UDC 004.8:343.97(574)
IRSTI 28.17.26

https://doi.org/10.55452/1998-6688-2025-22-1-84-93

[1]*Bitanov A.,
Master's student, ORCID: 0009-0001-5156-9972,
*e-mail asan.bitanov@gmail.com

[1]Kazakh-British Technical University, Almaty, Kazakhstan

FORECASTING THE NUMBER OF CORRUPTION CRIMES IN KAZAKHSTAN: A MACHINE LEARNING APPROACH

Abstract

This study aims to predict the number of corruption crimes in Kazakhstan using machine learning methods. The research is based on official monthly crime statistics collected from the Legal Statistics Portal, specifically the Report Form No. 3-K, which records corruption-related offenses since 2016 [3]. Three regression models were applied: k-Nearest Neighbors (kNN), Extreme Gradient Boosting (XGBoost), and Linear Regression. Model performance was assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) score. The findings indicate that Linear Regression achieved the highest predictive accuracy ($R^2 = 1.000$), followed by XGBoost ($R^2 = 0.9977$) and kNN ($R^2 = 0.9333$). These results suggest that machine learning models can effectively forecast corruption crime trends. This study highlights the potential of machine learning in corruption crime prediction. Future research can explore additional predictive features, alternative machine learning models, and real-time data integration to enhance forecasting accuracy.

Key words: Corruption, Machine Learning, Predictive Analytics, k-Nearest Neighbors, XGBoost, Linear Regression, Kazakhstan, Crime Prediction, Anti-Corruption, Regression Models.

Introduction

Corruption is one of the most pressing and widespread issues in the Republic of Kazakhstan. According to Transparency International's 2023 Corruption Perceptions Index (CPI), Kazakhstan received a score of 39, ranking 93rd out of 180 countries [1]. Although this score remains low, it represents an improvement compared to the previous year, reflecting the impact of the Anti-Corruption Policy Concept for 2022–2026, introduced by the government [2]. This policy framework outlines key strategies for combating corruption, incorporating international best practices and legal reforms.

Despite ongoing efforts, corruption continues to undermine public trust in government institutions, hinder economic growth, and exacerbate social inequality. It creates an environment where wealth and power are disproportionately concentrated, leaving marginalized groups more vulnerable. To address this challenge, Kazakhstan has increasingly focused on digitalization as a means of enhancing transparency and accountability [4]. Key initiatives include open data policies, E-government services, blockchain applications, and Big Data analytics to improve governance and public sector efficiency.

While these measures contribute to reducing corruption, they primarily focus on reactive approaches, identifying and addressing cases of corruption after they occur. To enhance predictive capabilities, this study explores the application of machine learning (ML) techniques for forecasting corruption crime trends. There is currently a lack of predictive models tailored to Kazakhstan's corruption landscape. This study aims to bridge this gap by developing and evaluating machine learning models to forecast corruption crime trends based on historical crime reports. ML algorithms can analyze historical crime reports, detect hidden patterns, and provide data-driven insights that may help policymakers take proactive measures in combating corruption.

Literature Review

A huge frame of empirical and anecdotal proof speaks to the deep and terrible effects of corruption. According to current United Nations statistics, for example, global corruption fees the worldwide financial system over 3.6 trillion USD annually [5].

Oxford Insights identifies synthetic intelligence as "the following frontier in anticorruption" because of its functionality to hit upon styles in datasets which might be too huge for human analysis. By leveraging AI to pinpoint elements of interest, humans can concentrate on examining specifics and investigating potential instances of misuse, fraud, or corruption [6].

Top-down tactics are primarily based totally at the perception that establishments are fashioned with the aid of using legal guidelines created with the aid of using political leaders. As a result, those anti-corruption projects attempt to result in extrade with the aid of using new legal guidelines, regulations, and approaches inside public administration. In contrast, bottom-up tactics view establishments as growing certainly via social norms, customs, traditions, beliefs, and values inside society. These efforts focus attention on information the cultural and societal context to perceive and help present projects that goal to lessen corrupt practices. This method mainly is based on the involvement of lively civil society groups and reporters who can act as watchdogs [7].

Another observes explores the phenomenon from a predictive analytics standpoint, the usage of present-day system getting to know strategies to pinpoint the maximum sizable predictors of corruption notion via more suitable nonlinear fashions with excessive predictive accuracy. In this study's multiclass classification modeling framework, the Random Forest algorithm (an ensemble-type machine learning method) emerged as the most accurate prediction and classification model, followed by Support Vector Machines and Artificial Neural Networks [8].

