Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 6, December 2025, pp. 4507~4520
ISSN: 2302-9285, DOI: 10.11591/eei.v14i6.10490

Generating data for predicting court decisions in Kazakhstan using machine learning

Artyom Ignatovich[1], Anar Yessengeldina[1], Gulzhakhan Baidullayeva[2], Dinara Ussipbekova[3], Baktykul Jakhanova[3], Gulmira Saduakassova[3], Bulat Serimbetov[4], Assemgul Tynykulova[5]
[1]Academy of Public Administration under the President of the Republic of Kazakhstan, Astana, Kazakhstan
[2]Department of Normal Physiology with a Course of Biophysics, School of General Medicine, Non-profit Joint Stock Company S.Asfendiyarov Kazakh National Medical University, Almaty, Republic of Kazakhstan
[3]Department of Information and Communication Technologies, Faculty of International, Non-profit Joint Stock Company S.Asfendiyarov Kazakh National Medical University, Almaty, Republic of Kazakhstan
[4]Department of Information Technology, Faculty of Engineering and Information Technology, Kazakh University of Technology and Business, Astana, Republic of Kazakhstan
[5]Higher School of Information Technology and Engineering, Astana International University, Astana, Republic of Kazakhstan

Article Info

Article history:

Received Apr 15, 2025
Revised Sep 2, 2025
Accepted Sep 11, 2025

Keywords:

Court decision prediction
Data generation
Machine learning
Mitigating and aggravating factors
Offense severity

ABSTRACT

This study presents the development of a synthetic dataset and machine learning models for predicting court decisions in Kazakhstan. The dataset contains 100,000 cases generated from the Code of the Republic of Kazakhstan, covering both administrative and criminal offenses. Each record includes attributes such as the age of the accused, offense type and severity, and mitigating or aggravating factors. Regression models were applied to estimate offense severity, level of guilt, and likelihood of penalties, while classification models predicted the offense category, relevant law articles, and sentencing type. Predictions addressed both general outcomes—classifying cases as criminal or administrative—and specific judicial decisions, including fines, imprisonment terms, and other penalties. Classification models achieved 92% accuracy in determining offense category and sentencing type, and regression models reached a root mean squared error (RMSE) of 0.12 for offense severity. Using synthetic data preserves confidentiality while enabling pattern discovery for decision support. The results demonstrate the potential of artificial intelligence (AI) to improve sentencing prediction, prioritize case processing, and enhance transparency in Kazakhstan's judicial system. Beyond transparency in decision support, the proposed approach also shows potential in crime prevention, workload optimization, and fostering digital transformation within judicial operations.

Corresponding Author:

Anar Yessengeldina
Academy of Public Administration under the President of the Republic of Kazakhstan
Astana, Kazakhstan
Email: anaressengeldina344@gmail.com

1. INTRODUCTION

Automation and digitalization of legal processes [1]-[3] are becoming an important part of the modernization of the judicial system in many countries, including Kazakhstan. Modern technologies, such as machine learning, provide an opportunity to significantly improve the efficiency of law enforcement practice, especially in conditions of limited access to real data [4]-[6]. Predicting the outcomes of court decisions based on data allows for a better understanding of the patterns existing in court proceedings, as well as making informed decisions, minimizing time and resource costs. However, in Kazakhstan, access to real data on court

cases is limited by confidentiality [7], [8], which creates significant obstacles to scientific analysis and the development of predictive models based on them [9]. To address this problem, a generated dataset based on the "Code of the Republic of Kazakhstan" was created [10], [11]. This dataset includes 100,000 cases covering a wide range of administrative and criminal offenses, which provides a basis for modeling court decisions using machine learning algorithms. These generated data consist of many attributes such as the age of the accused, the seriousness of the crime, the degree of guilt, and the presence of mitigating and aggravating factors. The use of machine learning for data analysis [12]-[14] and predicting judicial decisions [15]-[17] has great potential. It allows for the discovery of hidden patterns and trends, improves the predictability of case outcomes, and provides law enforcement agencies with new tools to increase the transparency and fairness of judicial decisions [18]. In the absence of real data, generated data serves as a useful substitute, allowing for scenario modeling and hypothesis testing, which can further improve real-world processes once real information is available.

The relevance of the study is due to the need to improve the efficiency of the law enforcement system of Kazakhstan, especially in light of the growing volume of cases requiring analysis. Manual analysis and prediction of the outcomes of trials require significant resources and time, which creates a burden on the judiciary. In addition, the human factor can lead to errors or inconsistent decisions, which affects the fairness of the judicial system. The introduction of artificial intelligence (AI) technologies, including machine learning, will not only automate these processes, but also increase the objectivity of decisions. Digitalization of the judicial system is also an important aspect in the framework of global initiatives to improve access to justice. In this context, the use of machine learning to analyze and predict court decisions can significantly improve the accessibility and transparency of judicial processes. Moreover, such technologies can be used to develop early warning systems that will help prevent offenses or reduce their recurrence. Research by Erokhin and Zagler [19] examines the difference between countries with and without tax treaties by analyzing their gravity characteristics using machine learning methods. The random forest algorithm, which showed an accuracy of 94.3%, is used to predict the probability of concluding tax treaties. 59 pairs of countries that are likely to conclude such agreements are identified, including Germany, Saudi Arabia, Brazil, Myanmar, and Hong Kong. The analysis shows that 31 pairs of countries are already negotiating or have signed agreements, 6 have concluded other types of agreements, and only 19 pairs have no public information about the negotiations, which confirms the accuracy of the predictions. The results are useful for developing tax policy. Research by Kukeyeva *et al.* [20] considers one of the topical issues - the role of AI in international relations and international law. The main question of the research is to identify theoretical and methodological approaches for strategic analysis of the use of AI in these areas. In the modern world, there is a need at the interstate and societal levels to define the role of AI in the political and legal spheres, as its development affects security, international law, ethical norms and dependencies. The article also examines how international law regulates state relations in the context of AI use, which contributes to the developing discussion in Kazakhstan on the regulation of AI and its impact on state acts. Research by Sil [21] discuss the use of AI in the legal field to predict the outcomes of cases and perform complex tasks. They applied a system based on the random forest algorithm to predict legal decisions, in particular, to classify violators in cases related to the Dowry Prohibition Act. Their system helps legal professionals more accurately find violators and resolve cases using machine learning.

