Received: 26 January 2026 | Revised: 6 August 2026 | Accepted: 21 September 2026 | Published online: 6 November 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–24
DOI: 10.47852/bonviewJCCE62029210

BON VIEW PUBLISHING

Identifying Structural Drivers of Crime Using Statistical Inference and Machine Learning Models

Alibek Barlybayev[1, 2], Mira Kaldarova[1,*], Rakhat Yerezhepov[1], Nurlan Botabayev[3], Madi Tugubayev[3] and Gulmira Shakhmetova[2]

[1] *Higher School of Information Technology and Engineering, Astana International University, Kazakhstan*

[2] *L.N. Gumilyov Eurasian National University, Kazakhstan*

[3] *Committee on the Legal Statistics and Special Accounts of the State Office of Public Prosecutor of the Republic of Kazakhstan, Kazakhstan*

Abstract: This study identifies structural socio-economic correlates of crime investigation activity in Kazakhstan and examines how interpretable composite indicators relate to it. The dependent variable is the annual volume of pre-trial investigations recorded in the Unified Register of Pre-Trial Investigations over 2015–2025. It is modeled against 70 explanatory factors that span socio-economic, demographic, institutional, legal, healthcare, education, agriculture, and digitalization domains. The analysis combines Pearson correlation, one-way analysis of variance, simple and multivariate linear regression, and regularized ensemble machine learning (Gradient Boosting, Random Forest, Ridge, Support Vector Machine). A category-based dimensionality reduction then aggregates the 70 factors into 8, and subsequently 10, standardized composite indices. Because only about 10 aggregated annual observations are available, findings are reported in associational rather than causal terms. The strongly negative cross-validated R2 of the ensemble models is flagged as direct evidence of overfitting in this small-*n*, high-dimensional setting. Digitalization, education, poverty, and healthcare indicators emerge as the leading correlates of crime. A parsimonious regression on the composite indices accounts for 97.6% of in-sample crime variance while improving interpretability and stability. Interpretable index-based models therefore appear better suited than high-capacity learners for policy-relevant inference in this setting.

Keywords: crime determinants, pre-trial investigations, socio-economic vulnerability, category-based modeling

1. Introduction

The identification of structural elements that determine criminal behavior stands as a primary research priority for criminology, public policy, and socio-economic studies [1]. Modern crime patterns demonstrate multiple factors. These include economic conditions, social vulnerability, institutional capacity, human capital development, and technological progress. The current criminological methods do not effectively explain how criminal behavior and police work patterns develop in contemporary society [2].

Administrative crime registries offer researchers a valuable dataset to study the operational mechanisms of these systems. The Unified Register of Pre-Trial Investigations (ERDR) operates as a complete national database [3]. This database monitors all investigative activities across Kazakhstan. The ERDR system monitors institutional reactions to criminal offenses. This enables researchers to analyze social behavior and government operational capacity within a single analytical framework [4]. The dual characteristics of ERDR make it an ideal research instrument that enables scientists to study how social economic development interacts with institutional power and technological progress to impact criminal offenses.

Research studies conducted in various nations show that education serves as a crime prevention tool [5]. Crime also declines when combined with employment programs, healthcare services [6], and digital communications systems. But poverty, unemployment, and unstable family environments create conditions for people to become criminals [7]. The growth of digital infrastructure together with digital literacy development has created a revolutionary change that affects how people access information, work opportunities, and public services and how they interact with governing systems. The development of crime prevention requires societies to build complete economic systems. These systems need to surpass traditional legal approaches to prevention. Empirical studies have two main limitations. First, researchers

***Corresponding author:** Mira Kaldarova, Higher School of Information Technology and Engineering, Astana International University, Kazakhstan. Email: mira_kaldarova@aiu.edu.kz

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

use broad social indicators that obscure diverse results [8]. Second, they employ disparate research approaches that prevent them from combining statistical analysis with modern machine learning methods [9].

This research fills these knowledge gaps through its complete system-based evaluation method, which studies factors that affect crime investigation patterns. The research analyzes more than 70 indicators, which include socio-economic, demographic, institutional, legal, healthcare, educational, agricultural, and digitalization data. Classical statistical methods and ensemble machine learning models such as Gradient Boosting and Random Forest were also used. The hybrid method enables to conduct causal analysis and predictive modeling.

The research introduces a novel methodological contribution through a category-based approach that reduces the dimensionality of the data. The analysis combines explanatory variables into 8 and 10 aggregated indices. The method decreases the data dimensions to a lower number. This approach yields interpretable results and helps mitigate the overfitting problems typical of high-dimensional crime data. The research expands this framework through social sphere analysis. The study separates poverty from family distress and family relationship factors.

This study offers three contributions. The first concerns digital development. Among the economic, legal, and social factors examined, the Digitalization composite index, built from indicators of digital literacy and internet access, shows the strongest negative association with the annual volume of pre-trial investigations in the ERDR, at an index-level correlation of $r = -0.952$. The second contribution concerns human capital. Advanced-education indicators are associated with lower crime and behave as a stabilizing correlate, and the number of doctoral students alone accounts for a large share of the observed variance ($r = -0.984$). The third contribution shows that social vulnerability is not monolithic. When the Social Sphere category is disaggregated, its poverty ($r = -0.912$) and family distress ($r = +0.856$) sub-dimensions relate to crime in opposite directions, which fits the case for differentiated rather than uniform social-policy targeting. All three findings are statistical associations estimated on aggregated observational data, and none should be read as a causal effect.

The present study is observational and correlational, not causal, and we state this plainly at the outset. The analysis draws on aggregated national annual series spanning 2015–2025, which yield only a small number of yearly observations against a large set of candidate factors. In this small-sample, high-dimensional setting, the estimated relationships capture statistical co-movement between structural indicators and crime. They do not identify the underlying generative or behavioral mechanisms. Confounding, reverse association, and shared temporal trends cannot be ruled out. For this reason, the manuscript uses associational language throughout and treats any mechanistic reading strictly as a hypothesis. The structural correlates reported here are meant to motivate a later causal agenda built on dedicated identification strategies, including structural equation modeling, instrumental-variable estimation, and difference-in-differences designs, ideally supported by multi-region and finer-grained spatiotemporal data able to sustain credible causal inference.

Within this observational scope, four research questions guide the study. The first (RQ1) asks which structural socio-economic, demographic, and institutional factors are most strongly associated with the annual volume of pre-trial investigations recorded in the ERDR. The second (RQ2) asks whether category-based dimensionality reduction, which aggregates the 70 individual factors into a compact set of composite indices, preserves explanatory power while improving the interpretability and stability of the resulting models. RQ3 turns to method choice in a small-n, high-dimensional regime: do statistical-inference models give more reliable and interpretable evidence than high-variance ensemble machine learning models, whose near-perfect in-sample fit coincides with strongly negative cross-validated performance? The fourth (RQ4) asks whether disaggregating social vulnerability into distinct sub-dimensions reveals heterogeneous, and sometimes opposing, associations with crime that stay hidden when social vulnerability is treated as a single composite.

2. Literature Review

The results of the study by Salmi and Kivivuori [10] demonstrate that juvenile delinquency is systematically associated with both social capital and individual-level characteristics. The research, which involved Finnish adolescents and utilized multivariate analysis, found that parental support, teacher authority, and peer relationships continued to predict adolescent criminal behavior. This was replicated when the researchers included environmental factors and individual characteristics, such as self-regulation and intellectual ability, in their analysis. Results indicate that personal characteristics in the analysis make these relationships weaker according to the study. Research shows that low self-control is a leading factor in delinquent behavior. This factor significantly increases the risk of committing crimes. Low cognitive abilities and poor academic performance also lead to an increased risk of delinquency. Findings show that intergenerational closure and parental control and teacher support only appear as important factors when studying these variables separately, but they do not show any protective effect when parents help their children access the workforce. Social capital explains juvenile delinquency as an independent factor. But its influence depends on personality characteristics. Therefore, researchers must develop combined analytical methods for studying delinquency.

The results of the study by Cheung and Cheung [11] show that low self-control is significantly associated with delinquency at the bivariate level among Hong Kong adolescents. But its explanatory power decreases once social factors are taken into account. Using variables from social bond, differential association, strain, and labeling theories in the analysis, the study's results show that self-control does not predict any types of offenses, including general, property, and violent crimes. Even if we take into account the self-control factor, social factors such as deviant relationships with peers, deviant beliefs, problems at school, strict parenting, and stressful life situations continue to be closely associated with deviant behavior. In Chinese culture, social and interpersonal factors have a greater influence on criminal behavior. This partially supports the general theory of the cultural independence of crime. The study demonstrates that cultural factors must be considered when studying adolescent delinquent behavior.

The results of the paper by Ingilevich and Ivanov [12] demonstrate that crime rates in urban settings can be predicted with high accuracy. Using crime data from Saint Petersburg, the authors show that Gradient Boosting consistently outperforms linear and logistic regression models across multiple crime categories, achieving high explanatory power ($R^2 \approx 0.9$) while avoiding unrealistic predictions such as negative crime counts. Feature selection based on the chi-square test reduces model complexity. Feature selection also reveals population size, the density of bars, churches, and schools as the most influential predictors. The presence of bars is found to be the strongest predictor of criminal activity.

The results of the research by Cadena Urzúa and Letelier Saavedra [13] show that high social-connotation crimes in the Metropolitan Region of Chile are significantly shaped by a combination of spatial, temporal, and contextual factors. Over 370,000 police reports were analyzed using a logistic regression model. The results showed that population density, overcrowding, rurality, the presence of green spaces, and per capita municipal funding lead to different outcomes for different crime types. Overcrowding increases the probability of violent crimes such as homicide, assault, rape, and robbery with violence, while higher municipal investment and the presence of green areas are generally associated with lower probabilities of interpersonal violence. Data show that criminal activity follows certain time patterns. Different crimes are committed more frequently on certain days of the week and months of the year. The location of a community, as well as its surrounding environment, determines crime rates. Areas with entertainment venues, shopping centers, transportation hubs, and pharmacies have higher crime rates. The findings demonstrate that criminal activity is distributed according to social patterns, geographic locations, and time periods. This supports the feasibility of implementing crime prediction methods and targeted prevention strategies.

The scientific research by Kester and Afoma [14] produced results that showed Bayesian inference and Markov chain models successfully measure cybercrime probability through combined analysis of socio-economic and geographic elements. Research on phishing attacks reveals that phishing attacks occur most frequently in communities with extroverted structures, when people earn either low or high incomes, when parents fail to provide proper guidance, and when hackers use advanced techniques. The study reveals that law enforcement power and religious beliefs with positive aspects reduce the chances of cybercrime occurrence. Users can apply conditional reasoning within a probabilistic model. This will allow them to determine the likelihood of crime in specific areas based on active causal factors. It will also allow them to predict future crime targets by examining past crime data. Markov transition analysis reveals that cybercrime follows certain patterns. These patterns combine temporal and spatial relationships. Attackers typically choose their next targets based on the locations of previous crimes. Bayesian-Markov models function as powerful decision-making systems. Such models allow analysts to predict cybercrime in banking and interbank information systems.

The results of the study by Liebe and Schwitter [15] demonstrate that violence against refugees in Germany follows clear spatial, temporal, and contextual patterns. Event-history (Cox) models show strong diffusion effects. Prior attacks in the same district and geographical proximity to recent incidents increase the likelihood of subsequent violence. The findings indicate that higher shares of foreigners at the district level are associated with lower risks of both personal and miscellaneous attacks. The competition-for-resources hypothesis is not supported. Unemployment exhibits no robust independent effect. Political opportunity structures and regional context play a differentiated role. The presence of populist and right-wing parties reduces violence in West Germany. But these parties are associated with higher risks in East Germany, particularly for less severe forms of violence.

The results of the article by Butt et al. [16] demonstrate that deep learning models with transfer learning outperform traditional statistical approaches in spatiotemporal crime prediction. Comparative experiments were conducted on large-scale crime datasets from Chicago, New York, and Lahore. Findings show that the bidirectional long short-term memory (BiLSTM) model achieves the lowest prediction errors (MAE, median absolute deviation (MAD), and mean squared error (MSE)). Low error values are observed for both weekly and monthly forecasts. The integration of transfer learning reduces execution time and further improves predictive accuracy.

The results of the paper by Sam et al. [17] indicate that social crime involvement among Malaysian students is influenced by a combination of psychological, social, and familial factors. Statistical analyses based on survey data from 383 respondents show that parental factors exhibit the highest mean scores as protective determinants. Parental behavior, communication, and support hinder student involvement in social crimes. Self-efficacy, peer pressure, mass media influence, and parental factors are all associated with social crime involvement. Multiple regression results demonstrate a strong overall model fit ($R^2 = 0.72$), with self-efficacy showing the strongest negative association with social crime, while mass media influence emerges as the most powerful positive predictor, followed by peer pressure.

The results synthesized in the review by Doğan [18] indicate that crime patterns in Turkey are best explained through a social disorganization framework. The empirical studies reviewed noted a link between unemployment, income inequality, labor market segmentation, regional economic disparities, and higher crime rates. Evidence on migration shows no uniform direct effect on crime. Political upheavals, economic crises, and governance challenges exacerbate social disorganization. These events undermine institutional trust and informal social control. These events create crime-prone environments, especially in rapidly growing urban and migrant-dense areas.

The results of the research by Jen [19] provide strong empirical support for the law of crime concentration at place in a non-Western urban context. Analysis of burglary incidents in Taipei City from 2015 to 2019 shows that crime is concentrated in a small number of micro-places. Approximately 51.3% of all burglaries occurred in only 5% of equivalent street-segment units. 25.9% were concentrated in just 1.7% of units. More than 80% of spatial units experienced no burglaries at all. Lorenz curves and Gini-based measures further confirm substantial spatial inequality in the distribution of burglary. The Gini coefficient (0.4967) indicates concentration despite overall low crime levels. These patterns remain stable across individual years. Burglary in Taipei is persistently clustered in specific locations.

The results of the study by Kennedy et al. [20] demonstrate a strong and systematic association between low food access and criminogenic environmental conditions in Newark, New Jersey. The study was conducted using risk modeling and multilevel regression analysis. Neighborhoods with small grocery stores and abandoned buildings are associated with higher rates of violent assault. Micro-level analyses reveal that proximity to bodegas increases the risk of aggravated assault by more than sevenfold. Meso-level models indicate that aggregated crime risk scores and abandoned properties remain predictors of food deserts. Physical environmental factors outweigh social and racial-ethnic variables in the combined regression model. Food insecurity and exposure to violence reinforce one another spatially.

3. Research Methodology

3.1. Analytical framework and system-level formalization

The annual volume of pre-trial investigations recorded in the ERDR is treated here as the observable output of a coupled

socio-technical system. Its behavior is governed by a small number of latent sectoral states rather than by any single indicator on its own. The thematic structure of the 70 explanatory factors suggests eight latent sectoral states for year *t*: digitalization D_t, human capital and education E_t, healthcare capacity H_t, labor market L_t, economic conditions Ec_t, social vulnerability S_t, legal-institutional capacity Lg_t, and agricultural conditions A_t. The data-generating process can then be written compactly as ERDR_t = f(D_t, E_t, H_t, L_t, Ec_t, S_t, Lg_t, A_t) + eps_t, where f is an unknown, possibly nonlinear mapping from the sectoral state vector to investigation activity, and eps_t is an idiosyncratic disturbance that absorbs measurement error and factors outside the modeled domains. No latent state is observed directly. Each one is operationalized by a composite index I_k constructed from the indicators assigned to sector k, so that I_{k, t} = (1/|c_k|) × sum_{j in c_k} z_{j, t}, where z_{j, t} = (x_{j, t} − mean_j)/sd_j is the standardized value of indicator j in year t and c_k is the set of indicators belonging to sector k. This construction is the formal basis for the category-based dimensionality reduction developed in Section 3.6, which collapses the 70-factor space into 8 (and subsequently 10) composite indices.