Another paper focuses on reducing out-of-sample prediction error and ensuring strong generalization to future unseen data. The authors demonstrate that corruption can be predicted with high accuracy by using a limited set of variables that are readily accessible to policymakers. Additionally, they provide a straightforward rule for identifying areas likely to experience corruption episodes [9].

Currently, records mining strategies are taken into consideration powerful techniques for detecting fraud and corruption in diverse sectors, together with credit score card transactions, financial institution accounts, and telecommunications. This paper explores a revolutionary method the use of genetic algorithms to enhance the detection rate [10].

Anohter study examined the impact of governmental AI adoption on financial regulatory intensity in China, revealing significant findings across 30 provinces and municipalities from 2012 to 2022. Governmental AI adoption for financial regulation significantly strengthens financial regulatory intensity. The institutional environment and government transparency have respective promotional and restraining influences on this process. Further tests reveal a nonlinear impact of governmental AI adoption for financial regulation on regional financial regulatory intensity [11].

In another article, they explore to which extent the use of AI in the public sector impacts these core governance functions. Findings from the review of a sample of 250 cases across the European Union, show that AI is used mainly to support improving public service delivery, followed by enhancing internal management and only in a limited number assist directly or indirectly policy decision-making. The analysis suggests that different types of AI technologies and applications are used in different governance functions, highlighting the need to further in-depth investigation to better understand the role and impact of use in what is being defined the governance "of, with and by AI" [12].

Another study aims to provide empirical evidence and insights into public perceptions concerning the use of AI in local government services. Their methodological approach involves collecting data via an online survey from the residents of three major Australian cities–i.e., Sydney, Melbourne, Brisbane–and Hong Kong (n = 850), and performing statistical analyses. They found that: (a) Ease of using AI is significantly and positively influenced by attitude towards AI; (b) Attitude towards

AI significantly and positively influences perceived usefulness of AI in local government services; (c) AI is seen useful in resource management and to improve delivery of service, reduction of cost to provide urban-service, improvement of public safety, and monitoring the effectiveness of strategies to manage environmental crisis, and; (d) AI is more positively perceived by Australians in comparison to Hong Kongers, indicating the impact of contextual and cultural differences [13].

To enhance the efficacy of corruption prevention and control in grass-roots government, this study introduces the concept of data platform management and integrates it with the "5W" (Who, What, When, Where, Why) analysis framework. The research is motivated by the observation that existing studies on corruption prevention primarily concentrate on the formulation of laws and regulations, neglecting the potential improvement in actual effectiveness through the utilization of data platforms and analytical frameworks [14].

To what extent can artificial intelligence techniques help distribute public spending to increase GDP, decrease inflation and reduce the Gini index? In order to respond to this question, this article proposes an algorithmic approach on how budget inputs (specific expenditures) are processed to generate certain outputs (economic, political, and social outcomes). The authors use the multilayer perceptron and a multiobjective genetic algorithm to analyze World Bank Open Data from 1960 to 2019, including 217 countries. The advantages of implementing this type of decision support system in public expenditures allocation arise from the ability to process large amounts of data and to find patterns that are not easy to detect, which include multiple non-linear relationships [15].

Another study unified efforts across social and technical disciplines by first conducting an integrative literature review to identify and cluster 69 key terms that frequently co-occur in the multidisciplinary study of AI. Then they build on the results of this bibliometric analysis to propose three new multifaceted concepts for understanding and analysing AIbased systems for government (AI-GOV) in a more unified way: (1) operational fitness, (2) epistemic alignment, and (3) normative divergence. Finally, we put these concepts to work by using them as dimensions in a conceptual typology of AI-GOV and connecting each with emerging AI technical measurement standards to encourage operationalization, foster cross-disciplinary dialogue, and stimulate debate among those aiming to rethink government with AI [16].

In another study they reviewed digital public services, crowdsourcing platforms, whistleblowing tools, transparency portals, distributed ledger technology, and artificial intelligence. They reviewed evidence on both the anti-corruption effectiveness of ICTs and their misuse for corruption. Research indicates that ICT play a significant role in supporting anticorruption efforts by enhancing public scrutiny. ICT contributes to these efforts by enabling the reporting of corruption, fostering transparency and accountability, and facilitating greater citizen participation and interaction with government entities. Through these mechanisms, ICT has been shown to strengthen oversight and improve the responsiveness of government to public concerns [17].