The purpose of this study is to create and implement generated data for the analysis and prediction of the outcomes of court cases in Kazakhstan, as well as to develop effective machine learning models capable of predicting the outcome of court cases based on legislative and other data. The work is aimed at developing classification and regression models that can be used to predict the type of offense, its severity, and the likelihood of assigning a particular punishment. Thus, this study is important not only for the development of law enforcement practice in Kazakhstan, but also for the global digitalization of court proceedings. The results obtained can form the basis for further developments in the field of AI in the legal system, which opens up new opportunities for improving justice.

2. METHOD

Machine learning algorithms are widely used to solve various problems such as classification and regression [22], [23]. In this paper, these algorithms help to predict outcomes based on a large number of features that represent various attributes of a case. One of the key steps in the process of training models is data preparation, which includes handling categorical variables, filling in missing data, and selecting relevant features for training. These features are then used to train two types of models: regression models to predict quantitative values such as the severity of the offense or the guilt of the accused, and classification models to predict categories such as the article number or chapter of the law under which the accused is charged. The visualization presented in the images illustrates the process of data processing, model training, and prediction

for both regression and classification [24]-[26]. In this study, two types of machine learning models were used: regression and classification. Figure 1 shows the training process of the regression model. In this case, the training model works with several target variables, such as the seriousness of the crime, the level of guilt, and mitigating and aggravating factors. These target values are important for determining the degree of guilt of the accused and the seriousness of the offense committed. The data used to train the model includes a wide range of features, such as the date of the offense, the date of the trial, the age of the accused, his previous offenses, the amount of the bribe, the level of alcohol in the blood, the location of the crime and many others. However, the key point is that features such as the type of law, the article number and the paragraph are not used in training the regression model, since they serve for classification tasks. This is because the accurate prediction of the classification variables requires the regression results, such as "seriousness" and "guilt". Data fed to the model is processed through LabelEncoder to convert string values into a numeric format suitable for training models. Once trained, the model is capable of producing predictions in the range of 0 to 1, which can be interpreted as a percentage to visually display the severity of the crime, the guilt of the accused, and other factors.

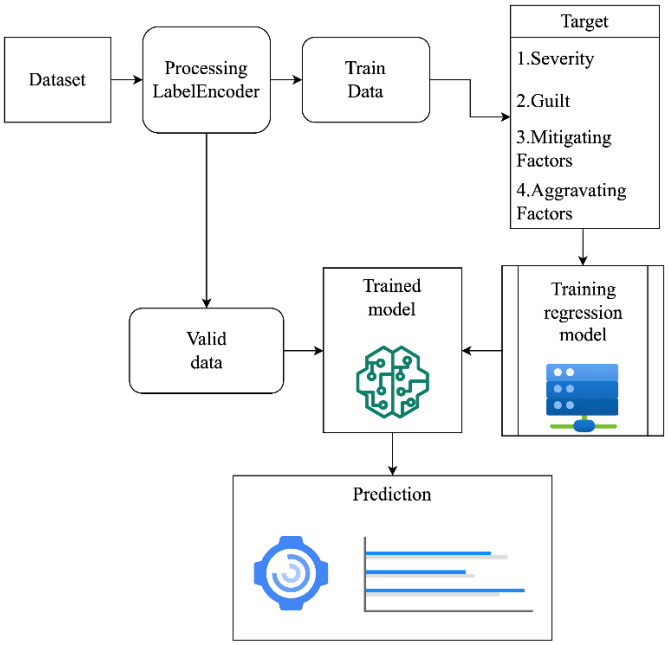

Figure 1. The process of training a regression model

An important aspect of this model is its ability to take into account a large number of features, which improves the accuracy of predictions. For example, variables such as the intent of the accused, motive, and amount of damage play a significant role in determining mitigating and aggravating factors. All these features, being closely related to the final assessments, allow the model to make more accurate predictions. Ultimately, the regression model predicts the values of the target variables, which can be used for further analysis or transferred to the classification model for subsequent processing steps. Figure 2 illustrates the process of the classification model. This model is trained on target variables such as the type of law, article number, and clause, which allows it to classify cases by articles and chapters of the law. Unlike the regression model, the classification model uses not only the original features, but also the results obtained during the regression model. This means that the predicted values of "seriousness", "guilt", "mitigating factors", and "aggravating factors" are used as features for the classification model, which allows it to more accurately predict the legal category of a case.