With only about 11 annual observations available for 2015–2025, estimating the full nonlinear mapping f is not feasible. We instead work with a first-order, static reduced form obtained by linearizing f around the observed operating range, ERDR_t = b0 + sum_k b_k I_{k, t} + eps_t, where b_k is the marginal association between composite index I_k and ERDR and b0 is the intercept. Within this approximation, the estimated coefficients b_k, the index-level Pearson correlations, and the analysis of variance (ANOVA) *F*-statistics reported below are read as the relative statistical weight of each sector in accounting for observed variation in pre-trial investigation activity. The specification is deliberately parsimonious. It trades the descriptive detail of the full system map for estimability and interpretability, and it is the specification for which all numerical results in this study are reported.

To make the cross-sector structure explicit and to motivate further structural analysis, the system is also represented as a weighted undirected graph $G = (V, \mathrm{Eu})$. The vertex set V contains the sectoral states and, at a finer resolution, the individual indicators. The edge set Eu connects pairs of vertices whose partial correlation, computed after conditioning on the remaining variables, exceeds a chosen threshold, with the absolute partial correlation serving as the edge weight. This graph-theoretic view summarizes which sectors share variation that is not mediated by other sectors. It provides a foundation for future structural equation modeling and for graph-theoretic descriptors of the determinant network, such as centrality, community structure, and path-based influence. With the present sample size, G serves as an organizing and descriptive device rather than an identified causal diagram. Estimating a sparse partial-correlation network and the associated structural model with valid uncertainty quantification is reserved for future work.

The static reduced form is the cross-sectional projection of an underlying dynamical system in which sectoral states co-evolve through annual feedback. As a methodological roadmap, this dynamic structure can be written as a discrete-time system x_{t + 1} = x_t + g(x_t, u_t), where x_t = (D_t, E_t, H_t, L_t, Ec_t, S_t, Lg_t, A_t) is the sectoral state vector, u_t denotes exogenous policy controls (e.g., digital inclusion, education, or healthcare capacity interventions), and g encodes the cross-sector transition dynamics. Such a formulation connects naturally to system-dynamics and agent-based representations of the determinant system, and once identified, it would permit scenario simulation and constrained policy optimization over u_t. We do not estimate g here. Identifying cross-sector transition dynamics and time-lagged feedback needs far longer series than we have, so that step waits on longer records and on multi-region panels. What the present framework delivers is the co-variation between composite sectoral indices and pre-trial investigation activity; the graph-theoretic, structural equation, and system-dynamics extensions outlined above remain the principal direction of future research.

3.2. Data preprocessing and feature engineering

The empirical analysis rests on a national administrative panel observed at annual frequency over the period 2015–2025. The dependent variable is the annual volume of pre-trial investigations recorded in the ERDR, used here as a measure of registered crime. On the explanatory side stand 70 indicators spanning socio-economic, demographic, institutional, legal, healthcare, education, agriculture, and digitalization domains. Because every series is aggregated to the calendar year, the effective sample comprises only about 10–11 yearly data points, so the number of candidate predictors far exceeds the number of temporal observations. That configuration shapes the analytical choices that follow: it favors parsimonious and regularized specifications, and it is the source of the overfitting documented in Section 4.3. The inferential limits it implies are stated in Section 5.

The 70 indicators are organized into eight substantive domains that reflect the structural dimensions hypothesized to be associated with pre-trial investigation activity: Education (10 indicators), Healthcare (8), Economy (15), Employment (6), Social Sphere (9, of which 8 vary over the observation window), Legal System (11), Agriculture (8), and Digitalization (3). This mapping serves two purposes. First, it imposes an interpretable thematic structure that organizes a large and heterogeneous indicator set into coherent socio-economic constructs. Second, it supplies the conceptual basis for the category-based dimensionality reduction introduced in Section 3.6, in which the indicators belonging to each domain are aggregated into a single composite index. The domain assignments are fixed in advance on substantive grounds rather than estimated from the data. That choice keeps the construction transparent and avoids capitalizing on chance associations.

Variable construction begins from raw administrative counts and rates harmonized to a common annual frequency. The ERDR aggregate series is the annual count of registered pre-trial investigations. The explanatory indicators are of two kinds: administrative counts, such as numbers of doctors, hospitals, doctoral students, or registered cattle, and rates or ratios, such as the poverty rate, the unemployment rate, and the digital-literacy share. During harmonization, the reporting period that had been recorded in error as “2025–2025” in the original data documentation was corrected to the full 2015–2025 interval. This is a metadata correction only and does not alter any underlying value. Indicators reported in heterogeneous units, including absolute counts, percentages, and monetary values such as the USD exchange rate, were aligned to a consistent annual definition before any modeling. Missing indicator-years were handled with procedures suited to short annual series: listwise exclusion of incomplete cases or linear interpolation between adjacent years where a series was otherwise continuous. Given the small number of observations, we do not assert a specific imputation count, and only indicators with adequate annual coverage over the study window were retained.

Because the indicators differ markedly in scale, sign, and units, each one was standardized to a *z*-score before aggregation and modeling, defined as z_j = (x_j – mean_j) / sd_j, where x_j is the raw value of indicator j and mean_j and sd_j are its sample mean and standard deviation computed over the available annual observations. Standardization places all indicators on a common, dimensionless scale with zero mean and unit variance. No single high-magnitude variable, such as a monetary or large-count series, can then dominate purely by virtue of its units. The step is also a precondition for the later composite indices. A domain index I_c is formed as the within-domain average of standardized indicators, I_c = (1/|c|)×sum_{j in c} z_j, an operation that is meaningful only when its constituent terms are unit-free and comparable in magnitude. Standardization thus plays a dual role: it stabilizes the regression and machine learning models, and it provides the formal precondition for the category-based dimensionality reduction developed in Section 3.6.

The data analyzed in this study were drawn from two official sources. The ERDR aggregate annual series of pre-trial investigations was retrieved from the ERDR portal of the Prosecutor General's Office of the Republic of Kazakhstan[1], to which the authors had institutional access through the Committee on Legal Statistics and Special Accounting. The 70 socio-economic, demographic, institutional, and sectoral indicators were obtained from the Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan and compiled by the Prosecutor General's Office. Access conditions for both sources, together with the availability of the derived composite indices, the harmonized analysis dataset, and the analysis code, are set out in the Data Availability Statement.

This study used data from the ERDR for the years 2015–2025. ERDR is an automated database that records information about criminal offense reports, procedural decisions made on them, actions taken, the progress of criminal proceedings, and applicants and participants in criminal proceedings. The register and the conditions under which its data can be accessed are described in the Data Availability Statement. The Prosecutor General's Office of the Republic of Kazakhstan also compiled data on socio-economic factors for the period 2015–2025. The dataset consists of 70 factors. List of factors: Criminal Sphere, Criminal Legislation, Criminal Procedure Code Legislation, Criminal Executive Code Legislation, Administrative Offenses, Civil Procedure Code Legislation, Construction, Preschool Education, Secondary Education, Post-Secondary Education, Higher Education, Master's Degree, Acreage, Wheat, Cattle, Small Horned Livestock, Horses, Agricultural Products, Employed in Agriculture, Farms, Cases Registered, Court Cases, Inflation, Food Prices, Non-Food Prices, Service Prices, Wages, Real Wage Index, USD Rate, RUB Rate, Labor Force, Employed Persons, Self-Employed, Unemployment, Total Unemployment, Youth Unemployment, Poor Population, Poverty Rate Avns, Dysfunctional Families, Children in Orphanages, Orphans, Orphanages, Marriages, Divorces, Registered for Narcology, Medical Organizations, Hospitals, Clinics, Private Clinics, Hospital Beds, Doctors, Medical Staff, Medical Staffing, Undergraduate Students, Master's Students, Doctoral Students, Residents, College Students, Court Funding, Court Payroll, Court_Material_and_Technical_Support, Digital Literacy, Internet Adults, Internet Houses, Loans, New_Building_Price, Resale_Price, Housing_Rent, Restaurants, and Cafes. No usable variation is available for the Orphanages series over 2015–2025, so its correlation with ERDR is undefined. The indicator is therefore kept in the list above but excluded from the correlation analysis, which leaves 69 factors carrying usable variance.

Figure 1 shows the annual dynamics of changes in legislation from 2015 to 2024. Dynamics are disaggregated by legal domain into criminal, administrative, and civil acts. The total number of legislative acts fluctuates from year to year but shows an overall upward trend over the study period, from 55 in 2015 to 63 in 2024 (peaking at 73 in 2023).

[1] https://erdr-public.kgp.kz/

Figure 1
Statistics on legislative changes

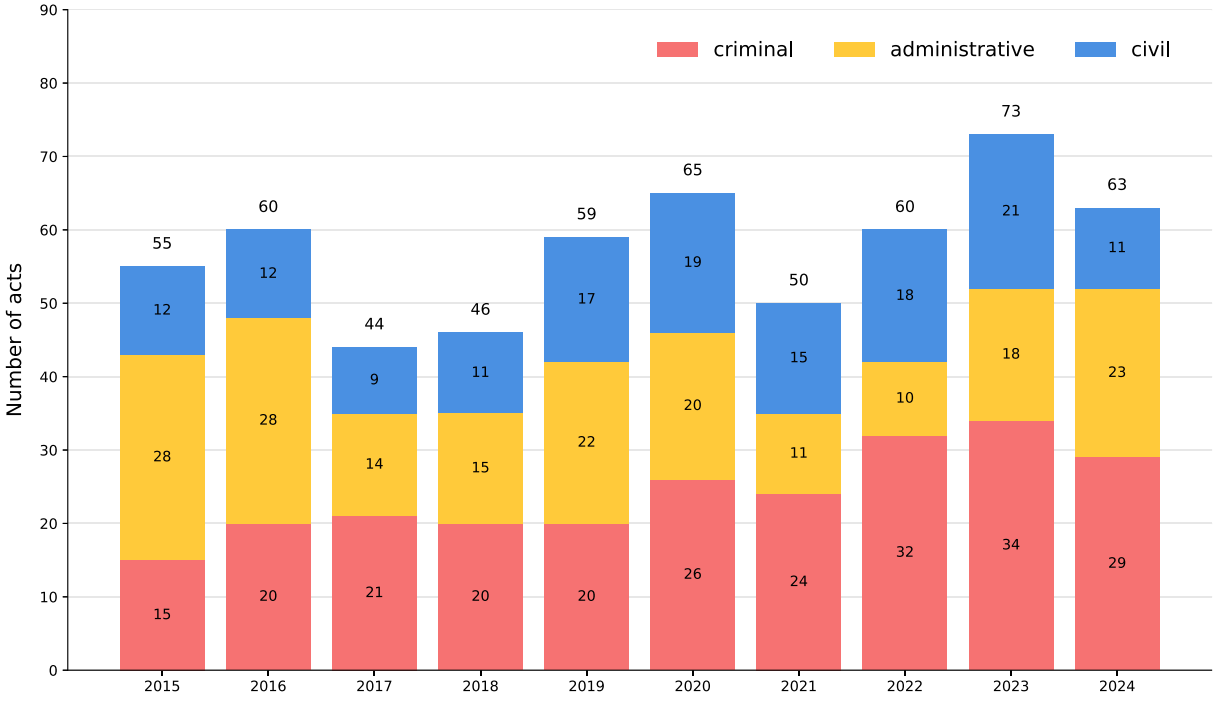


Figure 2 presents the temporal dynamics of registered crimes from 2015 to 2024. The data demonstrate a consistent downward trend.

3.3. Statistical inference and model evaluation methods

Pearson correlation [21] was used to assess the strength of linear relationships. The coefficient varied between −1 and +1. Negative values show inverse relationships. Positive values show direct relationships. The magnitude of $|r|$ was used to classify the strength of relationships (weak, moderate, strong, or very strong). Statistical significance was assessed using p-values with a threshold of $p < 0.05$.

Model performance was evaluated using the coefficient of determination (R^2) [22], root mean squared error (RMSE) [23], and mean absolute error (MAE) [24]. R^2 was used to assess the proportion of variance in crime. RMSE measured the average magnitude of forecast errors. MAE measured the mean absolute difference between observed and predicted values.

3.4. Model specifications and hyperparameter configuration

The following models were applied to model the relationships: Gradient Boosting [25], Random Forest [26], Ridge Regression [27], and Support Vector Machine (SVM) [28]. Gradient Boosting was used as an ensemble learning method. A Random Forest constructs multiple decision trees. This model uses bootstrap sampling and random feature selection. Ridge Regression is used as a regularized linear model. Ridge Regression includes an L2 penalty. This penalty reduces multicollinearity and stabilizes coefficient estimates. SVM is used to model nonlinear relationships. SVM projects data into a higher-dimensional feature space using kernel functions.

All predictive models were implemented in Python with the scikit-learn library, and their configurations are reported in full in Table 1 so that the results can be reproduced. The estimators fall into two families. The classical statistical component covers ordinary least squares linear regression, fitted in closed form by minimizing the residual sum of squares, together with one-way ANOVA based on the F-statistic, with significance assessed at $p < 0.05$. The machine learning component covers four estimators. Gradient Boosting is an additive ensemble of gradient-boosted regression trees with 100 estimators, a learning rate of 0.1, a maximum tree depth of 3, and the squared-error loss. Random Forest is a bootstrap-aggregated ensemble of 100 regression trees; each tree is trained on a bootstrap resample of the observations, each split is evaluated on a square-root subset of the candidate features (sqrt feature subsampling), and tree depth is left unrestricted. Ridge Regression adds an L2 penalty on the coefficient vector, with the regularization strength fixed at alpha = 1.0. Support vector regression (SVR) uses a radial basis function (RBF) kernel with the penalty parameter $C = 1.0$ and the epsilon-insensitive tube width set to epsilon = 0.1.

We did not pursue extensive hyperparameter optimization. A grid or randomized cross-validation search over so few annual cases would have inflated the researcher degrees of freedom and risked overfitting the validation folds themselves, returning configurations that look favorable by chance rather than by genuine generalization. Each estimator was instead run with the standard default settings listed above. The absence of tuning is a deliberate constraint that limits model flexibility and keeps the comparison across estimators transparent and reproducible. These defaults were applied uniformly to all 70 explanatory factors and then to the aggregated composite indices, so that differences in reported performance reflect the structure of the data and the inductive bias of each model rather than estimator-specific tuning effort.

