvements in these areas could facilitate greater time-efficiency of cohesion programs. Our work in general highlights the importance of considering a variety of factors, including less tangible ones, such as the quality of local governance. Our models revealed that some local dynamics, administrative capacities and the specific demographic structure of a municipality can significantly alter project expected delays.

The implications of these findings are relevant for policy makers. Interventions aimed at improving administrative capacity and strengthening institutional quality could improve the effectiveness of the use of funds. This implies that personalized and focused support strategies may be necessary to address the specific barriers that impede the effectiveness of cohesion programs, particularly in the education sector. Given the relevance of this sector, a centralization of procedures based on a national task force may be worth considering.

Appendix

Table A1: Hyperparameters in case of Classification

Model	IperParameters	Description	Implications
Elastic Net (EN)	Model type: $\alpha = 1$ Penalization parameter: $\lambda = 0.0003$	α at 1 indicates a pure Lasso model, favoring variable selection. α controls the degree of regularization to prevent overfitting.	A lower λ value means less regularization, which can be optimal for complex data where noise reduction is not a primary concern.
Random Forest (RF)	Number of variables per split: $mtry = 98$	*mtry* determines how many variables are examined at each split. A high value allows for greater diversity in the tree models.	A high *mtry* can improve accuracy but increase the risk of overfitting if the dataset is not sufficiently large or varied.
Gradient Boosting Machine (GBM)	Total number of trees: $n.trees = 150$ Interaction depth: $interaction.depth = 3$ Learning rate: $shrinkage = 0.1$ Minimum observations per node: $n.minobsinnode = 10$	$n.trees$ is the number of sequential trees. Interaction.depth manages the depth of the tree. *Shrinkage* reduces the contribution of each tree. *n.minobsinnode* sets the minimum observations per node.	$n.trees = 150$ with a shallow depth and shrinkage ensure gradual learning while avoiding overfitting the training data.
Neural Network (NN)	Number of units in hidden layer: $Size = 5$ Weight decay: $Decay = 0.1$	$Size$ indicates the number of neurons in the hidden layer which increases the model's ability to learn complex relationships. *Decay* helps to reduce overfitting by penalizing larger weights.	5 neurons offer good complexity without being too computationally demanding, while a decay of 0.1 prevents the model from overfitting the training data.
K-Nearest Neighbors (KNN)	Number of neighbors considered: $k = 15$	k represents the number of nearest neighbors considered for classification or regression.	A larger k provides more stabilized decisions that are less sensitive to fluctuations in the input data.
Support Vector Machine (SVM)	Gaussian kernel width: $\sigma = 0.0043$; Error penalization parameter: $c = 8$	σ influences the width of the Gaussian RBF kernel, c is the error penalization parameter.	A smaller leads to a finer decision boundary, while a high c severely punishes misclassifications, emphasizing correct classification.
Decision Tree (TREE)	Tree pruning complexity parameter: $cp = 0.0026$	*cp* is the complexity parameter that regulates the growth of the tree.	A lower *cp* allows more complex trees, potentially capturing better the relationships in the data, but at the risk of overfitting.
Naive Bayes (NB)	Laplace correction index: $Laplace = 0$ Use of kernel approach for density: $UseKernel$ = FALSE Smoothing parameter: Adjust = 1	*Laplace* and $UseKernel$ affect probability handling, *Adjust* modifies the kernel parameters for continuous estimates.	The configuration avoids the use of kernel, preferring a simpler and more direct approach to probabilities, useful in situations with good data distribution.