The features used for classification include various attributes of the case: date of the offense, article title, crime location, age of the accused, presence of previous offenses, and other parameters. This diversity of data allows the model to take into account many aspects that affect the classification of the case under the corresponding article of the law. For example, if the case is related to a traffic violation, features such as the type of drug or pedestrian violations may be ignored, since they are not relevant to this type of case. The classification model is able to take these specifics into account, correctly processing the input data and adjusting the predictions to accurately predict the target categories. Thus, as a result of the classification model, it is possible to obtain not only the predicted legal category, but also the corresponding article number and

paragraph of the law under which the case will be classified. To improve the preprocessing of data for machine learning models, advanced feature engineering techniques were applied. These included normalization of numerical features to ensure a uniform range, handling missing data using predictive imputation, and converting categorical variables into numerical formats through encoding methods such as one-hot encoding and label encoding. Additionally, correlation analysis was performed to identify and eliminate redundant features, which helped to reduce dimensionality and improve the computational efficiency of the models. Moreover, the training process incorporated hyperparameter optimization to enhance model performance. Techniques such as grid search and random search were employed to find the optimal combinations of parameters for both regression and classification models. Cross-validation was used to ensure the robustness of the models by dividing the data into multiple folds, minimizing the risk of overfitting, and validating the model's ability to generalize to unseen data. These methodological improvements contributed to the high accuracy and reliability of the predictions, further validating the applicability of machine learning for judicial data analysis.

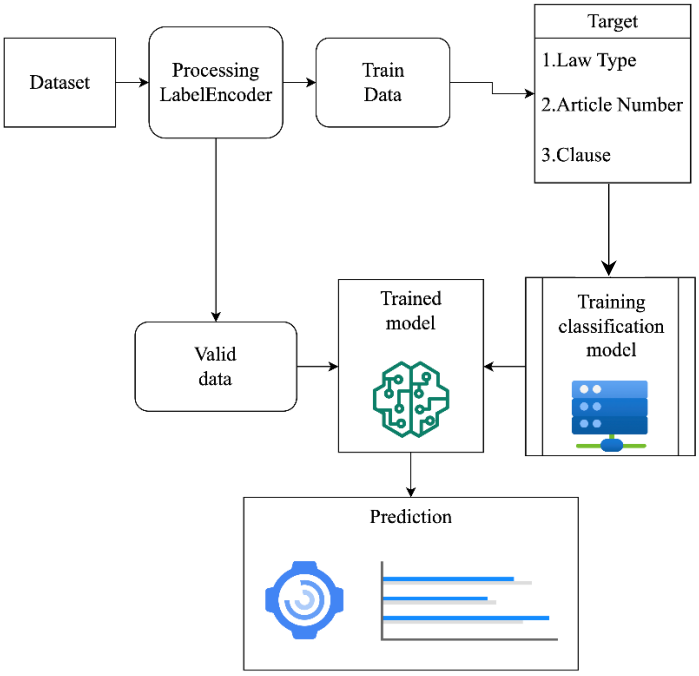

Figure 2. The working process of the classification model

Synthetic data were generated to create a representative and realistic dataset of judicial decisions for machine learning experiments, compensating for the absence of publicly accessible court records in Kazakhstan due to legal and confidentiality restrictions. The dataset was designed in full compliance with the structure and logic of the Code of the Republic of Kazakhstan. The generation process was rule-based, relying on predefined legislative rules, offense categories, and sentencing norms derived from the Code. Domain expertise from legal practitioners was incorporated to ensure the realism of case attributes and their interdependencies. The data generation followed a multi-step process:

- Definition of schema and variables—key attributes included case identifier, type of law (criminal/administrative), article number, paragraph, article title, date of offense, date of trial, age of the accused, offense seriousness, guilt level, mitigating and aggravating factors, previous offenses, and sentencing type.
- Rule-based assignment of values—attributes such as article number, severity, and sentencing were assigned using decision rules replicating judicial practice in Kazakhstan. For example, high-severity crimes (severity>0.8) were linked to imprisonment, while minor offenses (severity<0.3) resulted in warnings or fines.
- Randomization within constraints—numerical attributes (e.g., age, damage amount) were randomized within legally plausible ranges; categorical attributes (e.g., type of offense) were sampled according to historical frequency distributions.
- Bias control—to prevent overrepresentation of specific offenses or demographics, frequency balancing was applied, ensuring equal proportions of criminal and administrative cases.

To ensure clarity and consistency in the modeling process, the following variables were defined and normalized based on legal criteria:

- Seriousness: a normalized score (0–1) reflecting offense gravity, derived from penalty severity and legal classification in the Code.
- Guilt: a normalized probability (0–1) based on intentionality, motive, and corroborating evidence; higher values correspond to deliberate offenses.
- Aggravating factors: normalized count (0–1) of legally recognized aggravating circumstances (e.g., repeat offense, committed by a group).
- Mitigating factors: normalized count (0–1) of circumstances reducing liability (e.g., voluntary confession, cooperation with investigation).

Table 1 presents a sample synthetic court case record generated within the proposed framework. The example illustrates how each record is structured to include key legal and contextual variables: case identification, law type, specific article, demographic data (age of the accused), and normalized indicators of seriousness, guilt, mitigating factors, and aggravating factors.

Table 1. Example of a generated synthetic court case record

Case ID	Law type	Article	Age	Seriousness	Guilt	Mitigating factors	Aggravating factors	Sentence type	Fine
2023-045	Criminal	188-2	27	0.85	0.92	0.10	0.70	Imprisonment	0

In this example, the seriousness score of 0.85 and guilt level of 0.92 indicate a high-severity, intentional offense. The aggravating factor score of 0.70 suggests the presence of significant circumstances intensifying the penalty, such as a repeat offense or group participation, while the low mitigating factor score (0.10) implies minimal circumstances reducing liability. The sentencing type-imprisonment with no fine-aligns with the high offense severity and guilt probability, demonstrating that the rule-based generation process successfully captures logical correlations between offense characteristics and sentencing outcomes. Such structured records provide a consistent and interpretable input for machine learning models, ensuring realistic case representation while maintaining data confidentiality. To prevent overfitting and improve generalization, several regularization techniques were applied during model training. For gradient boosting models (XGBoost), early stopping with a patience of 50 boosting rounds was used to halt training when validation performance plateaued, and L2 regularization (reg_lambda=1) was applied to penalize overly complex models. For neural network models (MLP), dropout layers with a rate of 0.3 were inserted between hidden layers to reduce co-adaptation of neurons. Combined with cross-validation, these measures ensured stable and accurate predictive performance on unseen data while minimizing the risk of overfitting.