The defaults are not claimed to be optimal for the present task. They are meant to be defensible and neutral, and should

Figure 2
Number of registered crimes

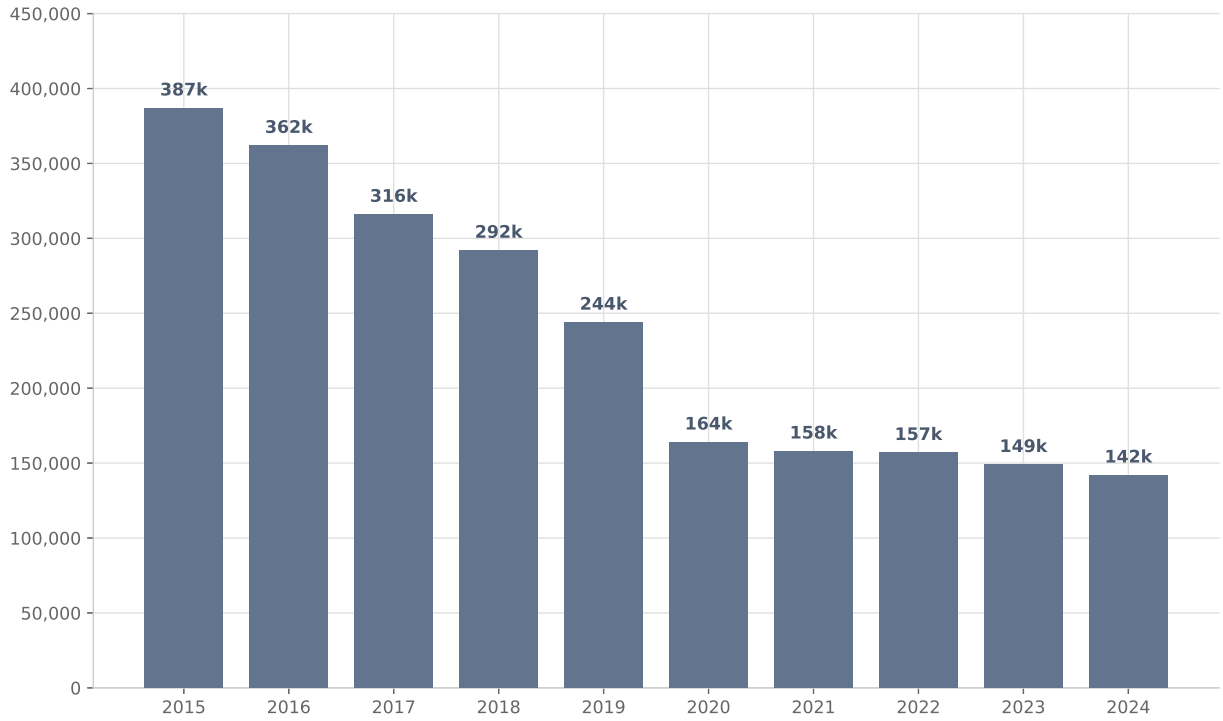


Table 1
Model specifications and hyperparameter settings

Model	Type	Key hyperparameters	Implementation
Gradient Boosting	Gradient-boosted regression trees	*n*_estimators = 100; learning_rate = 0.1; max_depth = 3; loss = squared error	scikit-learn (GradientBoostingRegressor)
Random Forest	Bagged regression-tree ensemble	*n*_estimators = 100; bootstrap = True; max_features = sqrt; max_depth = None (unrestricted)	scikit-learn (RandomForestRegressor)
Ridge	L2-penalized linear regression	Penalty = L2; alpha = 1.0	scikit-learn (Ridge)
SVM	Support vector regression	Kernel = RBF; C = 1.0; epsilon = 0.1	scikit-learn (SVR)
Linear Regression	Ordinary least squares	Closed-form least squares; no regularization	scikit-learn (LinearRegression)
ANOVA	One-way F-test	F-statistic; significance threshold $p < 0.05$	scipy/statsmodels

be read as a conservative baseline. Tuning would not in any case repair the overfitting reported in Section 4.3, which originates in the design rather than in the parameter settings. The time-aware validation protocol described in Section 3.7 and the reduced-complexity composite-index models are the appropriate remedies for the overfitting identified, while formal hyperparameter search is left to future work, where a larger or more granular panel, for example, regional or monthly data, would make such optimization statistically meaningful.

3.5. Linear regression and analysis of variance

Linear regression [29] was used to assess the relationship with crime. This method identifies a linear relationship between a dependent variable and a set of independent predictors. Linear regression (LR) estimates the model parameters using residual sum-square minimization. Linear regression is used in both a simplified, single-predictor form and a multivariate form. Model performance was assessed using the coefficient of determination (R^2) and error-based metrics, allowing for interpretation of the direction of the effect, its magnitude, and its explanatory power within a transparent and interpretable statistical model.

ANOVA [30] was used to assess the statistical significance of aggregated categories and disaggregated subcategories. ANOVA statistically evaluates differences in mean values of the dependent variable across groups defined by explanatory factors. Between-group and within-group variability are compared using the F-statistic. To determine statistical significance, p-values with a threshold of $p < 0.05$ were used.

3.6. Category-based dimensionality reduction: formal definition

To handle the imbalance between 70 explanatory factors and the handful of annual observations available, we use a theory-driven dimensionality reduction scheme that aggregates the original indicators into a small set of conceptually coherent composite indices. Let x_{j, t} denote the value of indicator j in year t, for j = 1, ..., 70 and t = 1, ..., T. Each indicator is first standardized across the observation window, $z_\{j, t\}$ = (x_{j, t} – mean_j) / sd_j, where mean_j and sd_j are the sample mean and standard deviation of indicator j. Standardization places all factors on a common, unit-free scale and keeps indicators measured in large administrative counts from dominating the aggregation. The index set {1, ..., 70} is then partitioned into K mutually exclusive and conceptually homogeneous categories C_1, ..., C_K (Education, Healthcare, Economy, Employment, Social Sphere, Legal System, Agriculture, and Digitalization), so that the union of the C_k recovers the full set and the categories are pairwise disjoint. Each composite index is the within-category mean of the standardized indicators, $I_\{k, t\}$ = $(1/|C_k|) \times$ sum_{j in C_k} $z_\{j, t\}$, where $|C_k|$ is the number of indicators assigned to category k. Indicators may be sign-aligned before aggregation so that a higher value of $I_\{k, t\}$ consistently denotes greater sectoral development, for example, by reversing the sign of poverty- or distress-type indicators within a category. This makes each index monotone and directly interpretable. The construction shrinks the design matrix from 70 columns to K composite indices. In the primary specification, K = 8, an 88.6% reduction in the number of regressors, and K = 10 once the Social Sphere category is disaggregated. The reduced-form regression linking the dependent variable to the composite axes is then ERDR_t = b0 + sum_{K = 1}^{K} b_k $I_\{k, t\}$ + eps_t, where ERDR_t is the annual volume of pre-trial investigations, b_k is the partial coefficient on composite index k, and eps_t is the disturbance term.

We prefer this aggregation to standard linear and nonlinear embedding techniques because of the constraints of the present setting. Principal component analysis (PCA) and nonlinear manifold-learning methods such as t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and deep autoencoders are highly flexible and unsupervised. With only ~11 observations, their estimated loadings, embeddings, and reconstructions become statistically unstable. They are sensitive to hyperparameter choices and random initialization, and they tend to fit idiosyncratic year-to-year noise rather than recoverable structure, while the resulting axes are hard to label or to translate into policy terms. The category-based index avoids these failure modes by replacing data-driven axis estimation with a fixed, theory-specified partition. No loadings are estimated from the data, so no variance is spent on learning the projection, and each axis carries a named, substantive meaning such as digitalization, healthcare capacity, or employment. The

construction is in fact a constrained special case of PCA. Within each category, the equal-weight standardized average $I_{k,t}$ is the first principal component under the constraint that all loadings are equal. This trades a small amount of explanatory efficiency for substantial gains in stability and interpretability, and it removes the estimation variance that an unconstrained projection would incur in this regime. We therefore treat PCA, *t*-SNE, UMAP, and autoencoder-based representations not as competitors but as the right tools for larger or spatially disaggregated panels, for example, district- or month-level data, where the number of observations is sufficient to estimate flexible embeddings reliably. Their application to such extended data is a direction for future work.

The proposed composite index is closely related to the standardized additive composite indicators routinely used in official statistics and in multidimensional development and well-being indices, in which normalized sub-indicators are averaged within thematic pillars to summarize a domain. Our contribution is to use this familiar aggregation as a deliberate regularization device in a data-scarce inferential setting. Relative to unconstrained dimensionality reduction techniques, it yields named, monotone, and policy-interpretable axes whose associations with the dependent variable are stable across specifications. This stability and interpretability advantage, rather than any single fitted coefficient, is the principal methodological contribution of the paper. The empirical behavior of the resulting indices, including their estimated associations with the volume of pre-trial investigations, is reported in Section 4.

3.7. Validation, uncertainty quantification, and robustness

The observations form an annual, temporally ordered series spanning 2015–2025, so conventional random *k*-fold cross-validation is inappropriate. Such folds let information from later years inform predictions about earlier years, which inflates apparent performance. Our validation schemes instead respect the temporal structure of the data. The first is a time-aware, expanding-window scheme: the model is fitted on an initial contiguous block of early years and evaluated on the immediately following year, after which the training window is extended forward one year at a time. This preserves the forward-only ordering and emulates genuine out-of-sample forecasting. Because the number of yearly observations is so small, we complement it with leave-one-out cross-validation, in which each year is withheld in turn and the model refitted on the remaining years. Even these schemes expose the unreliability of the flexible models. As reported in Section 4.3, the cross-validated coefficient of determination is strongly negative for every machine learning model despite near-perfect in-sample fit: Gradient Boosting attains an in-sample R2 of 0.9994 but a cross-validated R2 mean of − 370.25, with Random Forest at − 659.60 and Ridge at − 848.68. A negative cross-validated R2 means the fitted models predict held-out years worse than a constant mean. That is direct evidence of severe overfitting in this small-*n*, high-dimensional regime, and it explains our methodological preference for low-complexity, interpretable specifications. The parsimonious composite-index regressions, in particular, are far less sensitive to the curse of dimensionality.

Two complementary frameworks quantify the uncertainty around point predictions. The first is conformal prediction, which yields distribution-free, finite-sample prediction intervals with a guaranteed marginal coverage level, for example, 90%, under only an exchangeability assumption and without any parametric error distribution [31]. Given the limited number of observations, we prefer the Jackknife+ variant to a simple split-conformal procedure: it relies on leave-one-out residuals and so avoids sacrificing scarce data to a separate calibration set while retaining rigorous coverage guarantees. The second framework is Bayesian regression, in the form of Bayesian Ridge Regression together with a hierarchical formulation that exploits the category structure. It returns full posterior distributions over the coefficients, giving credible intervals that express parameter uncertainty directly and regularizing the estimates through informative priors [14]. Conformal intervals around predicted crime and Bayesian credible intervals on the index coefficients thus describe predictive and parametric uncertainty that the point estimates alone cannot convey. We present both as adopted and recommended methodology for the present and subsequent analyses, not as sources of additional headline statistics.

Robustness and sensitivity are assessed along three axes. Leave-one-year-out refitting removes each year in turn and re-estimates the models, which shows whether any single year disproportionately governs the conclusions. Coefficient-stability checks then track the sign, magnitude, and relative ordering of the estimated effects across these refits and across alternative specifications, flagging predictors whose contributions are fragile. The third axis is a one-at-a-time perturbation analysis: small controlled changes are applied to each input in turn, and we record the resulting change in the predicted crime and, above all, in the ranking of predictors by importance. This measures how sensitive the ordering is to local variation in the data. The checks agree with the empirical results already obtained. High-level rankings are qualitatively stable, with digitalization, education, and poverty consistently emerging as the structural dimensions most strongly associated with crime, while the relative positions of individual weak predictors shift across refits and should not be read as reliable orderings. This separation between stable aggregate signals and unstable individual factors supports interpreting results at the composite-index level rather than at the level of single raw variables.

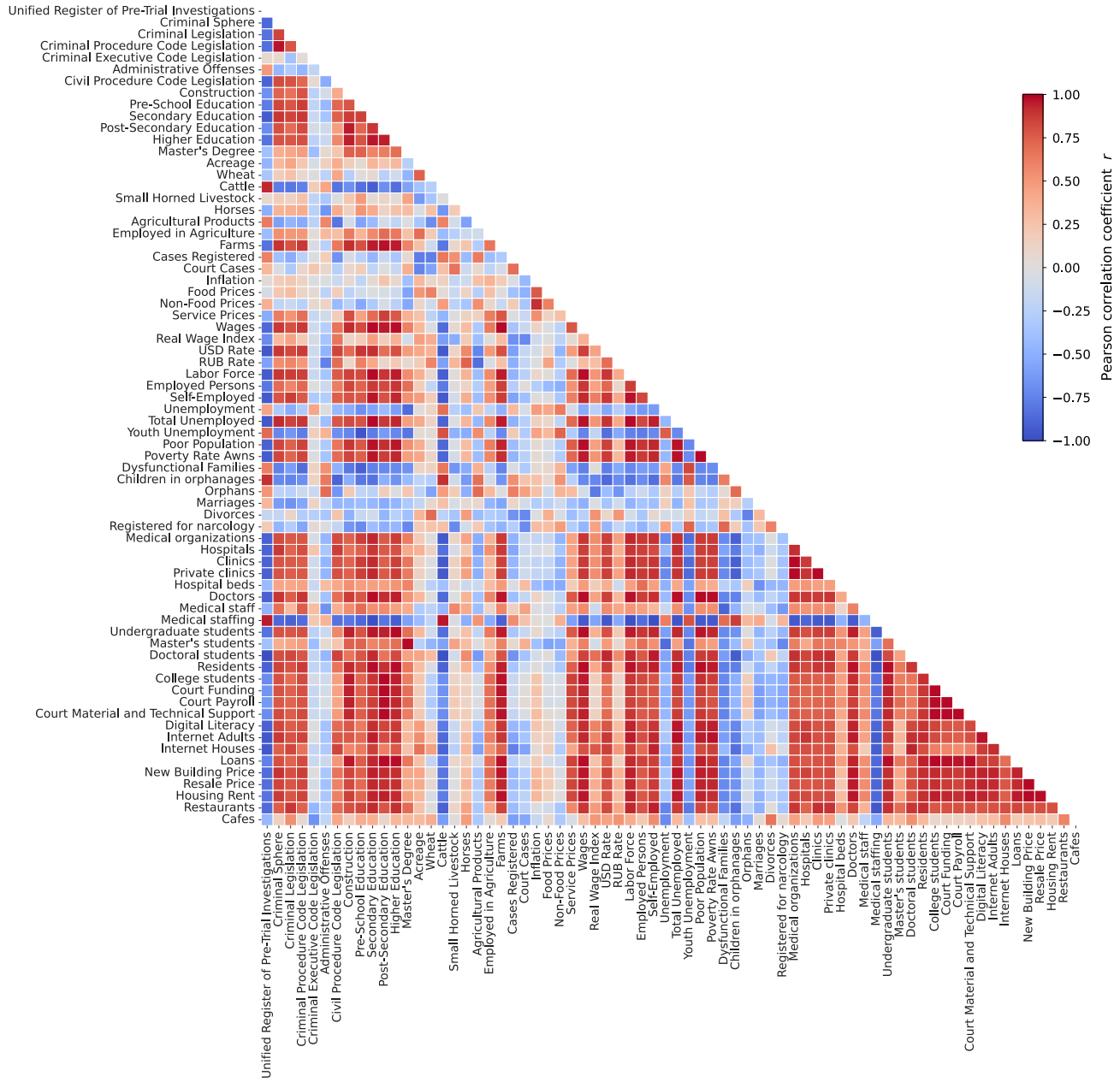
Socio-economic relationships need not be linear, so we also specify tools that test explicitly for nonlinear and threshold effects. Generalized additive models (GAMs) replace each linear term with a smooth penalized spline. The association between a predictor and crime is then estimated flexibly from the data while remaining additive and interpretable, with the degree of smoothness set by a penalty chosen through cross-validation. To probe discontinuities and regime changes, segmented (piecewise) and threshold regression models detect breakpoints at which the slope of an association changes. Such breakpoints would let us test substantive hypotheses, for instance, the existence of a minimum digital-literacy level below which the relationship with investigation activity differs in strength. Threshold autoregressive and self-exciting threshold formulations supply the matching time-series machinery for regime-dependent dynamics. These nonlinearity and threshold analyses are specified methodology and planned future work. They are not reported here as completed numeric results, and any threshold values would be estimated and validated within the time-aware and uncertainty-quantification framework described above.