Table A2: Descriptive statistics for the Features

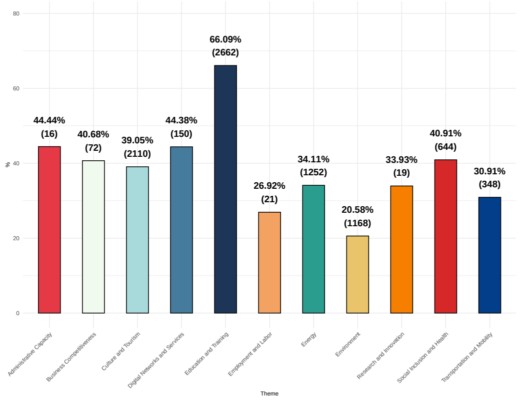
Outcome variables	**Mean**	**SD**	**Median**	**Min**	**Max**	**N.Obs.**	**Variable**	**Surch**
- For Classification								
Dummy Delay	0.38	0.49	0	0	1	22162	Dummy	Opencoesione.it
- For Regression								
Days Delay	171.86	395.02	0	-2372	2754	22162	Continues	Opencoesione.it
Endugenus on the Opencoesione dataset								
Business Competitiveness	0.01	0.09	0	0	1	22162	Dummy	Opencoesione.it
Circular	0.01	0.09	0	0	1	22162	Dummy	Opencoesione.it
Administrative capacity	0	0.04	0	0	1	22162	Dummy	Opencoesione.it
Announcement	0.26	0.44	0	0	1	22162	Dummy	Opencoesione.it
Cohesion Payments	327754.61	1098403.28	99548.86	0	51312827.85	22162	Continues	Opencoesione.it
Company Beneficiary	0.02	0.15	0	0	1	22162	Dummy	Opencoesione.it
Company Implementer	0.02	0.15	0	0	1	22162	Dummy	Opencoesione.it
Completion of 2000-2006 Programs	0.01	0.1	0	0	1	22162	Dummy	Opencoesione.it
Culture and Tourism	0.24	0.43	0	0	1	22162	Dummy	Opencoesione.it
Digital Networks and Services	0.02	0.12	0	0	1	22162	Dummy	Opencoesione.it
Direct Identification in the Program	0.08	0.28	0	0	1	22162	Dummy	Opencoesione.it
Direct Implementation Tools	0.26	0.44	0	0	1	22162	Dummy	Opencoesione.it
Education and Training	0.18	0.39	0	0	1	22162	Dummy	Opencoesione.it
Education Training and Supports for the Labor Market	0	0.03	0	0	1	22162	Dummy	Opencoesione.it
Educational Institute Beneficiary	0.13	0.33	0	0	1	22162	Dummy	Opencoesione.it
Educational Institute Implementer	0.13	0.33	0	0	1	22162	Dummy	Opencoesione.it
EM	0.01	0.11	0	0	1	22162	Dummy	Opencoesione.it
Employment and Labor	0	0.06	0	0	1	22162	Dummy	Opencoesione.it
Energy	0.17	0.37	0	0	1	22162	Dummy	Opencoesione.it
Energy Sector Infrastructure	0.07	0.25	0	0	1	22162	Dummy	Opencoesione.it
Enrolled from Abroad	322.91	1575.02	27	0	37262	22162	Dummy	Opencoesione.it
Enrolled from Inside	793.13	2797.65	110	0	65213	22162	Dummy	Opencoesione.it
Environment	0.26	0.44	0	0	1	22162	Dummy	Opencoesione.it
Environmental Infrastructure and Water Resources	0.36	0.48	0	0	1	22162	Dummy	Opencoesione.it
ERDF (European Regional Development Fund)	0.59	0.49	1	0	1	22162	Dummy	Opencoesione.it
ESF (European Social Fund)	0	0.02	0	0	1	22162	Dummy	Opencoesione.it
Extraordinary Maintenance Restoration and Renovation	0.63	0.48	1	0	1	22162	Dummy	Opencoesione.it
Framework Program Agreement APQ	0.14	0.35	0	0	1	22162	Dummy	Opencoesione.it
Infrastructure for Telecommunications and Information Technologies	0.01	0.11	0	0	1	22162	Dummy	Opencoesione.it
Infrastructure for the Equipment of Productive Areas	0.01	0.1	0	0	1	22162	Dummy	Opencoesione.it
Ministry Beneficiary	0.02	0.13	0	0	1	22162	Dummy	Opencoesione.it
Ministry Implementer	0.02	0.13	0	0	1	22162	Dummy	Opencoesione.it
Ministry Programmer	0.21	0.41	0	0	1	22162	Dummy	Opencoesione.it
Mountain Community Beneficiary	0.17	0.38	0	0	1	22162	Dummy	Opencoesione.it
Mountain Community Implementer	0.17	0.38	0	0	1	22162	Dummy	Opencoesione.it
Municipal Beneficiary	0.39	0.49	0	0	1	22162	Dummy	Opencoesione.it
Municipal Implementer	0.4	0.49	0	0	1	22162	Dummy	Opencoesione.it
Municipal Programmer	0.01	0.09	0	0	1	22162	Dummy	Opencoesione.it
Negotiation Procedure	0.22	0.42	0	0	1	22162	Dummy	Opencoesione.it
New Realization	0.17	0.37	0	0	1	22162	Dummy	Opencoesione.it
Ordinary Maintenance	0.07	0.25	0	0	1	22162	Dummy	Opencoesione.it
Other Activation Procedure	0.27	0.44	0	0	1	22162	Dummy	Opencoesione.it
Other Beneficiary	0.45	0.5	0	0	1	22162	Dummy	Opencoesione.it
Other different from ERDF ESF	0.41	0.49	0	0	1	22162	Dummy	Opencoesione.it
Other Implementer	0.44	0.5	0	0	1	22162	Dummy	Opencoesione.it
Other Intervention	0.14	0.34	0	0	1	22162	Dummy	Opencoesione.it
Other Programmer	0.22	0.41	0	0	1	22162	Dummy	Opencoesione.it
Percentage EU Funds	88.65	20.7	99.01	0	100	22162	Continues	Opencoesione.it
Private Sector Beneficiary	0.05	0.21	0	0	1	22162	Dummy	Opencoesione.it
Private Sector Implementer	0.05	0.21	0	0	1	22162	Dummy	Opencoesione.it
Private Sector Programmer	0	0.05	0	0	1	22162	Dummy	Opencoesione.it
Projects originated from other implementing tools QSN	0	0.07	0	0	1	22162	Dummy	Opencoesione.it
Provincial Beneficiary	0.09	0.28	0	0	1	22162	Dummy	Opencoesione.it
Provincial Implementer	0.09	0.28	0	0	1	22162	Dummy	Opencoesione.it
Provincial Programmer	0.02	0.15	0	0	1	22162	Dummy	Opencoesione.it
Public Entity Beneficiary	0.02	0.15	0	0	1	22162	Dummy	Opencoesione.it
Public Entity Implementer	0.02	0.15	0	0	1	22162	Dummy	Opencoesione.it
Public Notice	0.09	0.29	0	0	1	22162	Dummy	Opencoesione.it
Public Sector Beneficiary	0.94	0.23	1	0	1	22162	Dummy	Opencoesione.it