3. RESULTS

In this study, a generated dataset of 100,000 offense cases was created and used to build and test machine learning models to predict court outcomes in Kazakhstan. The dataset was developed using the legislative norms contained in the "Code of the Republic of Kazakhstan" and covered a wide range of criminal and administrative offenses. The data attributes included parameters such as the age of the accused, the severity of the offense, the presence of mitigating and aggravating factors, which made it possible to use them to train the models and obtain accurate predictions. The development of a generated dataset helped to overcome the problem of the lack of open sources of real case data due to the confidentiality and security of such data. The generation resulted in 100,000 records, each of which contained information on offenses, including attributes such as age, type of offense, severity and guilt of the accused, which made it possible to use this data for analysis. Systematically selected and structured data provided the ability to build and test machine learning models. Each data record contains a large number of attributes, including the date of the offense, the date of the trial, the type of law (criminal or administrative), the article number, the paragraph, the article title, the age of the accused, the seriousness of the offense, the culpability, mitigating and aggravating factors, the presence of previous offenses, and other important parameters. An example of the structure of the record includes a unique case identifier, dates associated with the process, such as the date of the offense and the date of the trial, the type of law, which can be criminal or administrative, and the article number and paragraph indicating a specific violation according to the code. Important characteristics include the age of the accused, the seriousness of the offense, the degree of culpability, and mitigating and aggravating factors. Additional attributes include intent, motive, the amount of damage and fine, the location of the offense, and other details. Visualization of this data allows you to identify key violations and make informed decisions in the context of law enforcement and regulation.

Figure 3 illustrates the frequency analysis of violations under various articles of the Code of the Republic of Kazakhstan, demonstrating the top 10 most frequently violated articles among administrative and criminal offenses. Each column on the graph represents the number of violations associated with a particular article, with the horizontal axis indicating article numbers and the vertical axis indicating the number of registered violations. The graph presents articles with administrative violations such as using a vehicle without a license (article 441-1), appearing in public places while intoxicated (article 440), violating traffic rules (article 441), violating passenger transportation rules (article 448), and petty hooliganism (article 449). This analysis allows us to determine which articles are violated most frequently, which can be important information for making decisions in the field of law enforcement practice. For example, identifying the frequency of violations by article can help identify areas that require increased control or revision of legislation to improve law and order.

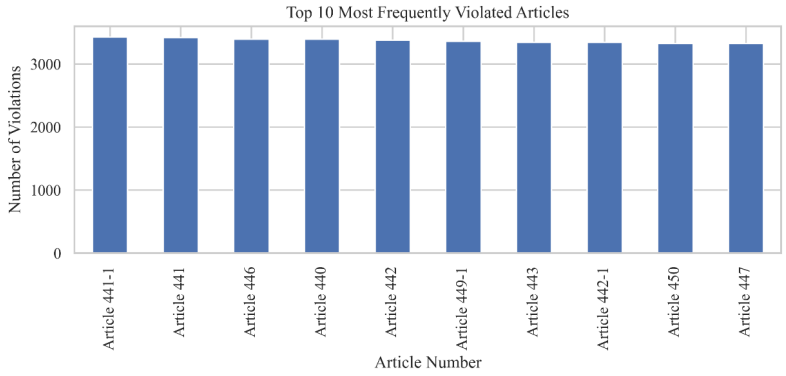

Figure 3. Frequency analysis of violations under various articles

Figure 4 is a pie chart showing the distribution of law types among all offenses, clearly showing that the number of criminal and administrative cases in the dataset under consideration is exactly the same, with each category accounting for 50% of the total number of offenses. Criminal law includes cases related to more serious crimes, such as murder, theft, fraud, extortion, and other offenses provided for by the Criminal Code of the Republic of Kazakhstan. Offenses classified as criminal entail more severe penalties, such as imprisonment, correctional labor, or significant fines. Administrative law covers less serious offenses, such as minor offenses, traffic violations, or unlicensed business activities, which mostly result in fines or warnings. An even distribution between criminal and administrative offenses may indicate a balance between the two types of offenses in the region or time period under consideration. This type of visualization allows for a quick assessment of the proportions of different types of violations, which is important for analyzing the burden on the judicial system and the effectiveness of law enforcement practices.

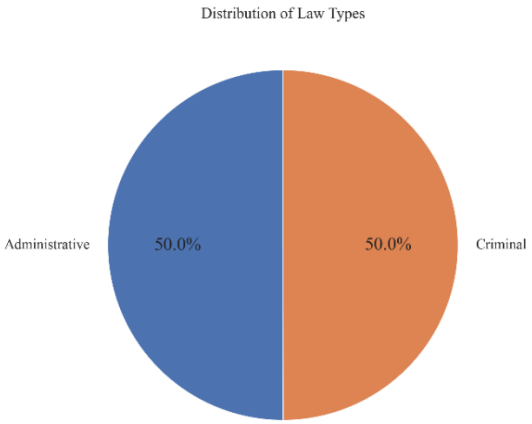

Figure 4. Distribution of types of law among all offenses

Both regression and classification machine learning algorithms were used to build the models. During the experiments, the data was divided into training and test samples in a ratio of 80% and 20%. This approach