4. Experiments and Discussions

4.1. Comprehensive analysis of factors influencing crime

Figure 3 shows a correlation heatmap. The lower-triangular correlation heatmap displays the pairwise linear relationships among a set of demographic and institutional indicators that belong to crime. The color scale ranges from strong negative

Figure 3
Correlation structure of socio-economic, legal, demographic, and institutional factors influencing crime



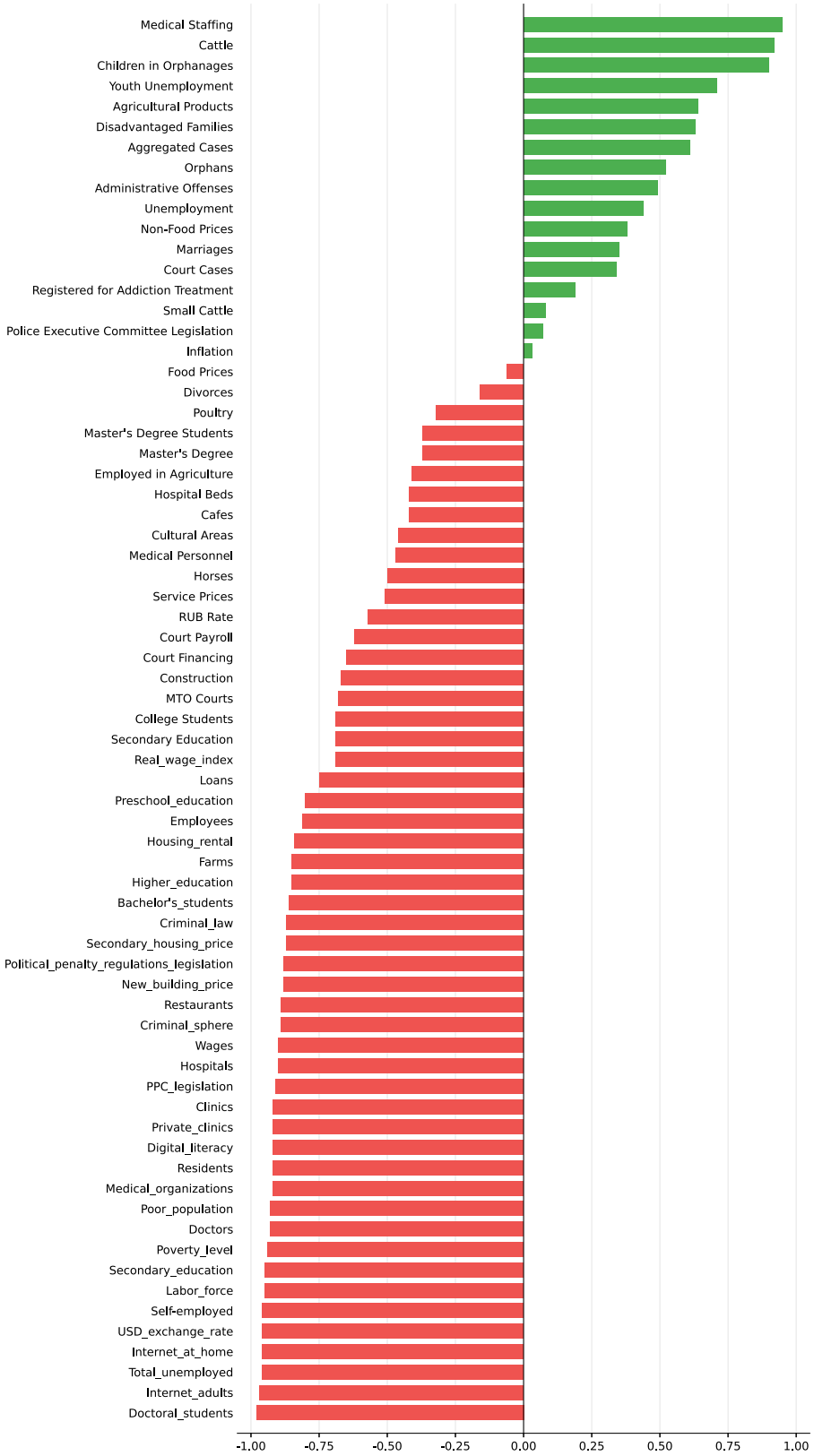
correlations in blue to strong positive correlations in red. The color intensity indicates the size of the correlation coefficient. The matrix covers ERDR together with the 69 factors that vary across the observation window; Orphanages is omitted for the reason given in Section 3.2.

Figure 3 shows that the variables cluster into thematic groups. Legal and judicial indicators such as laws, procedural laws, registered cases, and court measures show positive links. Legal and judicial indicators' strong positive links point to the criminal justice system being consistent. Social and economic indicators such as wages, employment, people working, and price indexes show links. Social and economic indicators' high positive links reflect the economic forces. We think the education-related variables (preschool, through levels) and the healthcare capacity indicators (hospitals, clinics, medical staff, and beds) form groups. The education-related variables and the healthcare capacity indicators have links. We observe that social vulnerability indicators such as unemployment, poverty, broken families, and orphaned children are associated with income, employment, education, and healthcare services. When the social vulnerability indicators rise, income, jobs, schooling, and health services fall. The exchange rate variables and inflation show patterns. Inflation links directly or indirectly to jobs and price measures.

Figure 4 shows the direction and how strong the influences of sociodemographic and institutional variables on crime are. It can be seen that the bars that go to the right of the zero line represent contributions.

A general analysis of the data suggests that social vulnerability and demographic stressors increase the number of investigations. Variables such as staffing, children in orphanages,

Figure 4
Directional influence of socio-economic, demographic, and institutional factors on crime



disadvantaged families, youth unemployment, and overall unemployment have positive numbers and point to a direct link between social instability and more investigative activity. The institutional and administrative indicators such as offenses, aggregated cases, court cases, and registration for addiction treatment also have effects. The institutional and administrative indicators such as offenses, aggregated cases, court cases, and registration for addiction treatment reflect that the legal system becomes more active under conditions of strain.

We notice that a broad set of development, education, healthcare capacity, and institutional update indicators show a negative effect. Higher education enrollment, including bachelor, master, and doctoral students, is linked to lower crime. Higher number of people working is linked to lower crime. Higher real wages, employment, higher digital literacy, and higher internet use all point to lower crime. Healthcare infrastructure variables, such as hospitals, clinics, medical organizations, and medical personnel, show contributions. Cultural and service sector development shows contributions.

To build high-quality crime prediction models, we needed to select the right variables. Using correlation analysis, we identified the top 20 important factors. These variables have a linear relationship with crime. The top 20 factors provide inputs for modeling. The variables capture a range of economic population, online, health, and government areas. The variables show that crime dynamics involve these factors.

The strongest negative correlation is observed for the number of students, with a Pearson correlation of − 0.984. The number of students is linked to crime intensity. Advanced education is linked to crime intensity. A larger student population and higher levels of educational attainment are associated with lower recorded crime, and these educational indicators are among the variables most strongly correlated with reduced pre-trial investigation activity. Signs related to inclusion and technological access, such as internet usage among adults (corr = −0.949) and internet availability at home (corr = −0.938). Digital literacy (corr = −0.892) shows strong negative correlations. Higher levels of digital integration and broader access to information resources are therefore associated with lower recorded pre-trial investigation activity.

Labor market characteristics are also crucial. Low unemployment (corr = −0.940), high self-employment (corr = −0.934), high labor force participation (corr = −0.928), and secondary education enrollment (corr = −0.926) reduce crime rates. Active participation in the labor market keeps crime low. Together, the results show that stable employment structures, educational attainment, and active participation in the labor market are key to stopping crime. In parallel, poverty-related indicators such as the poverty level (corr = −0.924) and the size of the poor population (corr = −0.897) further confirm the close linkage between socio-economic vulnerability and crime patterns.

Healthcare system capacity is another factor. The number of doctors (corr = −0.901), organizations (corr = −0.897), private clinics (corr = −0.891), and polyclinics (corr = −0.883) each show a link to crime-related outcomes. A higher overall healthcare staffing level is associated with a higher volume of pre-trial investigations (corr = 0.956). More developed healthcare infrastructure, as captured by the number of doctors, organizations, private clinics, and polyclinics, is correspondingly associated with lower levels of recorded crime.

The number of residents (corr = −0.893) is negatively associated with the volume of pre-trial investigations. The number of children in orphanages (corr = 0.911) shows that more children in orphanages go with crime. The number of cattle (corr = 0.933) shows that more cattle go with crime. The USD exchange rate (corr = −0.938) shows that a higher exchange rate goes with crime. The rural economy and external economic pressure on the tenge exchange rate may influence crime rates. Institutional and legal factors (corr = −0.880) indicate that rules and laws remain factors influencing crime outcomes.

4.2. Statistical significance of socio-economic factors on crime

Figure 5 shows results of an ANOVA test that checks how important the selected socio-economic, demographic, and institutional variables are for explaining the variation in crime. The horizontal bar chart in Figure 5 lists the F-statistics for each variable. Larger F-statistics in the chart mean the variable explains more of the variation. Larger F-statistics also mean the differences between groups are bigger than the differences inside groups.

The findings show high F-values. F-values explain the differences in investigation activity. The healthcare and education indicators stand out. The number of hospitals, the number of nursing staff, the number of doctors, and the enrollment numbers at undergraduate and doctoral levels rank among the indicators. The indicators point to the importance of capacity and human capital.

The socio-economic factors also show importance. Variables that measure poverty and the labor market matter a lot. The variables include the poverty rate, the population, the labor force size, self-employment, unemployment, and salary levels. The variables have high F-statistics. The high F-statistics confirm a link between the variables and changes in crime. The housing market indicators, such as building prices and secondary market prices, also contribute a lot. The financial variables, such as loans and the USD exchange rate, also contribute a lot. The housing market indicators and the financial variables together show that crime dynamics respond to the economy and financial situation.

The ANOVA results show that every factor listed is statistically significant at the $p < 0.01$ level. Each factor has power. The Hospitals factor has an F-statistic with F equal to 43.35 and p equal to 0.000184. F-statistic result points to healthcare infrastructure as the determinant among the variables. Here is a big group of economic, education, labor market, and health indicators. The list includes the USD exchange rate, medical staff, bachelor and doctoral students, residents, wages, farms, higher and secondary education, doctors, loans, self-employment, new-building and secondary-market housing prices, and labor force size. We note that these indicators share the same F-statistic ($F = 40.15$) and a low p-value ($p = 0.000229$). This indicates that each of these factors is strongly associated with variation in the dependent variable.

The research data reveal two main patterns because the population remains small and most people face financial difficulties ($F = 40.01$, $p = 0.000232$). The data indicate that financial difficulties create strong connections, which lead to criminal behavior. The research shows that youth unemployment rates ($F = 20.01$, $p = 0.001583$) and adult internet usage ($F = 19.28$, $p = 0.001422$) affect the situation, but their effects are weaker than those of other factors.

4.3. Machine learning–based analysis of key determinants

The multivariate modeling framework that analyzed crime produced Figure 6 to show the relative significance of chosen socio-economic, demographic, legal, and institutional factors.

Figure 5
ANOVA results for key determinants of crime

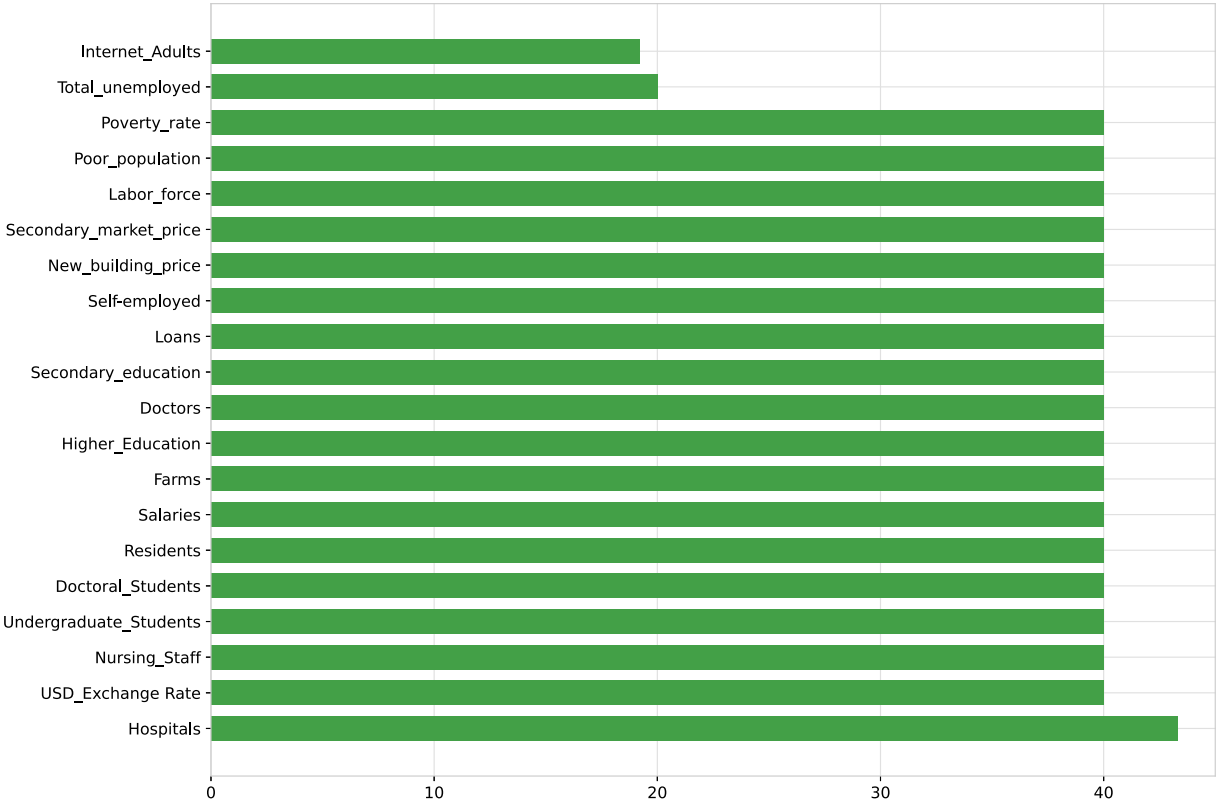
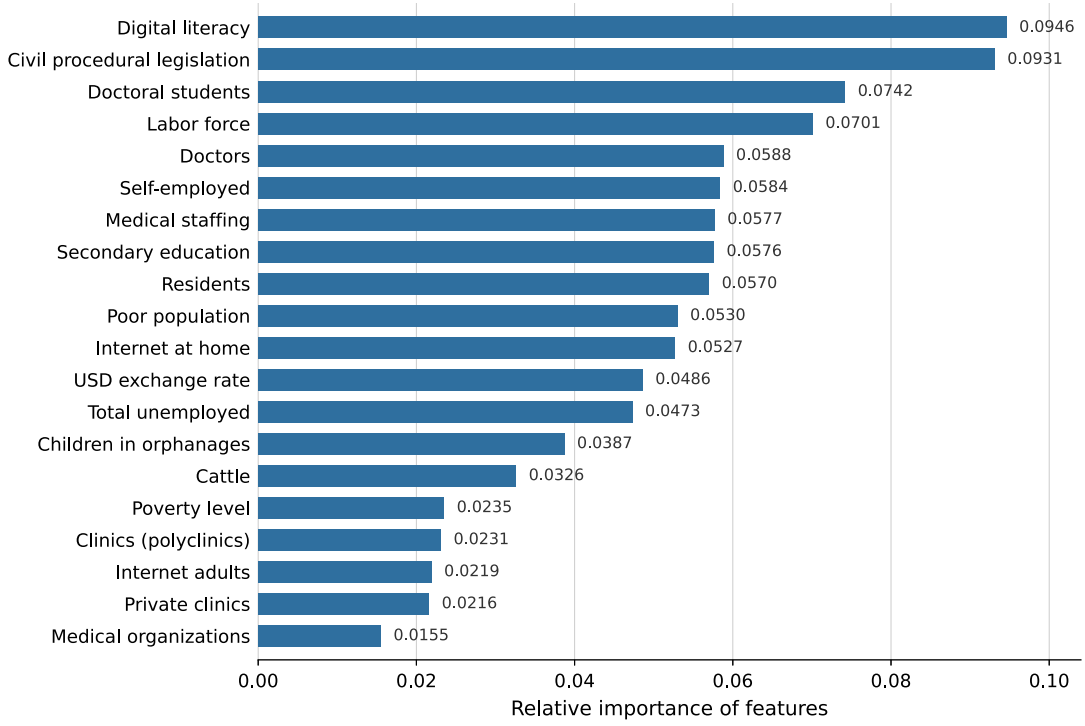


Figure 6
Relative importance of key predictors influencing crime



The research results indicate that digital literacy functions as the primary determining factor that affects results (importance = 0.0946) because it determines the path of crime investigation procedures. The legal framework, together with regulatory rules, establishes the number of pre-trial investigations that follow the structure defined by civil procedural laws (importance = 0.0931). The two factors that show high importance levels include doctoral student numbers (0.0742 importance rating) and labor force population (0.0701 importance rating) because they decrease crime severity by affecting educational standards and work participation rates.