Another study approaches the phenomenon from the predictive analytics perspective by employing contemporary machine learning techniques to discover the most important corruption perception predictors based on enriched/enhanced nonlinear models with a high level of predictive accuracy. Specifically, within the multiclass classification modeling setting that is employed herein, the Random Forest (an ensemble-type machine learning algorithm) is found to be the most accurate prediction/classification model, followed by Support Vector Machines and Artificial Neural Networks [18].

Another research aims to explore different legal Open Data; in particular, they explored the data set of the National AntiCorruption Authority in Italy on public procurement and the judges' sentences related to public procurement, published on the website of the Italian Administrative Justice from 2007 to 2022. The first goal was to train machine learning models capable of automatically recognizing which procurement has led to disputes and consequently complaints to the Administrative Justice, identifying the relevant features of procurement that correspond to certain anomalies. The second goal was to develop a recommender system on procurement to return similar procurement to a given one and find companies for bidders, depending on the procurement requirements [19].

Materials and Methods

This research was done with Python and special libraries, such as Pandas, SK-learn and Matplotlib. The Figure 1 demonstrates the methodology of this research.

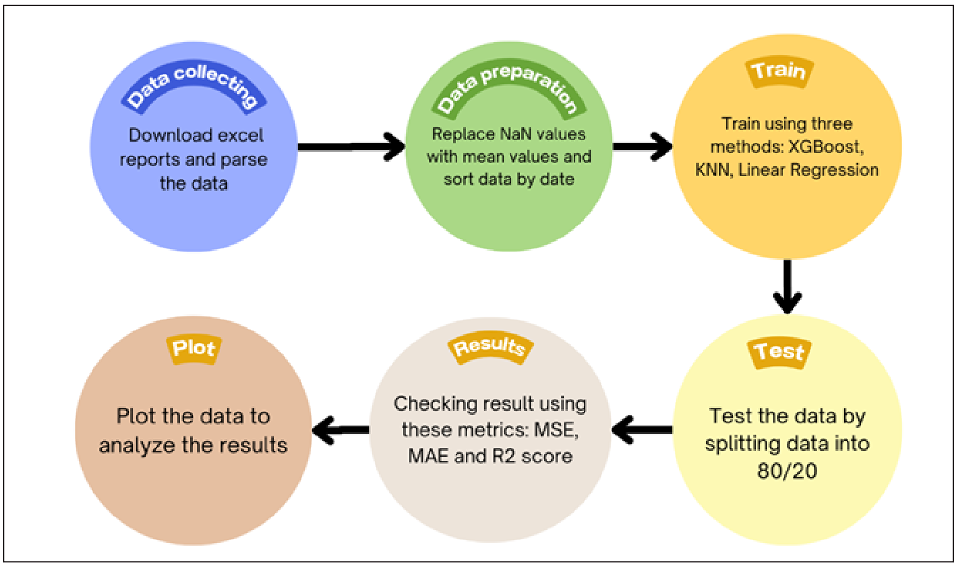
Data collecting
Download excel reports and parse the data
Data preparation
Replace NaN values with mean values and sort data by date
Train
Train using three methods: XGBoost, KNN, Linear Regression
Test
Test the data by splitting data into 80/20
Results
Checking result using these metrics: MSE, MAE and R2 score
Plot
Plot the data to analyze the results

Figure 1 – Methodology figure

The first step is data collecting. To collect data, we use open data from Legal Statistic web page. The Internet portal of legal statistics was created by the Committee on Legal Statistics and Special Accounts of the General Prosecutor's Office of the Republic of Kazakhstan in order to inform citizens about the state of crime in the country and in its individual regions, as well as to provide interactive services [20].

The report we are interested in is called Report form No. 3-K about corruption crimes. We manually download the monthly reports from November 2016 to March 2024. Thus, we have 81 data rows. There are many different items in the reports. For our research we use only the columns that are presented in Table 1.