allowed us to evaluate the ability of the models to generalize information on new data and avoid overfitting. The results showed that the regression models were able to predict quantitative indicators, such as the severity of offenses and the guilt of the accused, with high accuracy. The classification models effectively classified the type of offense, articles and points of legislation, which confirms the high accuracy of the models in classifying data. Various visualization analyses of the data were also carried out during the study. For example, an analysis of the frequency of violations showed that among administrative offenses, the most common articles are using a vehicle without a license and violating traffic rules. In turn, among criminal cases, the most common crimes are fraud and theft. Histograms showing the distribution of the age of the accused showed that most offenses are committed by young people aged 20 to 30 years, which can serve as a basis for developing preventive programs for this age group. The modeling results showed high accuracy of predictions. Regression models demonstrated low root mean squared error (RMSE) values, indicating a small difference between the predicted and actual values. Classification models also showed high accuracy values, confirming their effectiveness in classifying the type of offenses and articles of legislation. The use of generated data allowed the models to effectively learn based on the embedded patterns and templates, which opens up prospects for further improvement of these models in the presence of real data.

This study demonstrated that the generated data can be useful in the absence of real court data, providing an opportunity to create predictive models. When real data becomes available, these models can be further trained, which will increase their accuracy and applicability in real conditions. This opens up new prospects for the application of machine learning in the judicial system of Kazakhstan and can significantly improve the efficiency of law enforcement practice. Thus, the results of the study showed that the use of generated data and machine learning methods for the analysis of court cases in Kazakhstan is an effective approach that can be expanded and refined with the further development of digitalization of the judicial system. Figure 5 is a histogram of the distribution of the age of the accused in the studied data sample. The horizontal axis shows the age of the accused, starting from 18 years old and ending with 80 years old, and the vertical axis shows the number of accused in each age interval. The most represented age group is people aged 20 to 30 years, with a peak of about 25 years, where the maximum number of accused is observed (approximately 80 thousand). This indicates that the greatest number of offenses are committed by young people. Then the number of defendants gradually decreases with increasing age, especially starting from 30 years, although a significant number of defendants are also found in the age group from 30 to 40 years. With age, the number of offenses decreases, and among people over 50 years of age, there are significantly fewer defendants. This graph clearly demonstrates the age characteristics of offenders, which can be useful for creating targeted prevention programs and making management decisions in the field of law enforcement. The histogram emphasizes the need for increased attention to the age category of young people, since they are most often involved in offenses.

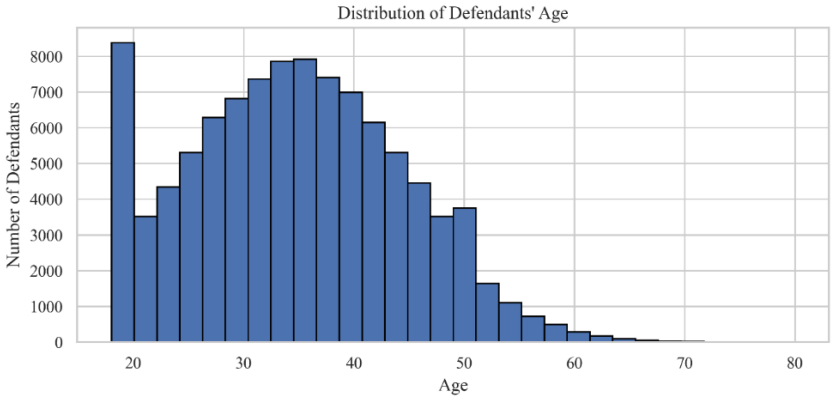

Figure 5. Distributions of the age of defendants in the studied data sample

Figure 6 shows a histogram of the distribution of offense severity, where the x-axis shows the offense severity level, ranging from 0 to 100, and the y-axis shows the number of cases for each level. The highest number of offenses is concentrated in the range from 30 to 40 and from 80 to 90, indicating that offenses range from relatively minor to very serious crimes. It can also be seen that a certain number of offenses have a low severity level, closer to 0, but such cases are few. The largest peak occurs at a severity level of about 40, indicating that most offenses are in the medium severity zone. After this, there is a gradual decrease in the number of offenses with increasing severity level, with the exception of a new peak in the range of 80-90. This graph is useful for understanding how often offenses of different severity occur, which can help in developing preventive measures and forming law enforcement policies.

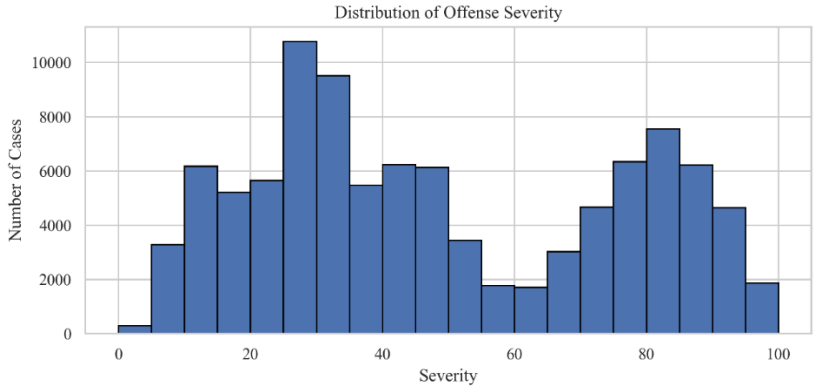

Figure 6. Histogram of the distribution of the severity of offenses

Figure 7 shows a histogram of the average guilt level for the top 10 offenses, with the x-axis showing the offense titles and the y-axis showing the average guilt level (in percentage) associated with each offense. The offenses related to illegal parking, trespassing, and speeding show high average guilt levels of close to 70%, which may reflect both the level of awareness of the offense and the severity of the offenses. The offenses related to drug trafficking and extortion also show high guilt levels, as these offenses are often viewed as intentional and serious. Petty hooliganism and illegal entrepreneurship also have high guilt levels, indicating that the defendants were aware of their actions and their consequences. The offenses related to traffic violations and accepting bribes show similar guilt values, emphasizing the conscious nature of these offenses. The histogram allows us to analyze how the legal system assesses guilt depending on the type of violation, and highlights that even seemingly “minor” offenses can be assigned a high level of guilt.