Public Sector Implementer	0.95	0.22	1	0	1	22162	Dummy	Opencoesione.it
Public Sector Programmer	1	0.05	1	0	1	22162	Dummy	Opencoesione.it
Public-Private Sector Beneficiary	0	0.03	0	0	1	22162	Dummy	Opencoesione.it
Public-Private Sector Implementer	0	0.03	0	0	1	22162	Dummy	Opencoesione.it
Regional Beneficiary	0.08	0.27	0	0	1	22162	Dummy	Opencoesione.it
Regional Implementer	0.08	0.27	0	0	1	22162	Dummy	Opencoesione.it
Regional Programmer	0.75	0.43	1	0	1	22162	Dummy	Opencoesione.it
Research and Innovation	0	0.05	0	0	1	22162	Dummy	Opencoesione.it
Research Technological Development and Innovation	0	0.04	0	0	1	22162	Dummy	Opencoesione.it
Retroactive Identification Procedure	0.05	0.21	0	0	1	22162	Dummy	Opencoesione.it
Social Inclusion and Health	0.07	0.26	0	0	1	22162	Dummy	Opencoesione.it
Social Infrastructure	0.46	0.5	0	0	1	22162	Dummy	Opencoesione.it
Total Beneficiaries	1.02	0.35	1	1	28	21997	Categorical	Opencoesione.it
Total Implementers	1.02	0.35	1	1	28	22148	Categorical	Opencoesione.it
Total Programmers	1	0.07	1	1	4	22162	Categorical	Opencoesione.it
Total Public Finance	394078.27	1621302.39	123286.04	171.45	156797676	22162	Dummy	Opencoesione.it
Total Rounds	1.14	0.35	1	1	2	19388	Dummy	Opencoesione.it
Transport Infrastructure	0.09	0.29	0	0	1	22162	Dummy	Opencoesione.it
Transportation and Mobility	0.05	0.22	0	0	1	22162	Dummy	Opencoesione.it
Variation of Structural Funds related to 2007-2013 programming	0.58	0.49	1	0	1	22162	Dummy	Opencoesione.it
Variation of the Cohesion Action Plan - Own Resources	0.13	0.33	0	0	1	22162	Dummy	Opencoesione.it
Variation of the Fund for Development and Cohesion	0.3	0.46	0	0	1	22162	Dummy	Opencoesione.it
Works, Systems And Equipment for Production Activities, Research and Social Enterprise	0.01	0.09	0	0	1	22162	Dummy	Opencoesione.it
Year	2011	2.07	2011	2007	2015	22162	Categorical	Opencoesione.it
Socio-Economic								
Per capita Income	14622.92	3453.19	13909.9	7103.54	51403.21	22162	Continues	Ministry of the Economy
Share of Poverty	0.42	0.11	0.43	0.03	0.76	22157	Continues	Ministry of the Economy
Demographic								
Births	394.66	1414.98	43	0	25958	22162	Continues	ISTAT
Cancellations for Abroad	108.79	507.36	15	0	15301	22162	Continues	ISTAT
Cancellations for Domestic	915.42	3218.09	124	0	47195	22162	Continues	ISTAT
Deaths	449.43	1591.52	54	0	27590	22162	Continues	ISTAT
Population at the beginning of the period	43566.62	151175.09	5186	47	2724565	22162	Continues	ISTAT
End-of-Period Population	43603.68	151537.2	5172	45	2760303	22162	Continues	ISTAT
Population between 50 and 100 k	0.07	0.26	0	0	1	22162	Continues	ISTAT
Population over 100 k	0.07	0.26	0	0	1	22162	Continues	ISTAT
Population Density	529.17	1215.16	133.68	0.86	12234.78	22113	Continues	ISTAT
Territorial dummies								
Center of Italy	0.1	0.3	0	0	1	22162	Dummy	ISTAT
Northern Italy	0.17	0.37	0	0	1	22162	Dummy	ISTAT
Southern Italy	0.73	0.44	1	0	1	22162	Dummy	ISTAT
Abruzzo	0.03	0.18	0	0	1	22162	Dummy	ISTAT
Basilicata	0.05	0.22	0	0	1	22162	Dummy	ISTAT
Calabria	0.09	0.29	0	0	1	22162	Dummy	ISTAT
Campania	0.27	0.44	0	0	1	22162	Dummy	ISTAT
Lazio	0.01	0.12	0	0	1	22162	Dummy	ISTAT
Liguria	0.04	0.2	0	0	1	22162	Dummy	ISTAT
Lombardy	0.03	0.16	0	0	1	22162	Dummy	ISTAT
Marche	0.03	0.16	0	0	1	22162	Dummy	ISTAT
Molise	0.02	0.13	0	0	1	22162	Dummy	ISTAT
Piedmont	0.04	0.2	0	0	1	22162	Dummy	ISTAT
Apulian	0.09	0.29	0	0	1	22162	Dummy	ISTAT
Sardinia	0.03	0.18	0	0	1	22162	Dummy	ISTAT
Sicily	0.15	0.35	0	0	1	22162	Dummy	ISTAT
Umbria	0.03	0.16	0	0	1	22162	Dummy	ISTAT
Veneto	0.02	0.14	0	0	1	22162	Dummy	ISTAT
Tuscany	0.03	0.17	0	0	1	22162	Dummy	ISTAT
Geographical								
Altimetric Zone	3.05	1.44	3	1	5	22162	Categorical	ISTAT
Altitude of Center in Meters	342.98	276.66	303	0	2035	22162	Continues	ISTAT
Area in sq. km	75.29	94.5	40.43	0.12	1287.39	22113	Continues	ISTAT
Coastal Municipality	0.2	0.4	0	0	1	22162	Dummy	ISTAT
Coastal Zones	1.52	1.19	1	0	3	22162	Dummy	ISTAT
Island Municipality	0.13	0.33	0	0	1	22162	Dummy	ISTAT
Institutional Quality and Government Effectiveness								
Intistutional Quality Index	0.4	0.24	0.38	0	1	22162	Dummy	Opencoesione.it
Corruption	0.7	0.22	0.71	0	1	22162	Dummy	Ministry of Interior
Government Effectiveness	0.32	0.15	0.31	0	1	22162	Continues	Ministry of Interior
Voice and accountability	0.43	0.22	0.4	0	1	22162	Continues	Ministry of Interior
Rule of Law	0.4	0.22	0.35	0	1	22162	Continues	Ministry of Interior
Regulatory index	0.41	0.22	0.39	0	1	22162	Continues	Ministry of Interior