The mid- to high-importance predictors include healthcare-related variables, which consist of doctor numbers (0.0588), self-employed medical and service sector workers (0.0584), medical staff availability (0.0577), and polyclinics (0.0231). The research data show that healthcare system capacity, together with accessibility levels, determines why investigative work patterns differ between different areas. Among the influential variables in the model are secondary education, with an importance value of 0.0576, and resident population size, with an importance value of 0.0570.

The poor population (0.0530), internet access at home (0.0527), USD exchange rate (0.0486), and total unemployment (0.0473) show moderate importance as socio-economic vulnerability indicators, which reveal how macroeconomic instability and household-level deprivation create indirect effects on criminal activities. The two social welfare stress variables, which include children in orphanages (0.0387) and poverty level (0.0235), make some contribution to the model although their impact is minimal.

The final set of indicators, which measure infrastructure and service delivery, shows cattle stock (0.0326), internet use among adults (0.0219), private clinics (0.0216), and medical organizations (0.0155) with lower but significant importance values.

Figure 7 illustrates the agreement between predicted and observed values of crime for three different modeling approaches: (a) Gradient Boosting, (b) Random Forest, and (c) Ridge Regression. The analysis includes three separate panels that show scatter plots between predicted values and observed data points, while the dashed diagonal line shows the perfect prediction line at prediction = observation.

The Gradient Boosting model in panel (a) shows high prediction accuracy because its data points create a dense cluster that runs along the identity line. The model shows excellent predictive capabilities because it produces accurate results with small methodological errors, which helps it detect intricate social-factor, legal-element, and institutional-framework relationships.

Panel (b) shows the results for the Random Forest model. The predictions show similar patterns to the identity line, but they deviate slightly more when the target variable reaches its highest values, which results in small errors between predicted and actual values. The two datasets show strong agreement because ensemble-based tree methods achieve high performance when analyzing crime investigation procedures.

Panel (c) presents the Ridge Regression results, which exhibit a more noticeable dispersion from the identity line compared to the ensemble models. The pattern demonstrates how Ridge Regression generates stable model estimates that combat multicollinearity yet lack the ability to identify complex nonlinear patterns in the data.

The evaluation in Figure 8 compares four predictive modeling techniques through their R^2 performance, which includes Gradient Boosting, Random Forest, Ridge Regression, and SVM. The central R^2 values appear in the bars, but the error bars display the complete range of values that emerged from validation fold tests and repeated evaluation assessments. The horizontal dashed line at $R^2 = 0$ serves as a threshold, which distinguishes models that explain data from models that perform inferior to a baseline mean predictor.

Table 2 presents the predictive results that four modeling methods, including Gradient Boosting, Random Forest, Ridge Regression, and SVM, produced for crime prediction. The model performance is assessed through three evaluation methods, which include R^2 values from the test set and cross-validation R^2 scores with their mean and standard deviation and error metrics consisting of RMSE and MAE.

The results indicate that Gradient Boosting delivers the highest test performance because it reaches an R^2 value of 0.999 while achieving the lowest RMSE of 2756.90 and MAE of 2504.88, which demonstrates its excellent ability to predict data within the sample. Random Forest achieves high test performance (R^2 = 0.989), but its error metrics show higher values than Gradient Boosting, which results in lower absolute prediction accuracy. The test R^2 value from the Ridge Regression model reaches 0.979, but the model produces higher RMSE and MAE values because linear models struggle to handle the complex nonlinear patterns in the data.

Figure 7
Comparison of predicted versus actual values of crime using ensemble and regularized regression models: (a) Gradient Boosting, (b) Random Forest, (c) Ridge Regression

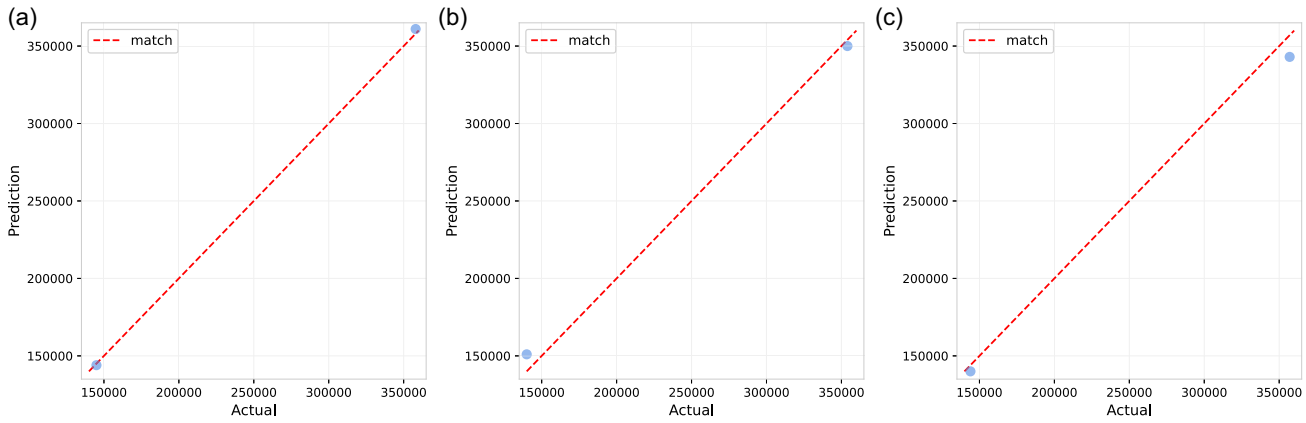


Figure 8
***R*²-based comparison of predictive models for crime**

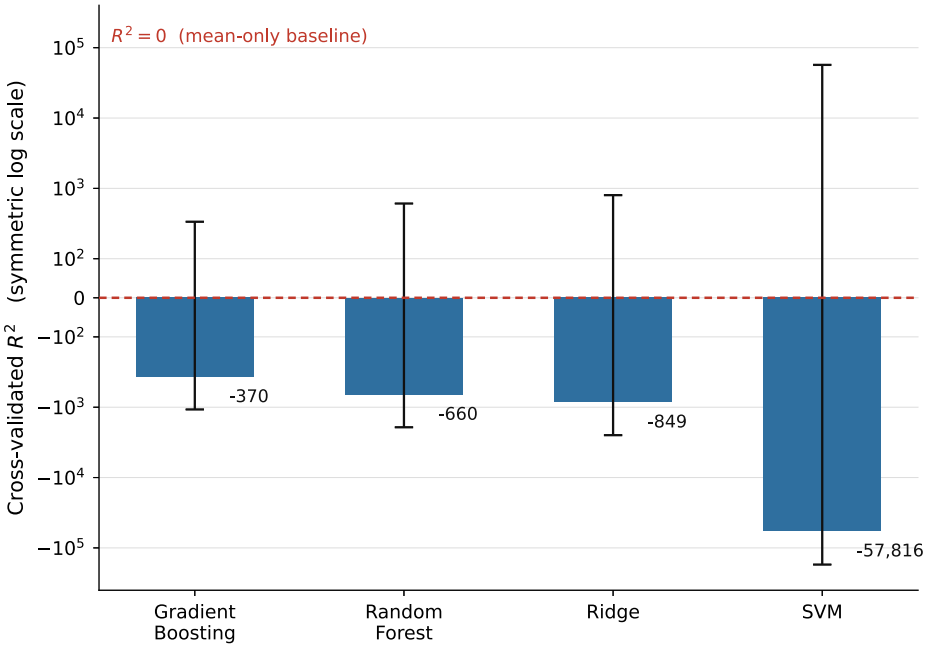


Table 2
Comparative performance metrics of predictive models

Model	Test R^2	CV R^2 (mean)	CV R^2 (std)	RMSE	MAE
Gradient Boosting	0.999380	–370.25	705.78	2756.90	2504.88
Random Forest	0.989087	–659.60	1268.80	11564.94	11550.72
Ridge	0.978542	–848.68	1649.58	16217.21	15292.56
SVM	–0.186492	–57816.07	114801.24	120590.58	110689.34

The SVM shows a negative test R^2 value of − 0.186, while its RMSE and MAE values reach extremely high levels, which proves the model fails to make accurate predictions while consistently missing the actual data points. The R^2 scores from cross-validation show negative means with big standard deviations which indicate that the models demonstrate high instability while failing to generalize well during validation tests. The test R^2 values approach perfection, but cross-validation results show strong negative performance, which indicates possible overfitting or insufficient data or vulnerable response to time-based and structural irregularities in the dataset.

The gap between these two sets of figures calls for caution rather than a reading that takes the headline accuracy at face value. Gradient Boosting, Random Forest, and Ridge Regression all fit the training data almost perfectly, with test R2 values of 0.9994, 0.9891, and 0.9785. On the held-out folds, the same models collapse, returning mean cross-validated R^2 values of −370.25, −659.60, and −848.68, each with a large standard deviation. The SVM fails on both counts (test R^2 = −0.1865, mean CV R^2 = −57,816.07). A negative cross-validated R^2 means that, on data the model did not see during fitting, it does worse than a naive predictor that always returns the sample mean. Pairing an almost perfect in-sample fit with strongly negative out-of-sample scores is a clear signature of overfitting. This is what flexible, high-capacity learners tend to do when seventy candidate predictors are fitted to an aggregated annual series: the ensembles memorize the training points instead of recovering a stable signal, so their apparent accuracy says nothing reliable about predictive skill or about which factors matter most. The substantive conclusions of this study therefore do not rest on these black-box predictors. Inference rests instead on the interpretable, low-dimensional category-based regression and the ANOVA, whose transparent structure suits the available sample far better. The numbers above are best read as in-sample descriptive fit rather than validated forecasts, and the conformal and Bayesian intervals specified in Section 3.7 are the tools we would use to report predictive uncertainty honestly.

4.4. Linear regression evidence on the factors driving crime

The analysis uses two explanatory variables which show the highest linear relationship with crime to build baseline regression models. The analysis selects two variables for baseline regression modeling because they show the most significant negative and positive relationships with crime. On this basis, two simple regression specifications are considered. Model 1 examines the bivariate relationship between crime and the number of doctoral students. Model 2 extends the previous specification through the addition of two new variables, which include the total number of doctoral students and the current medical staff numbers.

The selection of these variables is justified by their exceptionally strong Pearson correlation coefficients: the number of

doctoral students demonstrates a very strong inverse correlation with crime ($r = -0.984$), whereas medical staffing availability exhibits a very strong direct correlation ($r = 0.956$). The research framework includes essential elements that scientists employ to study the relationship between human capital development and institutional capacity and their impact on pre-trial investigation operations.

The bivariate relationships between crime and two essential explanatory variables that show the most significant opposing connections appear in Figure 9. The figure shows the bivariate relationships between crime and two essential variables that demonstrate the strongest opposing correlations between doctoral student enrollment and medical (nursing) staffing. The left section of the graph shows a scatter plot that includes a linear regression line that shows crime distribution based on doctoral student numbers. The right section shows how crime levels relate to medical staff numbers.

Panel (a) demonstrates a strong inverse linear relationship between doctoral student enrollment and crime. The Pearson correlation coefficient shows a strong negative relationship between doctoral student enrollment and pre-trial investigation activities because its value reaches −0.9841. This relationship is statistically significant, as confirmed by a large negative t-statistic ($t = -15.65$) and a p-value effectively equal to zero ($p < 0.001$). The regression line accounts for 96.8% of the total variance.

In contrast, Panel (*b*) reveals a strong direct linear relationship between medical staffing levels and recorded crime. The correlation coefficient shows a strong relationship between the variables ($r = 0.9563$), while the statistical test confirms its significance ($t = 9.25$, $p < 0.001$). The data indicate that medical facilities that employ more staff members will experience higher numbers of pre-trial investigations that get registered in their service areas.

These results indicate that the pursuit of higher education and the accumulation of human capital are statistically associated with lower recorded criminal activity. Medical staffing is the second salient factor and is positively associated with recorded crime; a plausible interpretation, advanced here as a hypothesis rather than a demonstrated mechanism, is that service areas with more extensively staffed and better-equipped facilities may have a greater capacity to detect, record, and report criminal offenses, which would warrant explicit testing in future causal-inference work.

Figure 10 provides a comparative assessment of two linear regression specifications used to model crime. First is a one-factor model based solely on doctoral student enrollment. Second is a two-factor model incorporating both doctoral students and medical staffing. The one-factor and two-factor models have their residual values shown in Panels (a) and (*b*) relative to their predicted values. The predictive ability of these models appears in Panel (*c*) through their R^2 values and RMSE measurements.