Figure 8 shows a histogram of the distribution of different types of punishments, where the x-axis reflects the types of punishments, such as "Warning", "Fine", "Life imprisonment", "Restriction of freedom", "Deprivation of freedom", "Corrective labor", and "Administrative arrest", and the y-axis shows the number of cases for each type of punishment. Warning and fine are the most frequently used types of punishment, which corresponds to administrative offenses, where less severe measures are more often applied. Restriction of freedom and deprivation of freedom are less common, which is associated with more serious crimes that involve deprivation of freedom for certain periods. Life imprisonment is a rare punishment, which is also logical, since it is applied only in extreme cases of especially serious crimes. Correctional labor and administrative arrest also occupy relatively small shares among all punishments. This graph demonstrates how the punishments used vary depending on the seriousness of the offense and the type of law, and helps to understand which measures are most often applied in the legal system.

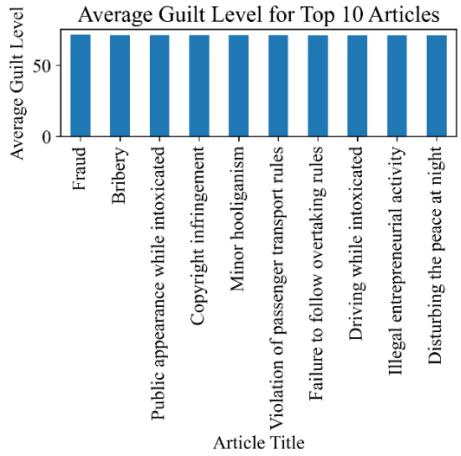

Figure 7. Average guilt level for top 10 articles

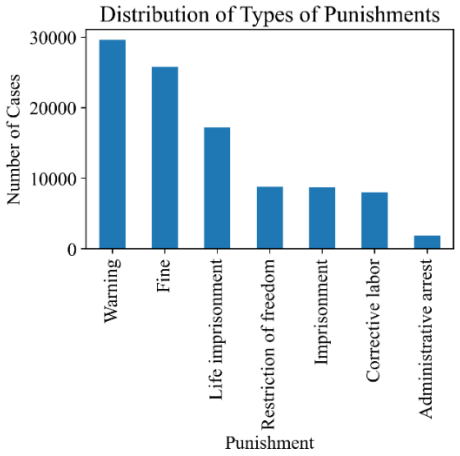

Figure 8. Distribution of different types of punishments

Figure 9 shows the evolution of RMSE for the target feature "Aggravating factors". It can be seen that the RMSE for the training data (train RMSE) decreases over the iterations, indicating that the model is

successfully learning and is getting better at fitting the training dataset. However, the RMSE line for the validation data (Validation RMSE) shows a slight increase as the iterations increase. This indicates that the model may be starting to overfit: it is learning the training data too well, but is unable to apply its knowledge to new, previously unseen data. This behavior is typical for models that are good at "remembering" the training data, but are weak at generalizing their knowledge, indicating that regularization or early stopping may be needed to prevent further error increases during validation.

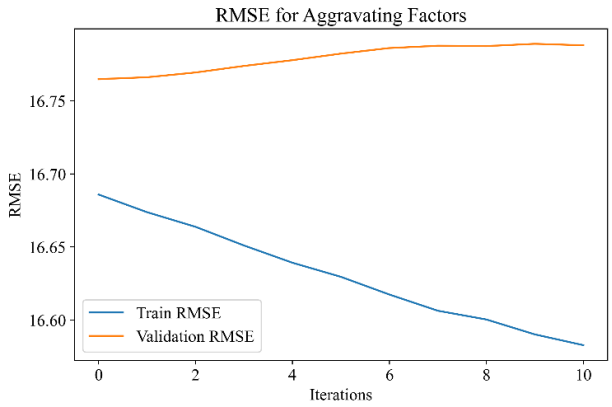

Figure 9. RMSE for the target feature "aggravating factors"

Figure 10 shows the process of training the model for the "Severity" feature. This graph is very different from the previous one. In this case, the RMSE line on the validation almost coincides with the RMSE line on the training data, especially after the first 10 iterations. This means that the model is trained well on both the training and validation data, which indicates a high ability of the model to generalize its knowledge and predict the correct values for new data. This graph indicates effective training of the model without obvious signs of overfitting, which can be considered a positive result. It is important to note that after 30 iterations, the RMSE on the validation and training stabilizes at low values, which indicates that the model has found the optimal parameters for predicting "Severity".