Local Policy and Administrative								
Age of Mayors	65.13	9.66	65	32	108	22153	Dummy	Ministry of Interior
Average Age of Advisors	58.82	4.61	59	39.8	76.42	22162	Dummy	Ministry of Interior
Average Age of Council	60.84	5.81	61	37	90.67	22162	Dummy	Ministry of Interior
Centre	0.01	0.09	0	0	1	19359	Dummy	Ministry of Interior
Centre-left	0.16	0.36	0	0	1	19359	Dummy	Ministry of Interior
Centre-right	0.16	0.36	0	0	1	19359	Dummy	Ministry of Interior
Civic List Text Analysis	0.19	0.39	0	0	1	22162	Dummy	Ministry of Interior
Civic Lists	0.68	0.47	1	0	1	19359	Dummy	Ministry of Interior
Coalition List Votes in I Round out of Voters	52.61	25.81	52.11	11.09	2952.79	19359	Dummy	Ministry of Interior
Coalition List Votes in I Rouout out of Electors	37.85	16.53	37.16	6.76	1698.73	19359	Dummy	Ministry of Interior
Election Year	2008.95	2.9	2009	2003	2015	19388	Categorical	Ministry of Interior
Five Stars MoViment	0	0.03	0	0	1	19388	Dummy	Ministry of Interior
Lists Over 60 Electors	0.03	0.16	0	0	1	19359	Dummy	Ministry of Interior
Lists Over 60 Voters	0.23	0.42	0	0	1	19359	Dummy	Ministry of Interior
Lists Over 80 Electors	0	0.04	0	0	1	19359	Dummy	Ministry of Interior
Lists Over 80 Voters	0.04	0.2	0	0	1	19359	Dummy	Ministry of Interior
Mayor Gender	0.07	0.26	0	0	1	22162	Dummy	Ministry of Interior
Mayor Over 60 Electors	0.02	0.14	0	0	1	19359	Dummy	Ministry of Interior
Mayor Over 60 Voters	0.22	0.42	0	0	1	19359	Dummy	Ministry of Interior
Mayor Over 80 Voters	0.04	0.19	0	0	1	19359	Dummy	Ministry of Interior
Mayors' Education Level	0.62	0.48	1	0	1	22162	Dummy	Ministry of Interior
Mayors Over 50	0.93	0.26	1	0	1	22153	Dummy	Ministry of Interior
Mayors Over 70	0.31	0.46	0	0	1	22153	Dummy	Ministry of Interior
Mayors Under 40	0	0.05	0	0	1	22153	Dummy	Ministry of Interior
Number Assessors	5.33	2.77	4	0	17	22039	Continues	Ministry of Interior
Number Councilors	18.45	10.5	16	6	60	22039	Continues	Ministry of Interior
Reelected Mayor in I Term after 2007	0.8	0.4	1	0	1	19388	Dummy	Ministry of Interior
Reelected Mayor in II Term after 2007	0.2	0.4	0	0	1	19388	Dummy	Ministry of Interior
Victorious in I Round Without Runoff	0.13	0.33	0	0	1	19388	Dummy	Ministry of Interior
Victorious in I Round	0.86	0.35	1	0	1	19388	Dummy	Ministry of Interior
Victorious in II Round	0.14	0.35	0	0	1	19388	Dummy	Ministry of Interior
Voters	34142.4	127152.1	4576	70	2359119	19359	Continues	Ministry of Interior
Votes	23878.32	85474.54	3340	49	1729287	19359	Continues	Ministry of Interior
Votes for Head Candidate in I Round Electors	37.98	9.77	37.53	10.09	86.17	19359	Dummy	Ministry of Interior
Votes for Head Candidate in I Round out of Voters	52.75	12.57	52.07	16.48	98.32	19359	Dummy	Ministry of Interior
Votes for Head Candidate in I/II Round out of Electors	38.6	9.17	38	12.95	86.17	19359	Dummy	Ministry of Interior
Votes for Head Candidate in I/II Round out of Voters	53.6	11.61	52.2	19.2	98.32	19359	Dummy	Ministry of Interior
Technical Office Funds								
Purchase of Goods and/or Raw Materials	16639.51	68808.69	2987.5	0	2832897	22162	Continues	Openbilanci.it
Purchase of Goods and/or Raw Materials on Percentage EU Funds	0.36	6.26	0.01	0	566.16	21728	Continues	Openbilanci.it
Purchase of Goods and/or Raw Materials on Total Financing	0.3	5.31	0.01	0	427.45	22162	Continues	Openbilanci.it
Purchase of Goods and/or Raw Materials per Altitude in Meters	434.84	6938.8	11.87	0	809620.8	22150	Continues	Openbilanci.it
Purchase of Goods and/or Raw Materials per Municipal Area in km	699.42	5374.97	62.1	0	490100.73	22116	Continues	Openbilanci.it
Purchase of Goods and/or Raw Materials per Per Capita Income	1.2	5.18	0.21	0	240.35	22162	Continues	Openbilanci.it
Purchase of Goods and/or Raw Materials per Population	8.06	75.98	0.38	0	6283.99	22162	Continues	Openbilanci.it
Purchase of Goods and/or Raw Materials per Share of Poverty	42208.51	175750.22	7175.08	0	5934025.11	22162	Continues	Openbilanci.it
Staff	617247.1	1807829.6	150866	0	35814598.2	22162	Continues	Openbilanci.it
Staff per Altitude in Meters	16052.43	138656.3	687.16	0	10091560	22154	Continues	Openbilanci.it
Staff per Population	280.59	1809.98	29.02	0	111535.25	22162	Continues	Openbilanci.it
Staff per Per Capita Income	44.43	134.47	10.75	0	2938.59	22162	Continues	Openbilanci.it
Staff per Share of Poverty	1568028.1	4708987.8	375110.1	0	109870066	22162	Continues	Openbilanci.it
Staff per Municipal Area in km	24164.17	107277.48	3884.38	0	4853272.98	22110	Continues	Openbilanci.it
Staff on Total Financing	14.36	458.29	0.7	0	61659.81	22162	Continues	Openbilanci.it
Staff on Total Cohesion Payments	18.3	597.71	0.79	0	78128.55	21724	Continues	Openbilanci.it
Service Performance	171640.2	1825730.9	15036	0	84394852.6	22162	Continues	Openbilanci.it
Service Performance per Altitude in Meters	5006.47	168139.3	63.42	0	23266960	22154	Continues	Openbilanci.it
Service Performance per Population	74.46	1155.92	2.19	0	93262.77	22162	Continues	Openbilanci.it
Service Performance per Per Capita Income	12.21	133.98	1.05	0	7222.25	22162	Continues	Openbilanci.it
Service Performance per Share of Poverty	440105.58	4772727.56	36345.39	0	270645995.47	22162	Continues	Openbilanci.it
Service Performance per Municipal Area in km	6872.76	118792.8	330.97	0	15043650	22110	Continues	Openbilanci.it
Service Performance on Total Financing	4.24	250.13	0.06	0	36588.74	22162	Continues	Openbilanci.it
Service Performance on Total Cohesion Payments	5.24	263.95	0.07	0	36977.66	21724	Continues	Openbilanci.it
Technical Office Funds on Population	426.03	3567.7	37.29	0	288676.48	22162	Continues	Openbilanci.it
Total Technical Office Funds	919352.1	3864733.5	190036	0	160484955.6	22162	Continues	Openbilanci.it
Total Technical Office Funds on Altitude in Meters	21181.37	167975.68	912.66	0	8321063.1	22150	Continues	Openbilanci.it
Total Technical Office Funds on Total Financing	15.18	279.04	0.91	0	37003.87	22162	Continues	Openbilanci.it
Total Technical Office Funds per Municipal Area in km	38153.5	276035.75	5070.61	0	29885466.59	22116	Continues	Openbilanci.it
Total Technical Office Funds per Per Capita Income	66.31	283.08	13.53	0	12589.45	22162	Continues	Openbilanci.it
Total Technical Office Funds per Share of Poverty	2336397.7	10183916.9	481043.6	0	495347415.3	22162	Continues	Openbilanci.it
Total Technical Office Funds on Cohesion Payments	17.71	306.89	1.02	0	39430.16	21728	Continues	Openbilanci.it