Table 1 – The example of one row from the dataset. The numbers of corruption crimes in November of 2023

Column	Value
Date	11-2023
Total corruption crimes	1397
Minor corruption crimes	52
Moderate corruption crimes	286
Grave corruption crimes	1000
Especially grave corruption crimes	59
Embezzlement or embezzlement of entrusted property	95
Fraud	144
Legalization (laundering) of money	1
Economic smuggling	0
Raiding	0

Continuation of table 1

Abuse of official authority	115
Abuse of power or official authority	16
Illegal participation in business activities	5
Obstruction of legitimate business activities	4
Taking bribes	367
Giving bribes	550
Mediation in bribery	62
Official forgery	26
Inaction in the service	6
Abuse of power	3
Abuse of power (duplicate)	3
Inaction of power	0

All data is collected in one CSV file. There are NaN (Not a Number) values in the dataset. We replace all the NaN values with mean. This step ensures that the dataset remains complete and suitable for analysis. We maintain the integrity of the data while minimizing the impact of missing entries on our next steps. In our dataset dates are presented in string data type. We convert them to int. The next step is dividing the dataset into features (denoted as X) and target variable (denoted as y). For the target variable we use the date column as it showed the best results. After defining the intention variable, we break up the records into schooling and attempting out gadgets the usage of 80/20 proportion. To enhance the performance of machine learning algorithms. Standardization of features involves subtracting the mean and scaling to unit variance. This step guarantees that every function contributes similarly to the version, stopping capabilities with large scales from disproportionately influencing the version outcomes.

k-Nearest Neighbors (kNN) Regression In kNN regression, the predicted value $\hat{y}$ for a query point *x* is the average of the values of its *k* nearest neighbors:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_j$$

where y_j are the target values of the *k* nearest neighbors of *x*.

Extreme Gradient Boosting (XGBoost) XGBoost builds an ensemble of wooden sequentially, in which each tree tries to correct the errors of the previous wooden. The intention characteristic to lower is:

$$L(\phi) = \sum_{i=1}^{n} l\left(y_j, \hat{y}_j^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k)$$

where: – l it is a discriminable convex loss function that part the contrast between prediction yˆ (t) i and the target yi. – Ω(f) is a regularization term to prevent overfitting.

Linear Regression Linear Regression fashions the connection among the structured variable y and unbiased variables X with the aid of using becoming a linear equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

where: – y is the dependent variable. – β0 is the intercept. – β1, β2, . . . , βp are the coefficients for each independent variable x1, x2, . . . , xp. – ϵ is the error term. The coefficients are found by minimizing the sum of squared errors (SSE):

$$SSE = \sum_{i=1}^{n} (y_j - \hat{y}_j)^2$$

where yˆi is the predicted value of the i-th observation.

Results and Discussion

To evaluate the performance of the regression models, we used three standard metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) score. These metrics assess the accuracy of predictions by measuring the deviation between actual and predicted values. The results for each model are presented in Table 2.

Table 2 – Performance comparison of kNN, XGBoost, and Linear Regression models

Regression Method	MAE	MSE	R2 Score
kNN Regressor	153.2000	39863.6659	0.9333
XGBoost Regressor	29.6092	1367.3824	0.9977
Linear Regression	0.6382	0.6382	1.0000

Mean Absolute Error (MAE) measures the average magnitude of errors between predicted and actual values. Lower MAE values indicate better performance. Mean Squared Error (MSE) calculates the squared differences between predicted and actual values, penalizing larger errors more significantly. A lower MSE suggests higher model accuracy. R-squared (R^2) score quantifies the proportion of variance in the dependent variable that is explained by the independent variables. An R^2 value closer to 1 indicates a better model fit.

To further analyze model performance, we compare the actual and predicted values using graphical representations for each regression method.