Figure 9
Inverse and direct effects of doctoral student enrollment and medical staffing on crime: (a) doctoral student enrollment, (b) medical staffing

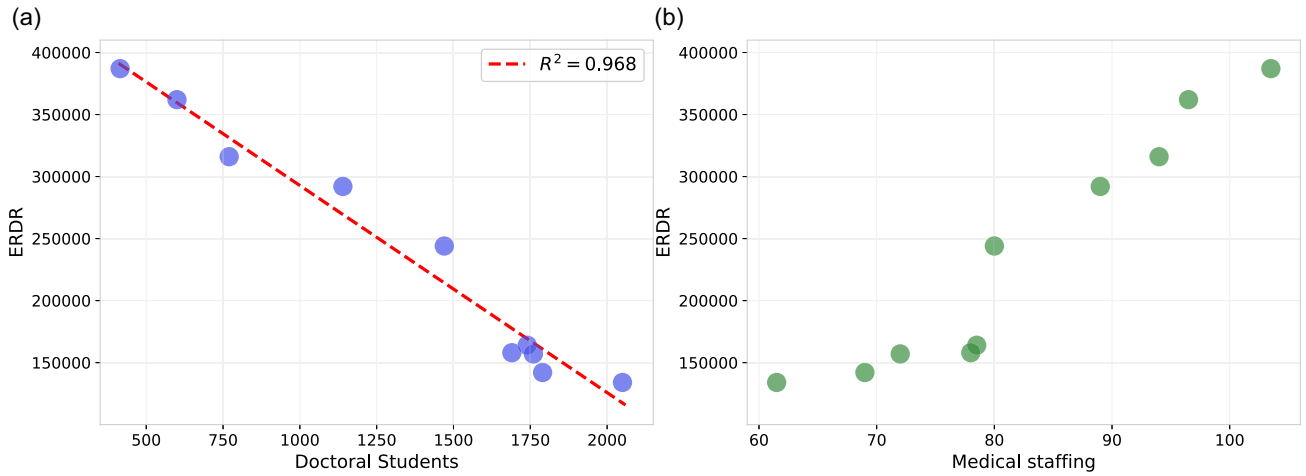
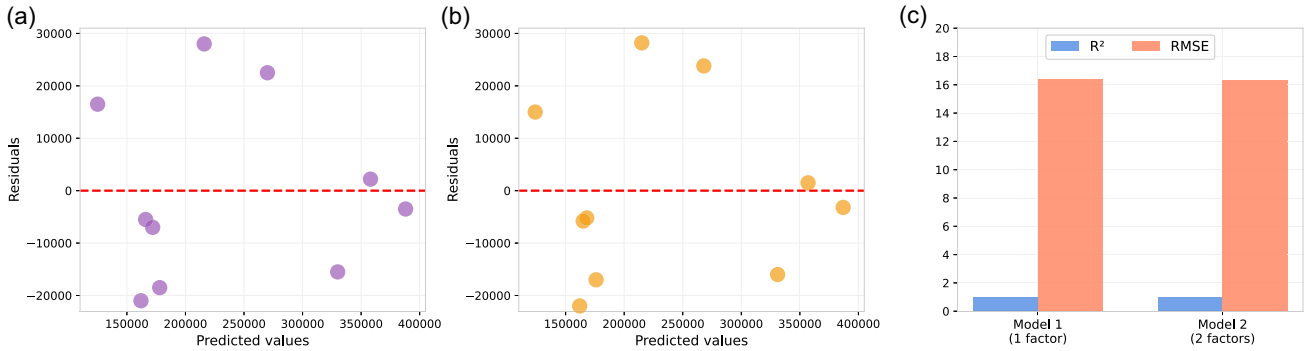


Figure 10
Residual diagnostics and performance comparison of one-factor and two-factor regression models for crime: (a) residuals of the one-factor model, (b) residuals of the two-factor model, (c) comparative performance (R^2 and RMSE) of both models



Panel (a) illustrates the residual distribution for Model 1. The residuals follow a symmetrical pattern that centers around the zero reference line without showing any significant pattern throughout the predicted value spectrum. The results show that the linear model fits the data properly while demonstrating the main negative connection between doctoral student enrollment numbers and crime. The model produces a high explanatory power (R^2 = 0.9684) while showing RMSE and MAE values of 16,462.53 and 14,046.38, respectively. The estimated regression equation, ERDR = 459,780.65 − 166.19 × Doctoral Students, confirms the strong negative effect of advanced education on pre-trial investigation volume.

Panel (b) presents the residual diagnostics for Model 2, which augments the baseline specification by including medical staffing. The residuals in this model remain centered around zero, while their spread is slightly smaller than in Model 1, indicating a modest improvement in model fit. This is supported by a small increase in explanatory power (R^2 = 0.9685) and reductions in both RMSE (16,428.88) and MAE (13,745.73). The corresponding regression equation, ERDR = 501,237.41 − 174.51 × Doctoral_Students − 367.86 × Medical_Staffing, suggests that, conditional on doctoral student enrollment, higher levels of medical staffing are associated with a further decrease in crime within the linear framework.

Panel (c) directly compares the two models using R^2 and RMSE metrics. The two models show equivalent explanatory strength, but the two-factor model produces slightly better results in error-based assessment methods.

Figure 11 shows how the actual crime measurements compare to the two regression model predictions, which run from the start to the end of the research duration. The figure shows the actual crime dynamics together with Model 1 predictions based on doctoral students only and Model 2 predictions based on doctoral students and medical staffing for visual model accuracy and temporal consistency evaluation.

crime series demonstrates a downward trend that begins with strong growth during the first period before its slope becomes less pronounced throughout the entire time period. The two predictive models successfully identified the main trend of the data by following the exact pattern of decrease that was observed in the data. At high crime levels, the model's forecasts closely match actual values. This is especially evident in the early years.

The time series in the middle section shows small differences between the predicted and actual values. In this section, the system undergoes a structural transition. The two models maintain their position within a small margin of error while they show the experimental data points where the curve bends and how the rate of decrease changes. The predicted series shows strong agreement with actual data during the later period when crime levels reach stable lower points.

Model 2 shows a slightly better fit than Model 1 because it better tracks annual changes that occur when the system undergoes changes. The model becomes better at understanding how institutions affect pre-trial investigation work because medical staffing information has been added to the system.

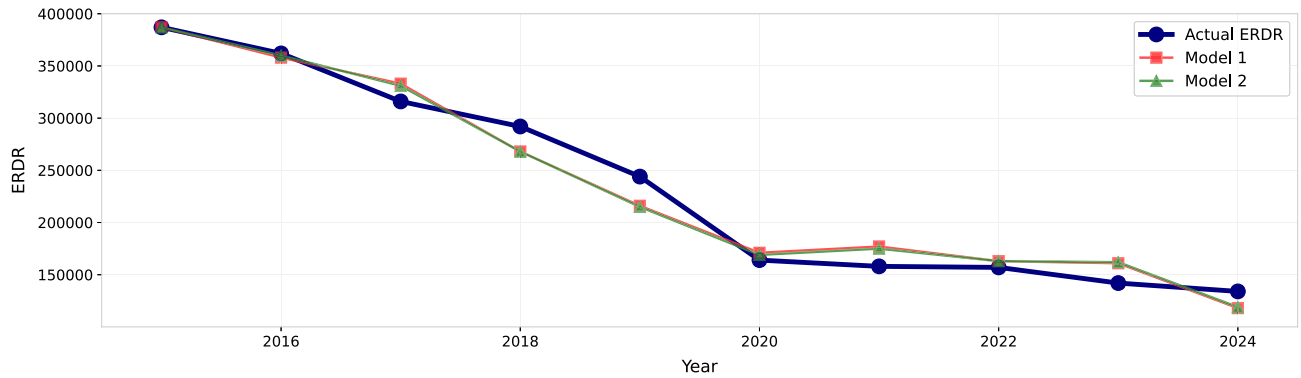
4.5. First category-based dimensionality reduction and regression modeling of crime

The research proposes an experimental system that will transform individual explanatory factors into generalizable elements to decrease the number of variables required for predictive models. To this end, the original set of 70 factors is grouped into eight conceptually coherent categories, followed by a categorized analysis of their influence on crime. The method works to maintain all important data points from the original variables while simplifying models for better understanding and reducing their complexity.

The factors are classified into the following categories:

1) Education (10 factors), including indicators of preschool, secondary, postsecondary, higher education, master's degree attainment, and additional education-related measures.
2) Healthcare (8 factors), encompassing medical organizations, hospitals, clinics, private clinics, hospital beds, and other indicators of healthcare capacity.
3) Economy (15 factors), covering construction activity, salary levels, real wage index, USD and RUB exchange rates, and other macroeconomic indicators.
4) Employment (6 factors), including labor force size, number of employees, self-employed population, unemployment rate, total unemployed, and related measures.
5) Social Sphere (9 factors), comprising poor population, poverty level, disadvantaged families, orphanages, children in orphanages, and additional social vulnerability indicators.
6) Legal System (11 factors), incorporating variables related to the criminal sphere, criminal law, criminal procedure code legislation, penal code legislation, administrative offenses, and other legal-institutional measures.
7) Agriculture (8 factors), including cultivated areas, wheat production, cattle, small cattle, horses, and other agricultural indicators.

Figure 11
Time-series comparison of actual and predicted values of crime



8) Digitalization (3 factors), represented by digital literacy, internet use among adults, and internet access at home.

The classification method enables researchers to develop eight composite categorical indices, which combine their respective factors into individual indices for Education and Healthcare and Economy and Employment and Social Sphere and Legal System and Agriculture and Digitalization. The feature space dimensionality decreases from 70 individual variables to 8 categorical indices, which results in an 88.6% reduction of dimensionality.

The scatter plots in Figure 12 show how crime relates to three aggregated indices through time while displaying both time-based and cross-time relationships between these indices and Digitalization and Healthcare and Employment. The study data points show yearly observations, which use color gradients to indicate the time progression from start to finish of the research period.

The Digitalization index shows a strong negative relationship with crime according to Panel (a). As digitalization increases over time, crime rates show a systematic decline. This demonstrates a close link between improved digital infrastructure and digital literacy and the reduction in the volume of pre-trial investigations. The regression line shows a steady negative linear pattern that extends from the beginning to the end of the study period.

Panel (b) presents the relationship between the Healthcare index and crime. Similar to Digitalization, the Healthcare index shows a strong inverse association with crime, indicating that higher healthcare capacity and accessibility are linked to lower levels of registered pre-trial investigations. The data show a negative pattern, but the scatter points near the regression line indicate multiple variables that affect this connection.

Panel (c) depicts the association between the Employment index and crime. The data show a direct negative connection between employment condition enhancements and crime value decreases. The Employment index shows slightly more unstable patterns throughout its data series, but it maintains a steady decline from start to finish.

The Pearson correlation coefficient shows the strength and direction of linear relationships between aggregated categorical indices and crime in Table 3. The table presents R values that demonstrate both the direction of relationships and provide a basic evaluation of their strength.

The research findings show digitalization maintains the strongest connection with crime ($R = -0.952$) through a very strong negative correlation. The research shows that people who have better digital skills and who have access to the internet will need fewer pre-trial investigations. The research shows that healthcare ($R = -0.860$), employment ($R = -0.850$), economic development ($R = -0.838$), and education ($R = -0.811$) maintain strong negative relationships with crime, which indicates that better institutional capacity, stable labor markets, and educated populations help decrease police investigations.

The legal system category demonstrates an inverse relationship ($R = -0.644$), which shows that better legal and regulatory frameworks lead to decreased crime levels, yet this connection remains weaker than the relationships with socio-economic and digital elements. The agricultural indicators show a moderate inverse relationship with crime dynamics with a correlation of − 0.425.

In contrast, the social sphere index is the only category displaying a positive relationship with crime ($R = 0.283$), although the strength of this association is weak. The direct relationship between social vulnerability levels and pre-trial investigation work exists, but the model only explains a limited amount of total variation.

Regression models were constructed using the aggregated categorical indices identified as significantly correlated with crime. The modeling strategy was based on a stepwise approach. Categories were added based on the strength of their correlation with crime to assess their impact on the model's explanatory power.

The performance results from Table 4 show how different regression models built with combined categorical variables affect the explanation of crime variation. The model performance is assessed through three evaluation metrics, which include the coefficient of determination (R^2) and the error-based measures of RMSE and MAE.

The Digitalization index in Model 1 explains 90.6% of crime variation through its strong predictive ability ($R^2 = 0.906$). The model contains a basic structure, which results in substantial prediction errors because it fails to account for multiple structural elements (RMSE = 28,301; MAE = 19,769).

The Healthcare index becomes a new variable that Model 2 adds to its basic model structure. The model shows better performance after this addition because R^2 reaches 0.926 and RMSE decreases to 25,093. The MAE shows a small increase, but the total decrease in squared error indicates better explanatory power.

The predictive power of Model 3 becomes significantly better when the Employment index is added to the model. The model explains more than 95% of the variance in crime ($R^2 = 0.952$) and shows significantly better error performance (RMSE = 20,362; MAE = 17,635) because it combines labor market conditions with digital and healthcare elements.

Model 4 contains the most detailed specification. This model combines six aggregate categories, including digitalization and healthcare, as well as employment and the economy, education, and the legal system. The model that combines various social and institutional factors produces the best prediction results ($R^2 = 0.962$) and minimal forecasting mistakes (RMSE = 18,125; MAE = 14,075) for understanding crime.

Figure 13 compares the actual crime series with the predictions of the four category-based regression models (Models 1–4) over the period 2015–2024.

The actual crime series shows a decreasing pattern from start to finish of the research duration, which starts with a significant drop during the first years before it continues to decrease until it reaches equilibrium in subsequent years. The four regression models show excellent agreement with the data points. They correctly identify both the direction and size of this trend throughout most of the observation period.

The Digitalization index serves as the basis for Model 1, which shows the overall declining trend but produces significant errors when economic structures undergo major transformations during the middle section of the data series. Model 2 analysis, which uses Digitalization and Healthcare indices together, produces better results for crime level prediction during years when institutional capacity effects start to appear more strongly. The Employment category in Model 3 produces better prediction results because it uses real data points that exist during both the time when employment rates decline and when employment rates remain stable. The fourth model, which combines six categories (Digitalization, Healthcare, Employment, Economy, Education, and Legal System), shows the greatest match with the actual crime series. The model produces its most accurate yearly data modifications.