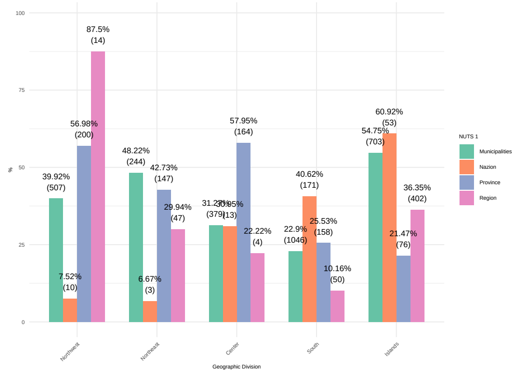
Figure A1: Others Descriptive Statistics

(a) Infrastructure Projects Delayed by Theme



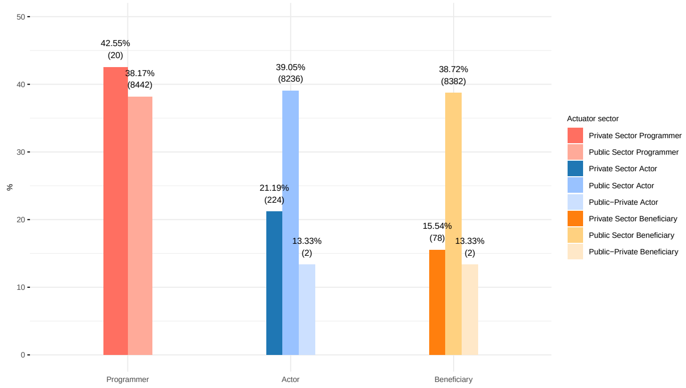
Source: Authors' elaboration on Opencoesione.it