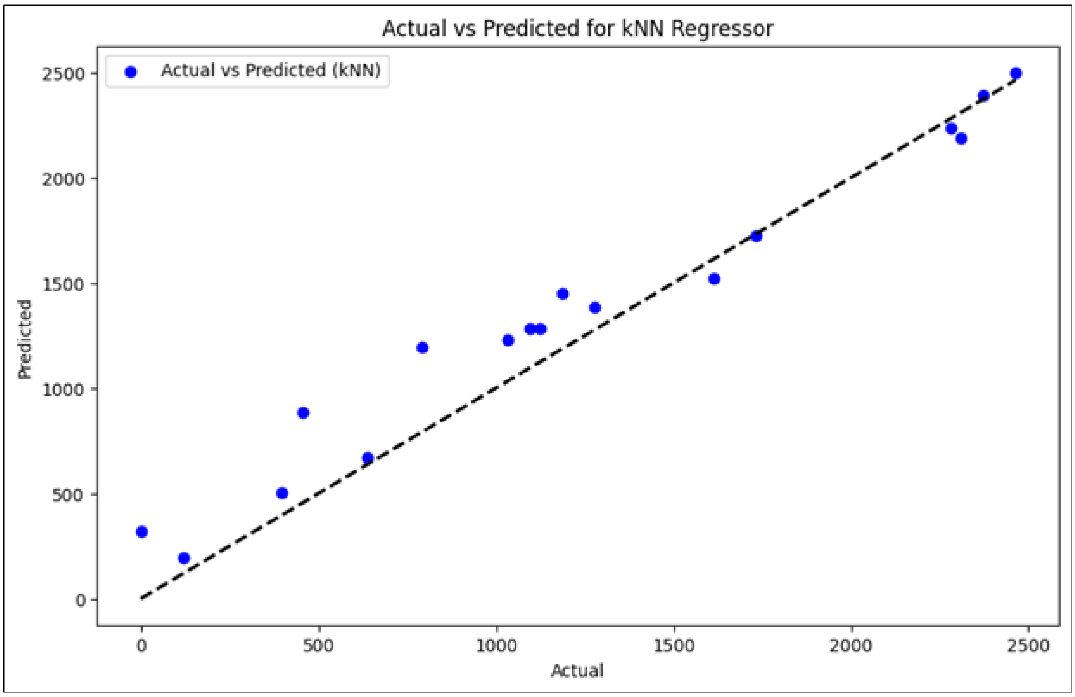

Figure 2 – Comparison of the actual data and predicted using KNN method

As shown in Figure 2, the kNN model captures the general trend in corruption crime numbers but struggles with extreme values and outliers. This suggests that while kNN can identify patterns in the data, it may not be the most reliable method for forecasting corruption crimes.

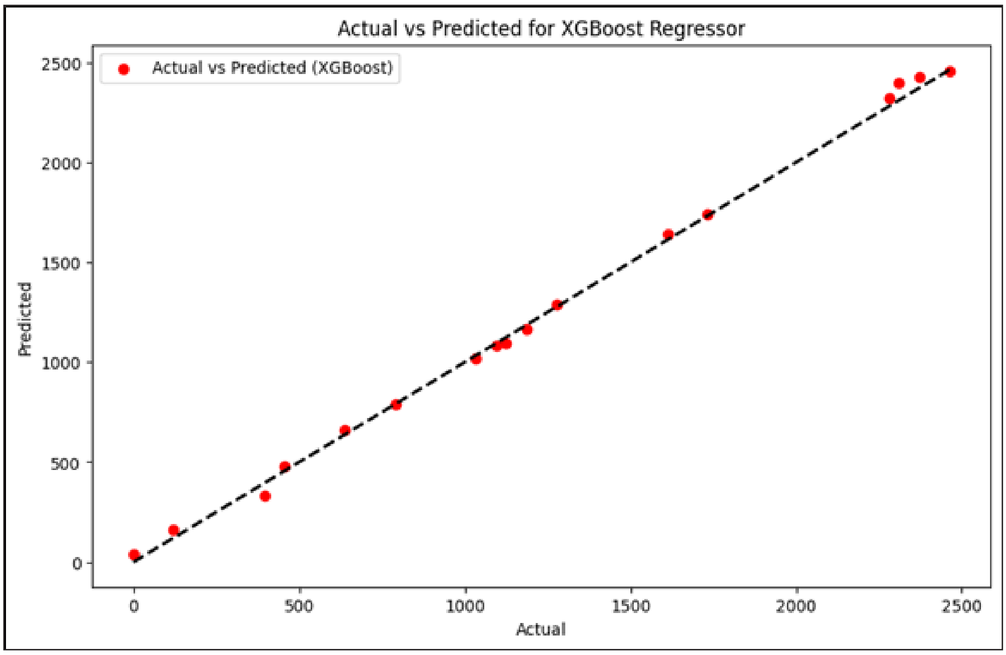

Figure 3 – Comparison of the actual data and predicted using XGBoost method

Figure 3 illustrates the performance of the XGBoost model, which closely aligns with actual values across most of the dataset. Although deviations are still present, particularly for extreme values, XGBoost demonstrates superior resilience in handling variations compared to kNN.