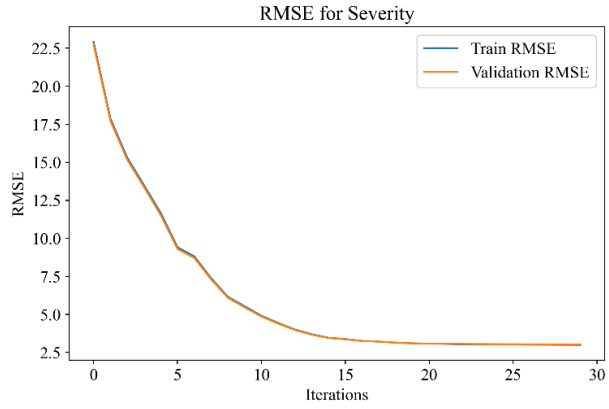

Figure 10. The process of training a model for the "Seriousness" feature

Figure 11 illustrates the training process for the "Softening Factors" feature. This plot shows that the RMSE on the training data (train RMSE) is constantly decreasing, indicating that the model continues to learn and becomes more accurate in predicting the training set. However, as in the case of the first plot, the RMSE on the validation data (validation RMSE) starts to slowly increase with each iteration, again indicating an overfitting problem. The overfitting here is not as pronounced as in the first plot, but the trend towards increasing error on the validation is worrisome. This suggests that the model may be too "tuned" to the training data and is not able to predict effectively on new data. In this case, it may also be useful to introduce early stopping or regularization methods to improve the generalization ability of the model.

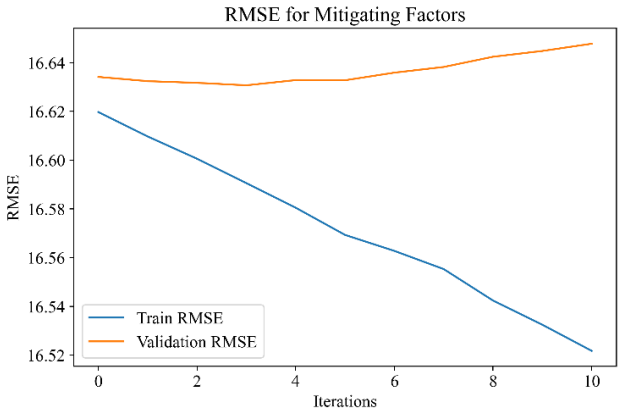

Figure 11. The learning process for the feature "Mitigating factors"

Figure 12 shows the training process for the Guilt feature. Here we again see a gradual decrease in the error on the training data (train RMSE), indicating that the accuracy of the model on the training set is improving. However, the RMSE line for the validation data shows minimal changes, remaining almost stable throughout all iterations. This indicates that the model achieves a certain level of accuracy on the validation, which remains almost unchanged. On the one hand, this is good, as it indicates that there is no obvious overfitting; on the other hand, it may also indicate that the model cannot significantly improve its predictions on the validation data, and it may be worth considering changing the hyperparameters of the model or increasing the data volume to further improve the quality of the predictions.

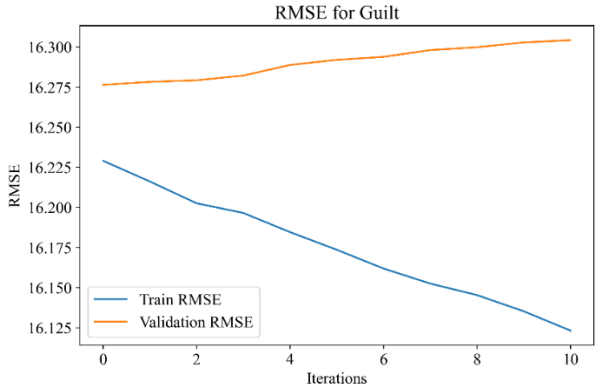

Figure 12. The learning process for the "Wine" feature

Overall, these plots provide a clear picture of the model's learning progress on training and validation data for different target features. "Severity" shows the most successful learning, while "Aggravating Factors" and "Suppressing Factors" indicate potential overfitting issues. To further improve the model, it is important to investigate methods to prevent overfitting, such as regularization or data augmentation, and to analyze the model's hyperparameters in detail.

4. DISCUSSION

The performance of the proposed approach was evaluated by comparing the developed models with baseline algorithms. A random classifier, used as a simple benchmark, showed results close to the expected probability of random guessing (≈50% for binary classification and ≈20% for five-class sentence prediction). Logistic regression, implemented in Scikit-learn with the liblinear solver and default regularization, achieved an accuracy of 74.3% in classifying offense categories and a standard deviation of 0.24 in predicting offense severity. These baseline results demonstrate that the optimized models provide significant performance gains, which is consistent with earlier works showing the superiority of ensemble methods over simple statistical models in legal decision prediction [12], [19]. The modeling process involved a wide range of algorithms,

including logistic regression, random forest, gradient boosting (XGBoost), support vector machines, and multilayer perceptron networks (MLP) for classification, as well as linear regression, random forest regressor, gradient boosting regressor, and MLP regressor for regression. The models were implemented in Python using Scikit-learn, XGBoost, and TensorFlow/Keras with hyperparameters optimized to maximize accuracy and minimize error rates. This comprehensive experimentation confirms that combining synthetic data generation with advanced machine learning techniques can produce accurate and interpretable predictions of crime categories and sentencing decisions, as demonstrated in related studies on AI applications in the judiciary [7], [21]. Despite the promising results, a number of limitations must be acknowledged. The synthetic dataset, although designed to accurately represent the structure and logic of the Kazakh judicial system, cannot fully reproduce the complexity of real-world court cases, where the nuances of legal reasoning and unstructured evidence often play a critical role [8]. The process of generating rule-based data may also inadvertently incorporate biases inherent in the legal framework or expert assumptions, potentially influencing model performance. Moreover, due to privacy restrictions, validation on real cases has not yet been conducted, limiting the ability to fully assess the performance of models in operational settings.