(b) Infrastructure Projects Delayed by NUTS and Implementing Body



Source: Authors' elaboration on Opencoesione.it

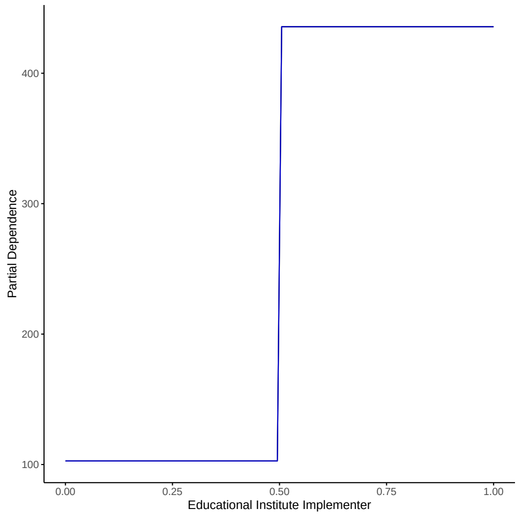
(c) Infrastructure Projects Delayed by Sectors and Execution Phases



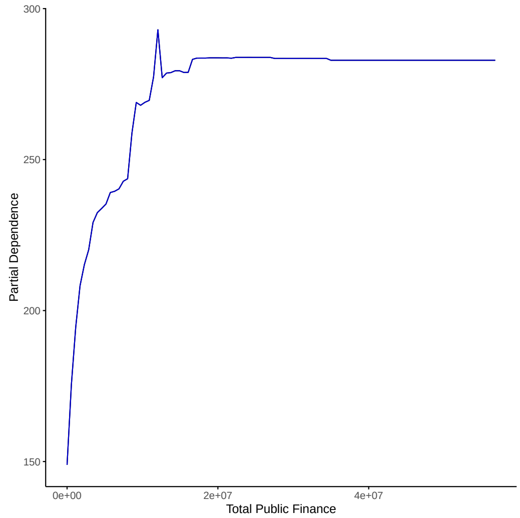
Source: Authors' elaboration on Opencoesione.it

Figure A2: PDP: Overview of Features Importance in Case of Regression

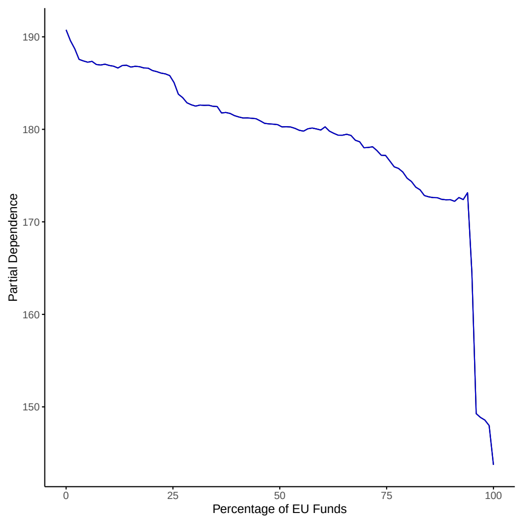
(a) PDP: Educational Institution Implementing Body in case of Regression



(b) PDP: Total Public Finance in case of Regression



(c) PDP: Percentages of EU funds in case of Regression



References

Accetturo, A. and G. De Blasio (2012). Policies for local development: An evaluation of italy's "patti territoriali". *Regional Science and Urban Economics 42*(1-2), 15–26.

Accetturo, A., G. De Blasio, and L. Ricci (2014). A tale of an unwanted outcome: Transfers and local endowments of trust and cooperation. *Journal of Economic Behavior & Organization 102*, 74–89.

Agenzia per la Coesione Territoriale (2018). Rapporto sui tempi di attuazione delle opere pubbliche.

Albanese, G., G. Barone, and G. De Blasio (2022). Populist voting and losers' discontent: Does redistribution matter? *European Economic Review 141*, 104000.

Albanese, G., V. Carrieri, A. Rita Ferrara, and M. M. Speziali (2024). Beyond gdp: assessing the impact of the 2007–13 eu cohesion policy. *Regional Studies*, 1–21.

Altman, D. G. (1990). *Practical statistics for medical research.* Chapman and Hall/CRC.

Andini, M., E. Ciani, G. de Blasio, A. D'Ignazio, and V. Salvestrini (2018). Targeting with machine learning: An application to a tax rebate program in italy. *Journal of Economic Behavior & Organization 156*, 86–102.

Antulov-Fantulin, N., R. Lagravinese, and G. Resce (2021). Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior & Organization 183*, 681–699.

Aschauer, D. A. (1989). Is public expenditure productive? *Journal of monetary economics 23*(2), 177–200.

Bachtler, J., P. Berkowitz, S. Hardy, and T. Muravska (2017). *EU Cohesion Policy.* Taylor & Francis.

Baltrunaite, A., T. Orlando, and G. Rovigatti (2023). The making of public works in italy: Institutional features and regional characteristics. *Italian Economic Journal*, 1–38.

Barone, G., A. D'Ignazio, G. De Blasio, and P. Naticchioni (2016). Mr. rossi, mr. hu and politics. the role of immigration in shaping natives' voting behavior. *Journal of Public Economics 136*, 1–13.