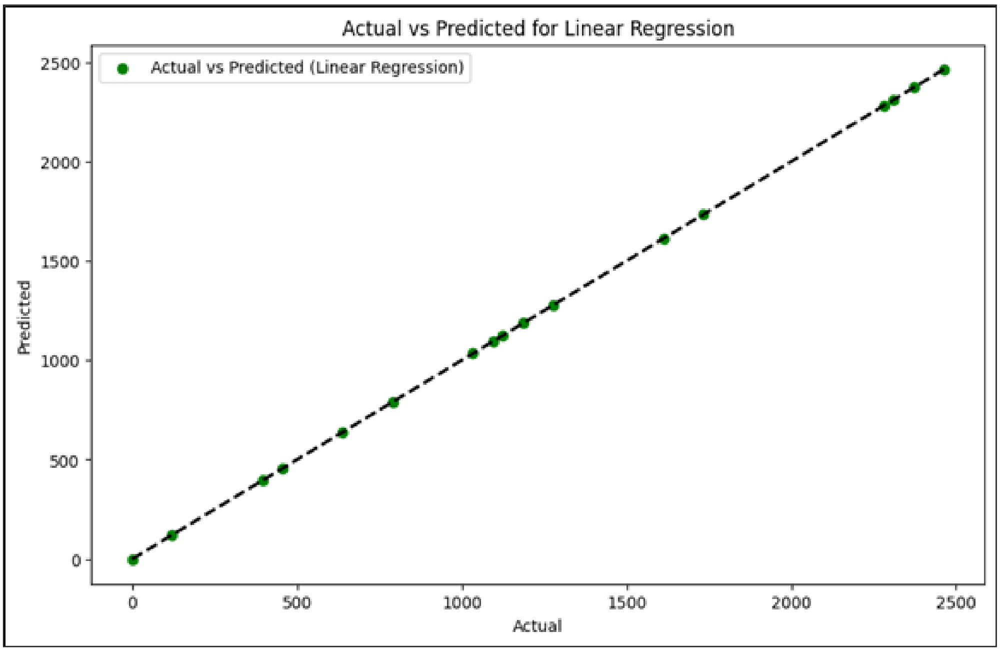

Figure 4 – Comparison of the actual data and predicted using Linear Regression method

The results in Figure 4 indicate that Linear Regression provides the most accurate predictions, with minimal deviations from actual corruption crime numbers. By fitting a linear equation to the dataset, the model successfully generalizes patterns and delivers highly precise forecasts.

Linear Regression demonstrated the highest predictive accuracy ($R^2 = 1.000$), indicating that the relationship between the independent variables and corruption crime numbers is well captured by a linear model. The extremely low error values (MAE = 0.6382, MSE = 0.6382) suggest that the dataset follows a relatively linear pattern, making this model the most suitable for forecasting corruption crime trends in Kazakhstan.

XGBoost also performed exceptionally well ($R^2 = 0.9977$), achieving low error values compared to kNN. This suggests that while non-linear relationships exist in the data, they are not significant enough to surpass the performance of a simple linear model. XGBoost's ability to handle complex patterns and outliers contributed to its high accuracy.

The kNN model showed the weakest performance ($R^2 = 0.9333$), with the highest MAE and MSE values. Although it captured general trends, its inability to handle extreme values and its sensitivity to variations in data distribution led to a lower predictive accuracy.

The results indicate that corruption crime trends in Kazakhstan follow a predominantly linear pattern, which explains why Linear Regression outperformed the other models. This suggests that corruption crime numbers are influenced by stable and predictable factors, rather than complex non-linear interactions. However, the strong performance of XGBoost highlights the presence of some non-linear elements in the dataset, which may require further investigation.

The relatively poor performance of the kNN model suggests that local variations and neighborhood-based approaches may not be the best strategy for predicting corruption crimes. This could be due to the high dimensionality of the dataset or the lack of localized clusters in corruption crime trends.

These findings suggest that future research should explore feature selection techniques to determine which variables contribute most to corruption crime trends. Additionally, incorporating external socio-economic indicators and testing hybrid models could provide deeper insights into the factors driving corruption in Kazakhstan.

Conclusion

This study examined the application of machine learning regression methods–k-Nearest Neighbors (kNN), XGBoost, and Linear Regression – for forecasting the number of corruption crimes in Kazakhstan. The models were evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) score, allowing for a comprehensive assessment of their predictive capabilities.

The results indicate that Linear Regression achieved the highest predictive accuracy, followed by XGBoost, which demonstrated strong resilience in handling variations in corruption crime trends. Although the kNN model effectively captured general patterns, it exhibited limitations in predicting extreme values and outliers. These findings suggest that machine learning techniques can serve as effective tools for forecasting corruption crime trends, with Linear Regression and XGBoost being particularly well-suited for this task.