Figure 12
Temporal and cross-sectional associations between key aggregated indices and crime: (a) index digitalization, (b) index healthcare, (c) index employment

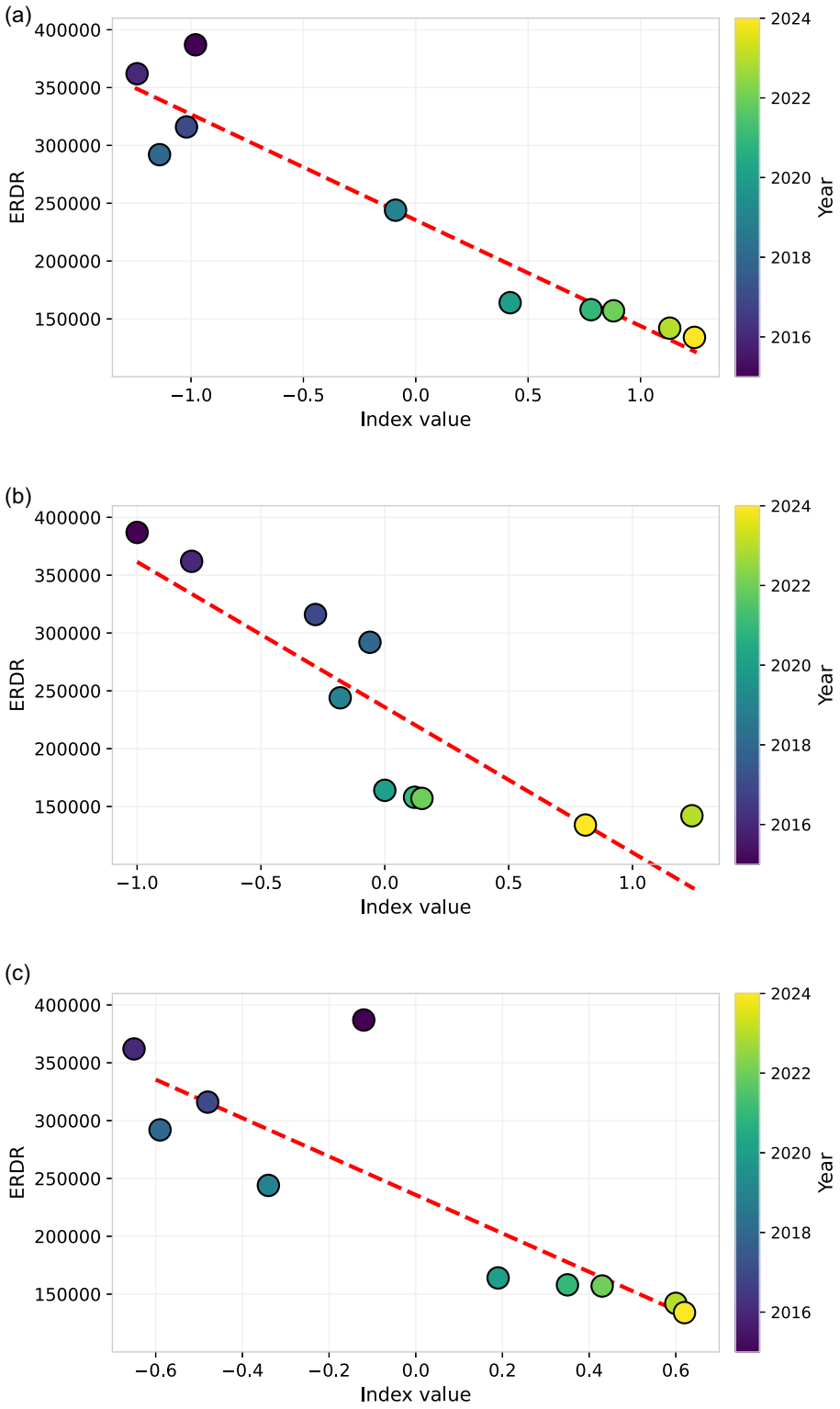


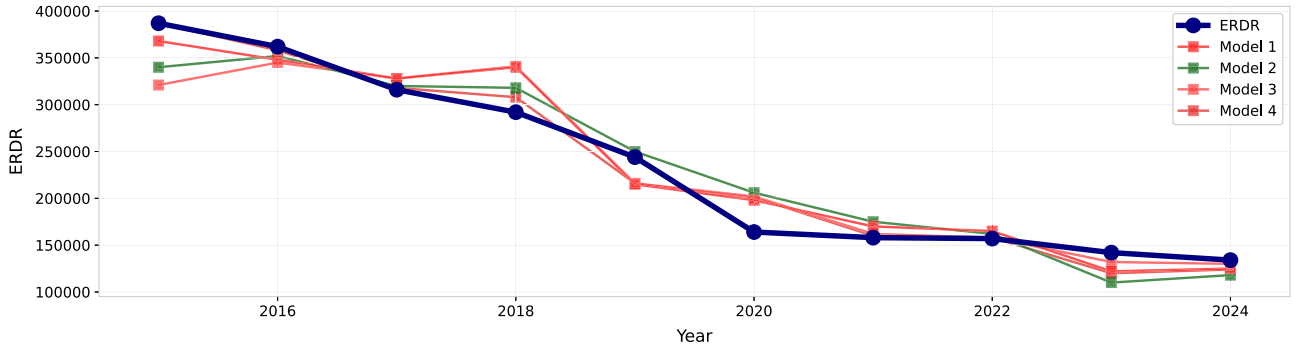
Table 3
Ranking of categories by strength of connection with crime

Index	*R*	Direction	Relation
Digitalization	–0.952	Inverse	Very strong
Healthcare	–0.860	Inverse	Very strong
Employment	–0.850	Inverse	Very strong
Economy	–0.838	Inverse	Very strong
Education	–0.811	Inverse	Very strong
Legal_system	–0.644	Inverse	Strong
Agriculture	–0.425	Inverse	Moderate
Social_sphere	0.283	Direct	Weak

Table 4
Performance evaluation of category-based regression models for crime

Model	Index	R^2	RMSE	MAE
Model 1 (1 category)	Digitalization	0.906480	28301.175505	19769.013802
Model 2 (2 categories)	Digitalization, Healthcare	0.926479	25093.294131	20354.355449
Model 3 (3 categories)	Digitalization, Healthcare, Employment	0.951589	20362.225082	17635.355888
Model 4 (6 categories)	Digitalization, Healthcare, Employment, Economy, Education, Legal System	0.961643	18124.789428	14075.076188

Figure 13
Temporal comparison of actual crime values and predictions from category-based regression models



The factor categorization method offers a number of methodological advantages. It reduces the number of variables from 70 individual factors to 8 aggregated indices. This reduces dimensionality. This transformation simplifies the model structure. However, the reduction preserves all the essential information from the original variables. The reduced feature space also helps reduce the risk of overfitting.

Based on these results, we will expand the categorization structure in a subsequent experiment. We will increase the number of aggregated indices to 10. This will help us understand whether a more detailed categorization can further improve explanatory power and forecasting accuracy.

4.6. Second category-based dimensionality reduction and regression modeling of crime

Dividing the Social_Sphere category into three distinct categories allowed us to achieve better forecasting results and more precise results. Prior to this improvement, the Social_Sphere index demonstrated a weak and unstable relationship with crime, with a correlation coefficient of + 0.283. Significantly stronger and more interpretable relationships emerged after disaggregation: Poverty $r = -0.912$; Family Distress $r = +0.856$; Family Relationships $r = +0.167$.

The Poverty subcategory shows a powerful negative connection with crime ($r = -0.912$), placing lower poverty alongside lower pre-trial investigation activity. Whether anti-poverty programs would themselves reduce crime is a separate question and a priority for the causal analysis outlined in Section 5.

Family Distress indicators, which include dysfunctional families and orphan numbers, demonstrate a strong positive relationship with crime ($r = +0.856$). The study shows that family instability growth and rising social vulnerability create conditions that result in higher pre-trial investigation work. The magnitude of this relationship highlights the critical importance of preventive and social support interventions targeted at at-risk families.

The Family Relationships subcategory demonstrates a weak positive association with crime through its variables. This

subcategory includes marriage and divorce, as well as substance abuse (r = +0.167). These factors have a limited impact on crime dynamics at the pooled data level. This makes them less important than indicators of poverty and family distress.

Table 5 shows how aggregate indices and social subcategories relate to crime using Pearson correlation coefficients. The table highlights significant heterogeneity in both the magnitude and sign of correlations across different areas.

The inverse relationship between crime and Digitalization shows the strongest connection because their correlation coefficient reaches −0.952. The data show that pre-trial investigation volumes decrease when digital literacy and internet accessibility levels reach higher levels. The data show a strong negative relationship with poverty because the correlation coefficient (r = −0.912) indicates that poverty reduction efforts will result in major reductions of crime. The analysis shows that healthcare (r = −0.860), employment (r = −0.850), economy (r = −0.838), and education (r = −0.811) maintain strong to very strong negative relationships, which demonstrate how institutional capacity, labor market stability, economic development, and human capital resources reduce investigative work.

Family Distress demonstrates a strong positive correlation with crime (r = +0.856). Family instability and social vulnerability directly impact the amount of work required for pre-trial investigation. The research results indicate that social policies that prevent problems from happening should focus on safeguarding families who face dangerous situations. The Legal System index shows a weak negative relationship (r = −0.644), which indicates that legal-institutional conditions have a small yet significant effect on crime.

The research shows that Agriculture (r = −0.425) maintains a moderate negative relationship with crime and Family Relationships (r = +0.167) demonstrates a small positive connection to crime. The results of the study show that agricultural efficiency indicators combined with typical family patterns do not significantly influence differences in crime.

Table 6 presents results from regression models that used different data aggregation levels and social subcategory information to predict crime variation.

The Digitalization index in Model 1 explains 90.6% of crime variance because it has an R^2 value of 0.906. Model 1 contains only the main variables, which leads to significant forecasting errors. This is because it does not take into account various social factors and institutional elements (RMSE = 28.301; MAE = 19.769).

Model 2 adds Poverty and Healthcare indices. The model performance improved when additional dimensions were included because R^2 reached 0.932 while RMSE and MAE decreased to 24,090 and 17,980, respectively. The research data reveal that economic poverty, together with healthcare service availability and digital technology implementation, provides a complete understanding of the situation.

Model 3 contains the complete specification, which uses all eight evaluation criteria including Digitalization, Poverty, Healthcare, Family Distress, Employment, Economy, Education, and Legal System. The model achieves superior explanatory power (R^2 = 0.976), and it produces the lowest prediction errors (RMSE = 14,401; MAE = 12,201), which leads to improved predictive accuracy. The research findings show that using specific social risk factors together with economic and institutional data produces the best results for understanding crime patterns.

Table 5
Correlation ranking of aggregated indices and social subcategories influencing crime

Index	Correlation
Digitalization	–0.952
Poverty	–0.912
Healthcare	–0.860
Family Distress	0.856
Employment	–0.850
Economy	–0.838
Education	–0.811
Legal System	–0.644
Agriculture	–0.425
Family Relationships	0.167

4.7. ANOVA-based assessment of category significance before and after social sphere disaggregation

ANOVA was conducted to assess the statistical significance of aggregated categories in explaining variation in crime, both before and after the disaggregation of the social sphere. ANOVA results prior to disaggregation indicate that several categories exhibit statistically significant effects on crime. Healthcare shows a very strong effect (F = 40.15, p < 0.001). Then followed Economy, Employment, and Digitalization (each F = 11.10, p = 0.007), and Education (F = 9.49, p = 0.011). But Agriculture, Legal System, and the aggregated social sphere do not demonstrate statistically significant effects (p > 0.05). This indicates the limited explanatory power of treating social factors as a single composite category.

The statistical model became more streamlined after disaggregation. Healthcare produces the most significant effect, which determines the final result because it has the highest significant effect (F = 40.15, p < 0.001). The Poverty category shows the highest level of significance because it has an F value of

Table 6
Regression model performance with expanded aggregated and social subcategory indices influencing crime

Model	Index	Count	R^2	RMSE	MAE
Model 1	Digitalization	1	0.906480	28301.175505	19769.013802
Model 2	Digitalization, Poverty, Healthcare	3	0.932242	24089.728200	17980.174937
Model 3	Digitalization, Poverty, Healthcare, Family Distress, Employment, Economy, Education, Legal System	8	0.975784	14401.326752	12201.190841

40.01, which is less than 0.001. The research categories Economy, Employment, and Digitalization show significant statistical relationships ($F = 11.10$, $p = 0.007$), while Education also demonstrates statistical significance ($F = 9.49$, $p = 0.011$). The research data show that Family Distress reaches statistical significance at a 0.037 probability level when researchers divide their data into separate groups ($F = 5.47$, $p = 0.037$). The research data indicate that family instability, together with social vulnerability factors, produces the observed patterns of crime.

In contrast, agriculture, the legal system, and family relations remain statistically insignificant ($p > 0.05$). These factors have a relatively weak impact on crime within the analyzed framework.

4.8. Discussion

The research conducted a complete assessment that investigates various elements at different points to identify which variables affect crime. Clear patterns in crime investigations, as well as significant nonlinear and system-wide relationships, were identified. These relationships are difficult to detect using standard case-by-case analysis methods.

The analysis of all analytical layers shows digitalization functions as a primary factor that reduces criminal activities. The analysis shows that digital literacy, internet access, and household connectivity are negatively correlated with crime. This confirms the high significance of the features in the machine learning model and the significant explanatory power of the category-based regression analysis. The research results confirm what theory predicts about digital inclusion because it provides people with better access to information, work possibilities, public services, and social connections, which decreases their chances of committing crimes.

Healthcare systems need complete assessments of their capacity to understand how it affects their operations. Healthcare indicators show a strong negative correlation with crime. This indicates that improved access to healthcare services and infrastructure development is associated with lower volumes of pre-trial investigations. The medical staffing data show a different pattern. It shows a positive relationship with recorded crime. The difference between these two indicators demonstrates how institutional capacity influences crime reporting. Crimes are more effectively detected and documented in regions with well-developed healthcare systems. This explanation is supported by the results of regression and time-series analysis. The expansion of the healthcare system leads to a decrease in the crime reporting rate. Crime rates remain low over long periods. At the same time, the healthcare system improves problem detection over short observation periods.

The system operates at its most stable level through human capital development, which achieves its highest point during advanced educational stages. The enrollment data for doctoral students and higher education programs demonstrate negative relationships with crime, which appear both in statistical correlation tests and linear regression models that predict most of the measured variables. The study's findings support criminological theories. This theory posits that education acts as a preventative strategy to reduce crime. Education improves employment prospects and shapes social values. Education makes criminal activity more costly. The education system maintains its stability in results across various model specifications. The education system functions independently of short-term economic fluctuations.

The economy needs two vital components, which consist of stable labor market conditions and proper macroeconomic stability. The study demonstrates that workers who are employed, receive actual wages, participate in the labor market, run their own businesses will experience decreased crime levels, but poverty measures will create positive and significant relationships. The main methodological achievement of the study is the successful division of the social sphere into three separate components: Poverty, Family Distress, and Family Relationships. Before the analysis took into account social factors, there was no statistical relationship between them. We then identified the relationship between poverty reduction and family distress as opposing factors in the analysis. The research shows that poverty exists in direct opposition to crime because financial backing and rising income levels decrease police investigation activities, but family problems including being an orphan and having unstable family relationships lead to increased criminal behavior.

The machine learning results confirm these findings by showing how different factors create complex relationships with each other and which factors generate the most substantial effects. The in-sample predictive results of Ensemble models including Gradient Boosting and Random Forest show near-perfect accuracy, which proves that ERDR data contains complex nonlinear relationships with its influencing factors. The negative cross-validation R^2 values are why we do not lean on raw predictive accuracy, and why the policy reading rests on the category-based regression and the ANOVA instead.

Dimensionality reduction method is the main analytical approach in this study. The model retains its explanatory power by reducing 70 individual variables to 10 composite indices. The use of composite indices improves both the stability and interpretability of the model. The models that unite digitalization with poverty, healthcare, employment, education, economic conditions, and legal institutions produce the most accurate predictions while minimizing their prediction errors, thus validating an integrated systems approach to crime prevention.

The research results indicate that crime reduction policies need to focus on digital inclusion, education spending, poverty reduction, family assistance programs, healthcare system advancement, and institutional development. There is no strong statistical correlation between the agricultural sector and individual legal indicators. This suggests that broad-based socio-economic strategies will yield better results than targeted legal or business-focused approaches.

The boundaries of what these results can support, and the data and identification strategies needed to move past them, are set out in Section 5 rather than rehearsed here.

The analysis shows that crime emerges through a network of digital elements, social systems, economic systems, and institutional frameworks. The study's findings show that crime prevention is most effective through planned infrastructure construction. Independent prevention methods are ineffective. However, they contribute to evidence-based policymaking.