Becker, S. O., P. H. Egger, and M. Von Ehrlich (2010). Going nuts: The effect of eu structural funds on regional performance. *Journal of Public Economics 94*(9-10), 578–590.

Becker, S. O., P. H. Egger, and M. Von Ehrlich (2012). Too much of a good thing? on the growth effects of the eu's regional policy. *European Economic Review 56*(4), 648–668.

Becker, S. O., P. H. Egger, and M. Von Ehrlich (2013). Absorptive capacity and the growth and investment effects of regional transfers: A regression discontinuity design with heterogeneous treatment effects. *American Economic Journal: Economic Policy 5*(4), 29–77.

Becker, S. O., P. H. Egger, and M. Von Ehrlich (2018). Effects of eu regional policy: 1989-2013. *Regional Science and Urban Economics 69*, 143–152.

Bloise, F., P. Brunori, and P. Piraino (2021). Estimating intergenerational income mobility on sub-optimal data: a machine learning approach. *The Journal of Economic Inequality 19*(4), 643–665.

Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006*.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition 30*(7), 1145–1159.

Breiman, L. (2001). Random forests. *Machine learning 45*, 5–32.

Bryk, A. and B. Schneider (2002). *Trust in schools: A core resource for improvement.* Russell Sage Foundation.

Bucci, M., G. Ivaldi, G. Messina, L. Moller, and E. Gennari (2021). Infrastructure gaps in italy: A case-by-case measurement. *Bank of Italy Occasional Paper* (635).

Calderón, C., E. Moral-Benito, and L. Servén (2015). Is infrastructure capital productive? a dynamic heterogeneous approach. *Journal of Applied Econometrics 30*(2), 177–198.

Carrieri, V., R. Lagravinese, and G. Resce (2021). Predicting vaccine hesitancy from area-level indicators: A machine learning approach. *Health Economics 30*(12), 3248–3256.

Cawley, G. C. and N. L. Talbot (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research 11*, 2079–2107.

Cerqua, A., R. Di Stefano, M. Letta, and S. Miccoli (2021). Local mortality estimates during the covid-19 pandemic in italy. *Journal of Population Economics 34*(4), 1189–1217.

Cerqua, A. and G. Pellegrini (2018). Are we spending too much to grow? the case of structural funds. *Journal of Regional Science 58*(3), 535–563.

Cerulli, G. (2023). *Fundamentals of Supervised Machine Learning: With Applications in Python, R, and Stata*. Springer Nature.

Charbuty, B. and A. Abdulazeez (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends 2*(01), 20–28.

Chaurey, R. and D. T. Le (2022). Infrastructure maintenance and rural economic activity: Evidence from india. *Journal of Public Economics 214*, 104725.

Chicco, D., M. J. Warrens, and G. Jurman (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science 7*, e623.

Christensen, P., P. Francisco, E. Myers, H. Shao, and M. Souza (2024). Energy efficiency can deliver for climate policy: Evidence from machine learning-based targeting. *Journal of Public Economics 234*, 105098.

Coco, G. and R. Lagravinese (2021). Cohesion policy and public investment in the eu. In F. Cerniglia, F. Saraceno, and A. Watt (Eds.), *The great reset: 2021 European public investment outlook*, pp. 203. Open Book Publishers.

Coco, G., R. Lagravinese, and G. Resce (2020). Beyond the weights: a multicriteria approach to evaluate inequality in education. *The Journal of Economic Inequality 18*, 469–489.

Coppola, G., S. Destefanis, G. Marinuzzi, and W. Tortorella (2018). European union and nationally based cohesion policies in the italian regions. *Regional Studies*.

Coviello, D. and M. Mariniello (2014). Publicity requirements in public procurement: Evidence from a regression discontinuity design. *Journal of Public Economics 109*, 76–100.

Crescenzi, R. and M. Giua (2020). One or many cohesion policies of the european union? on the differential economic impacts of cohesion policy across member states. *Regional Studies 54*(1), 10–20.

Davoudi, S. and I. Strange (2009). *Conceptions of space and place in strategic spatial planning*. Routledge London.

De Blasio, G., A. D'Ignazio, and M. Letta (2022). Gotham city. predicting 'corrupted'municipalities with machine learning. *Technological Forecasting and Social Change 184*, 122016.

De Mol, C., E. De Vito, and L. Rosasco (2009). Elastic-net regularization in learning theory. *Journal of Complexity 25*(2), 201–230.

Delogu, M., R. Lagravinese, D. Paolini, and G. Resce (2024). Predicting dropout from higher education: Evidence from italy. *Economic Modelling 130*, 106583.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837–845.

D'Alpaos, C., M. Moretto, P. Valbonesi, and S. Vergalli (2013). Time overruns as opportunistic behavior in public procurement. *Journal of Economics 110*, 25–43.

Egwim, C. N., H. Alaka, L. O. Toriola-Coker, H. Balogun, and F. Sunmola (2021). Applied artificial intelligence for predicting construction projects delay. *Machine Learning with Applications 6*, 100166.

Espinoza, R. and A. F. Presbitero (2022). Delays in public investment projects. *International Economics 172*, 297–310.

European Commission (2022). *8th Cohesion Report: Cohesion in Europe towards 2050*. Luxembourg: Publications Office of the European Union, 2022.