Future research will focus on optimizing model parameters to enhance predictive performance and improve robustness. Additionally, identifying relevant features that influence corruption levels in the country will be a key direction. By incorporating socio-economic indicators, policy changes, and external factors, future studies aim to uncover the underlying causes of corruption trends rather than just predicting their occurrence. The integration of real-time data streams could further improve the responsiveness of corruption crime forecasting models.

This research underscores the potential of machine learning in supporting anti-corruption efforts by providing data-driven insights into corruption crime trends. By leveraging predictive analytics, policymakers and law enforcement agencies can adopt more proactive strategies to combat corruption, enabling more informed decision-making and targeted interventions.

REFERENCES

1 Transparency International, 2023 corruption perceptions index: Explore the [Online]. Available: https://www.transparency.org/en/cpi/2023/. [Accessed: Mar. 11, 2025] .

2 Government of Kazakhstan, Anti-corruption policy concept of the Republic of Kazakhstan for 2022–2026 [Online]. Available: https://www.gov.kz/memleket/entities/anticorruption/activities/17527?lang=en. [Accessed: Mar. 11, 2025].

3 Prosecutor General of Kazakhstan, Order of the prosecutor general of Kazakhstan on approval of the report form No. 3-K. [Online]. Available: https://adilet.zan.kz/rus/docs/V1600014126. [Accessed: Mar. 11, 2025].

4 Ministry of Digital Development, Innovations and Aerospace Industry of Kazakhstan, The digitalization against corruption [Online]. Available: https://www.gov.kz/memleket/entities/mdai/press/article/details/139736?lang=ru. [Accessed: Mar. 11, 2025].

5 Ash E., Galletta S., and T. Giommoniю A machine learning approach to analyze and support anti-corruption policy [Online]. Available: https://www.RePEc.org. [Accessed: Mar. 11, 2025].

6 U4 Anti-Corruption Resource Centreю Artificial intelligence – a promising anti-corruption tool in development settings? [Online]. Available: https://www.u4.no/publications/artificial-intelligence-a-promising-anti-corruption-tool-in-development-settings. [Accessed: Mar. 11, 2025].

7 Köbis N., Starke C., and I. Rahwanю Int. Journal of Anticorruption Researchб 2021, vol. 5, pp. 89–102. https://doi.org/10.48550/arXiv.2102.11567.

8 Lima M.S.M. and D. Delen. Government Information Quarterly, 2020, vol. 37, p.101407. https://doi.org/10.1016/j.giq.2019.101407.

9 de Blasio G., D'Ignazio A., and M. Letta. Technological Forecasting and Social Change, 2022, vol. 184, p. 122016. https://doi.org/10.1016/j.techfore.2022.122016.

10 Rashid A.R., Ali Z.L., and G. H. A. Alshmeel, Proc. 4th Int. Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2022. https://doi.org/10.1109/HORA55278.2022.9799970.

11 Pan M. and D. Li. International Review of Financial Analysis, 2024, vol. 84, p. 102367. https://doi.org/10.1016/j.irfa.2023.102367.

12 Van Noordt C. Government Information Quarterly, 2022, vol. 39, p. 101680. https://doi.org/10.1016/j.giq.2022.101680.

13 Yigitcanlar T., Li R.Y.M., and P.B. Bulu. Government Information Quarterly, 2023, vol. 40, p. 101733. https://doi.org/10.1016/j.giq.2022.101733.

14 Wang Z., Guan X., Sun L., and D. Zhang. Heliyon, 2023, vol. 9, p. e13765. https://doi.org/10.1016/j.heliyon.2023.e13765.

15 Valle-Cruz D., Fernández-Cortez V., and J.R. Gil-Garcia, Government Information Quarterly, 2021, vol. 38, p. 101600. https://doi.org/10.1016/j.giq.2021.101600.

16 Straub V.J., Morgan D., and J. Bertot. Government Information Quarterly, 2023, vol. 40, p. 101749. https://doi.org/10.1016/j.giq.2022.101749.

17 Adam I. and M. Fazekas. Information Economics and Policy, 2021, vol. 54, p. 100880. https://doi.org/10.1016/j.infoecopol.2020.100880.

18 Lima M.S.M. and D. Delen, Government Information Quarterly, 2020, vol. 37, p. 101407. https://doi.org/10.1016/j.giq.2019.101407.

19 Nai R., Meo R., and G.P. Pinna. Government Information Quarterly, 2022, vol. 39, p. 101679. https://doi.org/10.1016/j.giq.2022.101679.