5. Limitations and Future Research Directions

The principal limitation is the design itself. Every result rests on aggregated national time series for 2015–2025, which give only 10–11 yearly observations against 70 candidate explanatory factors, and a sample of that shape leaves the analysis correlational. The Pearson coefficients, ANOVA contrasts, regression slopes, and machine learning feature importances reported above all measure statistical association rather than causal effect. Three concerns follow from this. Because the design offers no causal identification, even the strongest patterns we report, such as

the −0.952 correlation for the Digitalization index or the eight-index model that accounts for 97.6% of crime variance, cannot be read as the effect of any policy. Confounding is the next, and it is serious: most socio-economic, demographic, and institutional indicators drift together over a single decade, so any of them, or an unobserved driver behind all of them, could be doing the work we attribute to a particular index. The opposite signs of the Healthcare index and medical staffing associations are a reminder that more than one mechanism can sit inside a single aggregate. Non-stationarity then compounds both problems, because national series tend to share trends that flatter in-sample fit. The clearest symptom is the distance between near-perfect in-sample accuracy (Gradient Boosting test R2 = 0.9994) and a strongly negative cross-validated R2 (mean −370.25), and that distance is the central reliability caveat of this work. These limits are why we keep every empirical statement associational and admit causal language only where a mechanism is named as a hypothesis.

What would it take to move from association toward mechanism? Mostly better data and stronger identification, and almost everything we propose here waits on the disaggregated regional and monthly records we do not yet hold; none of the figures in this paper depend on it. Causal identification is the first priority. Structural equation modeling would let us state a hypothesized causal graph among the composite indices and test it against the data [1]; instrumental-variable estimation would help where endogeneity and omitted variables are likely; and difference-in-differences designs could use policy or legislative changes that reach different regions at different times. When a single intervention can be dated, a Bayesian structural time-series counterfactual, along the lines of the causal impact approach, gives a cleaner quasi-experimental estimate of its effect. Identification of this kind pairs naturally with dynamic models. System-dynamics and agent-based simulations would turn the static associations into scenario experiments, letting feedback, delay, and threshold behavior among digitalization, employment, poverty, and investigative activity play out under competing intervention paths.

Several of these ambitions are, at bottom, requests for finer data. Replacing the national aggregate with regional and monthly series would make room for spatial point-process models of individual crime events and for geographically weighted regression, which can show where the index-crime relationships tighten or loosen across the country [19, 20]; the same resolution would support vector autoregression (VAR) and autoregressive integrated moving average (ARIMA) models of the temporal dynamics and lead-lag structure that annual totals erase [23, 24]. This is also the point where the present study meets our companion point-level spatiotemporal model, which already carries the event geography the aggregated design lacks. Better data would settle a question we cannot touch now, namely whether the patterns travel. Every parameter here comes from one national series, so refitting and validating the indices across regions, and ideally across jurisdictions, is the only honest way to learn how far they generalize.

A handful of refinements matter whatever the scale of the data. Point predictions from a sample this small mean little without calibrated uncertainty, so future estimates should report it directly, whether as Bayesian posterior intervals [14] or as distribution-free conformal prediction, which has the advantage of valid finite-sample intervals in exactly this small-n regime [31]. The linearity built into the composite-index regressions also deserves a second look: GAMs and threshold specifications can loosen the constant-slope assumption and, more usefully, find the points of diminishing or accelerating association where an intervention would matter most. Given more observations, nonlinear dimensionality reduction using t-SNE, UMAP, or autoencoders might recover latent structure among the 70 factors that an equal-weight linear average cannot, without discarding the interpretability the composite indices were designed to protect. The eventual payoff is prescriptive rather than descriptive. Once effects and their uncertainty can be estimated credibly, multi-objective and constrained optimization could search the digitalization, education, healthcare, employment, and anti-poverty levers for intervention mixes that stay cost-effective under real budget and feasibility limits. Each of these steps points the same way: from the interpretable but correlational baseline reported here toward evidence identified, uncertainty-aware, and spatially resolved enough to guide policy. For now, they are proposed work, and the conclusions of this study remain deliberately associational.

6. Conclusion

This study presents a systemic analysis of socio-economic factors and their impact on the dynamics of crime investigations in Kazakhstan in the period 2015–2025. Classical statistical methods were integrated with modern machine learning methods. A categorical approach to factor dimensionality reduction was also applied. This helped deepen the methodological understanding of crime investigation dynamics.

Digitalization is the most influential factor associated with reducing the level of crime risks. Indicators related to digital literacy, internet access, and household internet connectivity demonstrate a strong inverse relationship with the volume of pre-trial investigations. Correlation analysis, regression modeling, and feature importance analysis using machine learning also confirm this. The pattern fits pathways such as improved access to information, government services, and labor market opportunities, though digital integration remains a candidate correlate rather than a demonstrated mechanism.

The development of higher education acts as another stabilizing force. Enrollment in higher education and obtaining a doctorate degree show a strong negative correlation with criminality. These results support criminological theories that education is a long-term preventative factor for criminal behavior. Education enhances social integration and increases the opportunity costs of recidivism. Employment stability, good economic conditions, high labor force participation rates, good wages, and favorable macroeconomic indicators significantly reduce the activity of investigative bodies.

An analysis of the healthcare system reveals an interesting effect on crime. Expanding health infrastructure and improving accessibility reduce the overall rate of recorded crime. Increasing the number of healthcare personnel is positively correlated with recorded crime rates. This is likely due to increased institutional capacity to detect and report crimes, rather than an increase in crime per se. This suggests that factors should be functionally interpreted differently.

A key methodological contribution of the study lies in the disaggregation of the social sphere. Social vulnerability as a single composite index may mask fundamentally different effects. Once decomposed into poverty, family distress, and family relationships, clear and opposing relationships emerge. Poverty shows a strong inverse association with crime, which makes anti-poverty policy a plausible lever and a target for the causal work described

in Section 5. Family problems show a strong positive relationship with crime. The role of family instability and social vulnerability is critical in stimulating investigative activity. Family relations indicators show only weak associations with crime. This factor has lower explanatory power.

70 individual variables were condensed into 8–10 composite indices. This category-based dimensionality reduction approach improves the interpretability and stability of the model. It also preserves its explanatory power. Regression models built using aggregate indices explain up to 97.6% of crime variance. These models outperform simpler specifications. They also reduce the risk of overfitting observed in multivariate machine learning models. Ensemble methods such as Gradient Boosting and Random Forest achieve near-perfect accuracy on a sample basis. But performance is inconsistent when cross-validating.

Crime prevention and investigation workload management cannot rely solely on legal or law enforcement measures. Instead, effective strategies require a comprehensive policy. This policy must integrate: digital inclusion, education, job creation, poverty reduction, healthcare development, and targeted family support programs. Crime does not arise as an isolated legal phenomenon. It arises from the interaction of socio-economic, institutional, and technological systems.

Section 5 gives the limitations of the design, together with the panel, regional and longer-horizon data that would be needed to move beyond it; the ERDR itself does not yet reach that horizon. Even so, the approach proposed here offers a scalable foundation for evidence-based crime prevention policy and for the analytical use of national administrative crime registries.

Funding Support

This work was supported by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Grant No. AP32715381.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data are available from the corresponding author upon reasonable request.

Author Contribution Statement

Alibek Barlybayev: Conceptualization, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Mira Kaldarova:** Methodology, Validation, Formal analysis, Investigation, Resources, Data curation. **Rakhat Yerezhepov:** Software, Validation, Resources, Data curation, Visualization. **Nurlan Botabayev:** Validation, Formal analysis. **Madi Tugunbayev:** Methodology, Resources, Supervision, Funding acquisition. **Gulmira Shakmetova:** Writing – review & editing, Software, Validation, Data curation, Visualization.

References

[1] Yaseen, H. F., & Man, S. (2025). Criminality trends through socio-economic factors: A two decade panel ARDL analysis of upper-middle-income nations (2000–2021). *Crime & Delinquency*. Advance online publication.https://doi.org/10.1177/00111287251397995

[2] Hong, M., Cho, E., & Chang, J. (2025). Crime behavior prediction using process mining and deep learning: Toward a scalable script analysis framework. *KSII Transactions on Internet and Information Systems*, *19*(9), 3167–3183. https://doi.org/10.3837/tiis.2025.09.018

[3] Sultanov, N. (2016). The procedural model of criminal prosecution in the modern criminal procedure legislation of the Republic of Kazakhstan. *Journal of Advanced Research in Law and Economics*, *7*(5), 1179–1186.

[4] Mukhamadieva, G. N., Zhanibekov, A. K., Apsimet, N. M., & Alimkulov, Y. T. (2025). Integration of artificial intelligence into criminal procedure law and practice in Kazakhstan. *Laws*, *14*(6), 98. https://doi.org/10.3390/laws14060098

[5] Albahli, S., & Albattah, W. (2023). Crime type prediction in Saudi Arabia based on intelligence gathering. *The Computer Journal*, *66*(8), 1936–1948. https://doi.org/10.1093/comjnl/bxac053

[6] Hegland, T. A. (2025). Medicaid coverage of psychological treatment prevents crime. *Health Economics*, *34*(9), 1560–1577. https://doi.org/10.1002/hec.4976

[7] Henderson, H., Bourgeois, J. W., Smith, S., & Ferguson, C. J. (2024). Economic correlates of crime: An empirical test in Houston. *Journal of Criminal Justice*, *95*, 102306. https://doi.org/10.1016/j.jcrimjus.2024.102306

[8] Ersöz, F., Ersöz, T., Marcelloni, F., & Ruffini, F. (2025). Artificial intelligence in crime prediction: A survey with a focus on explainability. In *IEEE Access*, *13*, 59646–59674. https://doi.org/10.1109/ACCESS.2025.3553934

[9] Alsubayhin, A. A., Alzahrani, B., & Ramzan, M. S. (2024). Review of crime prediction through machine learning. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(5s), 273–281.

[10] Salmi, V., & Kivivuori, J. (2006). The association between social capital and juvenile crime: The role of individual and structural factors. *European Journal of Criminology*, *3*(2), 123–148. https://doi.org/10.1177/1477370806061967

[11] Cheung, N. W. T., & Cheung, Y. W. (2008). Self-control, social factors, and delinquency: A test of the general theory of crime among adolescents in Hong Kong. *Journal of Youth and Adolescence*, *37*(4), 412–430. https://doi.org/10.1007/s10964-007-9218-y

[12] Ingilevich, V., & Ivanov, S. (2018). Crime rate prediction in the urban environment using social factors. *Procedia Computer Science*, *136*, 472–478. https://doi.org/10.1016/j.procs.2018.08.261

[13] Cadena Urzúa, P., & Letelier Saavedra, L. (2018). Factores determinantes de los delitos de mayor connotación social en la región metropolitana. Análisis en base a un modelo de regresión logística [Determining factors of high social connotation crimes in the metropolitan region. A logistic model analysis]. *Política Criminal*, *13*(26), 1170–1189. https://doi.org/10.4067/S0718-33992018000201170

[14] Kester, Q.-A., & Afoma, E. J. (2021). Crime predictive model in cybercrime based on social and economic factors using the Bayesian and Markov theories. In *2021*

International Conference on Computing, Computational Modelling and Applications, 165–170. https://doi.org/10.1109/ICCMA53594.2021.00034

[15] Liebe, U., & Schwitter, N. (2021). Explaining ethnic violence: On the relevance of geographic, social, economic, and political factors in hate crimes on refugees. *European Sociological Review*, *37*(3), 429–448. https://doi.org/10.1093/esr/jcaa055

[16] Butt, U. M., Letchmunan, S., Hassan, F. H., & Koh, T. W. (2024). Leveraging transfer learning with deep learning for crime prediction. *PLoS One*, *19*(4), e0296486. https://doi.org/10.1371/journal.pone.0296486

[17] Sam, T. H., Isa, K., Tong, P. K., Palpanadan, S. T., Hong, H. C., & Bhutasang, N. (2024). Exploring the significant factors affecting the social crimes of Malaysian students. *Journal of Ecohumanism*, *3*(4), 2567–2581. https://doi.org/10.62754/joe.v3i4.3776

[18] Doğan, F. İ. (2024). Socioeconomic, migration, and political upheaval factors in crime: Applying the social disorganization theory in Turkey. *Sociology Compass*, *18*(11), e70012. https://doi.org/10.1111/soc4.70012

[19] Jen, I.-C. (2025). Examining the crime concentration: Spatial patterns of burglary crime in Taipei, Taiwan. *International Criminology*, *5*(1), 84–100. https://doi.org/10.1007/s43576-025-00163-7

[20] Kennedy, A. R., Kennedy, L. W., Giménez-Santana, A., & Caplan, J. M. (2025). Mapping food access and related risk factors in forecasting social disorder and crime. *Applied Spatial Analysis and Policy*, *18*(1), 4. https://doi.org/10.1007/s12061-024-09608-4

[21] Wu, S., Yang, D., Ge, W., & Chen, B. (2023). Risk prediction model of bank telecommunication fraud based on XGBoost. In *International Conference on Cyber Security, Artificial Intelligence, and Digital Economy, 12718*, (p. 127182H). https://doi.org/10.1117/12.2681646

[22] Dutta, K., Parui, S., Shaw, S., Biswas, S., & Das, S. (2025). Crime analysis and management system. *SN Computer Science*, *6*(7), 799. https://doi.org/10.1007/s42979-025-04319-0

[23] Huang, Y., Yang, B., Ren, X., Lu, Y., Lan, M., & Gong, X. (2025). Spatio-temporal cokriging crime predictions using social media data: A multi-type case study in San Jose, California. *Computational Urban Science*, *5*(1), 72. https://doi.org/10.1007/s43762-025-00233-6

[24] Fan, Y., Hu, X., & Hu, J. (2025). Research on a crime spatiotemporal prediction method integrating Informer and ST-GCN: A case study of four crime types in Chicago. *Big Data and Cognitive Computing*, *9*(7), 179. https://doi.org/10.3390/bdcc9070179

[25] Neelambaram, B., Kumar, M. S., Ganesh, D., Raghuram, A. V., ShyamSundar, M., & Uppu, T. P. (2024). Analysis and prediction of crime hotspots using machine learning with stacked generalization approach. In *2024 International Conference on Computational Intelligence for Green and Sustainable Technologies*, 1–6. https://doi.org/10.1109/ICCIGST60741.2024.10717502

[26] Shohan, F. T., Akash, A. U., Ibrahim, M., & Alam, M. S. (2025). A data-driven approach for predicting crime occurrence using machine learning models. *International Journal of Data Science and Analytics*, *20*(6), 5755–5774. https://doi.org/10.1007/s41060-025-00799-8

[27] Ganesan, R., & Ravichandran, S. (2024). Performance analysis for crime prediction and detection using machine learning algorithms. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(14s), 348–355.

[28] Fu, L., Afizan, Marjani., Jazri, M., Saepudin, H., S, & Muslih, M. (2025). Machine learning based prediction of burglary crime probability. In *2025 International Conference on Metaverse and Current Trends in Computing*, 1–8. https://doi.org/10.1109/ICMCTC62214.2025.11196503

[29] Singh, M., Jindal, G., Oberoi, A., & Dhangar, R. (2024). Improving crime detection through Geo-MDA: A hybrid linear regression approach in data mining. *International Journal of Performability Engineering*, *20*(8), 469–477. https://doi.org/10.23940/ijpe.24.08.p1.469477

[30] Lersch, K. M., & Hart, T. C. (2014). Environmental justice, lead, and crime: Exploring the spatial distribution and impact of industrial facilities in Hillsborough County, Florida. *Sociological Spectrum*, *34*(1), 1–21. https://doi.org/10.1080/02732173.2014.857184

[31] Berk, R. A., Kuchibhotla, A. K., & Tchetgen, E. (2024). Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. *Sociological Methods & Research*, *53*(4), 1629–1675. https://doi.org/10.1177/00491241231155883

How to Cite: Barlybayev, A., Kaldarova, M., Yerezhepov, R., Botabayev, N., Tugubayev, M., & Shakhmetova, G. (2026). Identifying Structural Drivers of Crime Using Statistical Inference and Machine Learning Models. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE62029210