Fabra, N., A. Lacuesta, and M. Souza (2022). The implicit cost of carbon abatement during the covid-19 pandemic. *European Economic Review 147*, 104165.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters 27*(8), 861–874.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Girma, S. and D. Paton (2024). Using double-debiased machine learning to estimate the impact of covid-19 vaccination on mortality and staff absences in elderly care homes. *European Economic Review 170*, 104882.

Gondia, A., A. Siam, W. El-Dakhakhni, and A. H. Nassar (2020). Machine learning algorithms for construction projects delay risk prediction. *Journal of Construction Engineering and Management 146*(1), 04019085.

Guzzardi, D., E. Palagi, A. Roventini, and A. Santoro (2023). Reconstructing income inequality in italy: New evidence and tax system implications from distributional national accounts. *Journal of the European Economic Association*.

Hand, D. J. and R. J. Till (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning 45*, 171–186.

He, C., A. Milne, and A. Ataullah (2023). What explains delays in public procurement decisions? *Economic Modelling 121*, 106201.

Hearst, M. A., S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf (1998). Support vector machines. *IEEE Intelligent Systems and their applications 13*(4), 18–28.

Iammarino, S., A. Rodriguez-Pose, and M. Storper (2019). Regional inequality in europe: evidence, theory and policy implications. *Journal of economic geography 19*(2), 273–298.

Incaltarau, C., G. C. Pascariu, and N.-C. Surubaru (2020). Evaluating the determinants of eu funds absorption across old and new member states–the role of administrative capacity and political governance. *JCMS: Journal of Common Market Studies 58*(4), 941–961.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *Science 353*(6301), 790–794.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction policy problems. *American Economic Review 105*(5), 491–495.

Kramer, O. and O. Kramer (2013). K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors*, 13–23.

Krstajic, D., L. J. Buturovic, D. E. Leahy, and S. Thomas (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics 6*, 1–15.

Kuhn, M. (2015). Caret: classification and regression training. *Astrophysics Source Code Library*, ascl–1505.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, R. C. Team, et al. (2020). Package 'caret'. *The R Journal 223*(7).

Lagravinese, R. (2015). Economic crisis and rising gaps north–south: evidence from the italian regions. *Cambridge journal of regions, economy and society 8*(2), 331–342.

Landis, J. R. and G. G. Koch (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Leduc, S. and D. Wilson (2013). Roads to prosperity or bridges to nowhere? theory and evidence on the impact of public infrastructure investment. *NBER Macroeconomics Annual 27*(1), 89–142.

Liu, Y., Y. Wang, and J. Zhang (2012). New machine learning algorithm: Random forest. In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, pp. 246–252. Springer.

Ministero per lo Sviluppo Economico (2007). Quadro strategico nazionale per la politica regionale di sviluppo 2007-2013. Technical report.

Murphy, K. P. et al. (2006). Naive bayes classifiers. *University of British Columbia 18*(60), 1–8.

Natekin, A. and A. Knoll (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics 7*, 21.

Nifo, A. and G. Vecchione (2014). Do institutions play a role in skilled migration? the case of italy. *Regional Studies 48*(10), 1628–1649.

Palcic, D., E. Reeves, D. Flannery, and R. R. Geddes (2022). Public-private partnership tendering periods: an international comparative analysis. *Journal of Economic Policy Reform 25*(2), 156–172.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia 4*(2), 1883.

Ramey, V. A. et al. (2021). *The macroeconomic consequences of infrastructure investment*, Volume 28215. University of Chicago Press Chicago.

Reeves, E., D. Palcic, D. Flannery, and R. R. Geddes (2017). The determinants of tendering periods for ppp procurement in the uk: an empirical analysis. *Applied economics 49*(11), 1071–1082.

Resce, G. and C. Vaquero-Piñeiro (2022). Predicting agri-food quality across space: A machine learning model for the acknowledgment of geographical indications. *Food Policy 112*, 102345.

Rish, I. et al. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Volume 3, pp. 41–46. Citeseer.

Rodríguez-Pose, A. (2013). Do institutions matter for regional development? *Regional studies 47*(7), 1034–1047.

Rodríguez-Pose, A. and E. Garcilazo (2015). Quality of government and the returns of investment: Examining the impact of cohesion expenditure in european regions. *Regional Studies 49*(8), 1274–1290.

Sansone, D. and A. Zhu (2023). Using machine learning to create an early warning system for welfare recipients. *Oxford Bulletin of Economics and Statistics 85*(5), 959–992.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks 61*, 85–117.

Steinwart, I. and A. Christmann (2008). *Support vector machines.* Springer Science & Business Media.

Surubaru, N.-C. (2017). Administrative capacity or quality of political governance? eu cohesion policy in the new europe, 2007–13. *Regional Studies 51*(6), 844–856.

Sutskever, I., O. Vinyals, and Q. V. Le (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems 27.*

Tharwat, A. (2021). Classification assessment methods. *Applied computing and informatics 17*(1), 168–192.

Titl, V., D. Mazrekaj, and F. Schiltz (2024). Identifying politically connected firms: A machine learning approach. *Oxford Bulletin of Economics and Statistics 86*(1), 137–155.

Tosun, J. (2014). Absorption of regional funds: A comparative analysis. *JCMS: Journal of Common Market Studies 52*(2), 371–387.

Wong, H. L., Y. Wang, R. Luo, L. Zhang, and S. Rozelle (2017). Local governance and the quality of local infrastructure: Evidence from village road projects in rural china. *Journal of Public Economics 152*, 119–132.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology 67*(2), 301–320.