20 Portal of legal statistics and special accounts [Online]. Available: https://qamqor.gov.kz/. [Accessed: Mar. 11, 2025].

[1]*Битанов А.,
магистрант, ORCID: 0009-0001-5156-9972,
*e-mail: asan.bitanov@gmail.com

[1]Қазақстан-Британ техникалық университеті, Алматы қ., Қазақстан

ҚАЗАҚСТАНДАҒЫ СЫБАЙЛАС ЖЕМҚОРЛЫҚ ҚЫЛМЫСТАРЫНЫҢ САНЫН БОЛЖАУ: МАШИНАЛЫҚ ОҚЫТУ ТӘСІЛІ

Аңдатпа

Бұл зерттеу машиналық оқыту әдістерін қолдану арқылы Қазақстандағы сыбайлас жемқорлық қылмыстарының санын болжауға бағытталған. Талдау ресми ай сайынғы қылмыс статистикасына негізделген, атап айтқанда, 2016 жылдан бері сыбайлас жемқорлық құқық бұзушылықтарын тіркеуге арналған №3-К есеп нысаны пайдаланылды. Зерттеу барысында үш регрессиялық модель қолданылды: k ең жақын көршілер әдісі (k-NN), градиенттік бустинг (XGBoost) және сызықтық регрессия. Модельдердің тиімділігі орташа абсолюттік қате (MAE), орташа квадраттық қате (MSE) және детерминация коэффициенті (R^2) метрикалары бойынша бағаланды. Нәтижелер сызықтық регрессияның ең жоғары болжамдық дәлдікке жеткенін көрсетті ($R^2 = 1.000$), одан кейін XGBoost ($R^2 = 0.9977$) және k-NN ($R^2 = 0.9333$). Бұл нәтижелер машиналық оқыту модельдерінің сыбайлас жемқорлық қылмыстарының динамикасын тиімді болжай алатынын дәлелдейді. Зерттеу машиналық оқытудың сыбайлас жемқорлыққа қарсы күрестегі болжау мүмкіндіктерін айқындайды. Болашақ зерттеулерде қосымша болжамдық факторларды қарастыру, балама модельдерді сынау және нақты уақыттағы деректерді интеграциялау ұсынылады.

Тірек сөздер: сыбайлас жемқорлық, машиналық оқыту, болжау аналитикасы, ең жақын көршілер әдісі, XGBoost, сызықтық регрессия, Қазақстан, қылмыс болжамы, сыбайлас жемқорлыққа қарсы күрес, регрессиялық модельдер.

[1]*Битанов А.,
магистрант, ORCID: 0009-0001-5156-9972,
*e-mail: asan.bitanov@gmail.com

[1]Казахстанско-Британский технический университет, г. Алматы, Казахстан

ПРОГНОЗИРОВАНИЕ КОЛИЧЕСТВА КОРРУПЦИОННЫХ ПРЕСТУПЛЕНИЙ В КАЗАХСТАНЕ: ПОДХОД НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Аннотация

Данное исследование направлено на прогнозирование количества коррупционных преступлений в Казахстане с использованием методов машинного обучения. Анализ основан на официальных ежемесячных данных о преступности, собранных с портала правовой статистики, в частности данных из отчета № 3-К, который фиксирует случаи коррупционных преступлений с 2016 г. Были применены три регрессионные модели: метод k ближайших соседей (kNN), градиентный бустинг (XGBoost) и линейная регрессия. Оценка моделей проведена по метрикам средняя абсолютная ошибка (MAE), среднеквадратичная ошибка (MSE) и коэффициент детерминации (R^2). Результаты показали, что линейная регрессия достигла наивысшей точности прогнозирования ($R^2 = 1.000$), за ней следуют XGBoost ($R^2 = 0.9977$) и kNN ($R^2 = 0.9333$). Эти данные подтверждают, что модели машинного обучения могут эффективно предсказывать динамику коррупционных преступлений. Исследование демонстрирует потенциал машинного обучения в прогнозировании коррупционных преступлений. В дальнейшем можно изучить дополнительные предикторы, протестировать альтернативные модели и интегрировать анализ с потоковыми данными для повышения точности прогнозирования.

Ключевые слова: коррупция, машинное обучение, прогностическая аналитика, метод ближайших соседей, XGBoost, линейная регрессия, Казахстан, прогнозирование преступлений, антикоррупция, регрессионные модели.

Article submission date: 05.11.2024