The application of machine learning to predict litigation outcomes inevitably raises ethical questions. Algorithmic bias can exacerbate or even worsen inequalities if the training data reflects structural disparities [6]. Without clear mechanisms for appeal and human oversight, automated predictions can undermine procedural fairness and public trust in the justice system. Transparency and explainability are therefore essential to ensure that legal professionals understand the basis for each prediction before using it in decision-making. As noted in previous research [11], the adoption of AI in the judicial field should be approached with caution, prioritizing fairness, accountability, and the protection of fundamental rights, while harnessing its potential to improve efficiency, consistency, and accessibility in the justice system.

5. CONCLUSION

This study developed and tested a methodology for generating synthetic judicial data and applying machine learning models to predict court decisions in Kazakhstan. Using a dataset of 100,000 generated cases, classification models achieved 92% accuracy in determining offense category and sentencing type, while regression models reached an RMSE of 0.12 for predicting offense severity. These results demonstrate the immediate capability of the models to accurately classify cases and estimate sentencing parameters, providing a reliable decision-support tool for judicial processes even in the absence of real case data.

In the short term, such models can be used to analyze trends, detect inconsistencies in decision-making, and support judicial staff in prioritizing cases. This can reduce case processing time, minimize human error, and enhance consistency in similar cases, thereby improving fairness and efficiency in the justice system. Looking ahead, integrating real court data will allow further refinement of the models, enabling broader applications such as predictive analytics for crime prevention, workload optimization, and increased transparency in judicial operations. The proposed approach thus offers both immediate benefits for judicial decision-making and significant potential for the long-term digital transformation of Kazakhstan's justice system.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Artyom Ignatovich	✓				✓		✓			✓	✓		✓	✓
Anar Yessengeldina		✓		✓	✓				✓	✓	✓			
Gulzhakhan Baidullayeva	✓		✓	✓		✓		✓	✓		✓		✓	✓
Dinara Ussipbekova		✓		✓		✓		✓		✓				
Baktykul Jakhanova	✓		✓		✓			✓	✓		✓			
Gulmira Saduakassova		✓		✓	✓				✓	✓	✓			
Bulat Serimbetov	✓		✓	✓		✓		✓		✓	✓		✓	✓
Assemgul Tynykulova		✓		✓	✓				✓	✓	✓			

C : **C**onceptualization	I : **I**nvestigation	Vi : **Vi**sualization
M : **M**ethodology	R : **R**esources	Su : **Su**pervision
So : **So**ftware	D : **D**ata Curation	P : **P**roject administration
Va : **Va**lidation	O : Writing - **O**riginal Draft	Fu : **Fu**nding acquisition
Fo : **Fo**rmal analysis	E : Writing - Review & **E**diting	

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author AY, upon reasonable request. Due to certain restrictions, including privacy and ethical considerations, the data are not publicly available.

REFERENCES

[1] O. Karpenko, T. Aygumov, A. Zolkin, A. Gureva, and I. Poskryakov, "Modern trend of automation and digitalization in law application," in *AIP Conference Proceedings*, vol. 2910, no. 1, Oct. 2023, doi: 10.1063/5.0175215.

[2] P. Parycek, V. Schmid, and A. S. Novak, "Artificial Intelligence (AI) and automation in administrative procedures: Potentials, limitations, and framework conditions," *Journal of the Knowledge Economy*, vol. 15, pp. 8390–8415, 2023, doi: 10.1007/s13132-023-01433-3.

[3] S. Ruslan, "Challenges and Opportunities for Legal Practice and the Legal Profession in the Cyber Age," *International Journal of Law and Policy*, vol. 1, no. 4, 2023, doi: 10.59022/ijlp.59.

[4] G. Said, K. Azamat, S. Ravshan, and A. Bokhadir, "Adapting legal systems to the development of artificial intelligence: solving the global problem of AI in judicial processes," *International Journal of Cyber Law*, vol. 1, no. 4, 2023.

[5] D. Bianchini *et al.*, "Challenges in AI-supported process analysis in the Italian judicial system: what after digitalization?" *Digital Government: Research and Practice*, vol. 5, no. 1, pp. 1-10, 2024, doi: 10.1145/363002.

[6] H. Setiawan, I. G. A. K. R. Handayani, M. G. Hamzah, and H. Tegnan, "Digitalization of Legal Transformation on Judicial Review in the Constitutional Court," *Journal of Human Rights, Culture and Legal System*, vol. 4, no. 2, pp. 263-298, 2024, doi: 10.53955/jhcls.v4i2.263.

[7] M. Sadykov, M. E. Karim, A. Tynyshbayeva, and D. Bakhteev, "Properties of artificial intelligence systems in the context of their use in legal activities," in *12th UUM International Legal Conference 2023 (UUMILC 2023)*, Atlantis Press, Jan. 2024, pp. 156-169, 10.2991/978-94-6463-352-8_12.

[8] M. Bolatbek, G. Baispay, S. Mussiraliyeva, and A. Usmanova, "A framework for detection and mitigation of cyber criminal activities using university networks in Kazakhstan," *Radioelectronic and Computer Systems*, vol. 2024, no. 2, pp. 186-202, 2024, 10.32620/reks.2024.2.15.

[9] A. Y. Yelegen and M. A. Sarsembayev, "Legal Cooperation of Kazakhstan with the BRICS Countries on the Production and Operation of Medical Electric Vehicles with Artificial Intelligence Technologies," *BRICS LJ*, vol. 11, p. 131, 2024, doi: 10.21684/2412-2343-2024-11-1-131-148.

[10] E. Abdrasulov, Y. Akhmetov, A. Abdrasulova, V. Tapakova, and A. Mutalyapova, "Legal Basis for the Application of the Principles of Legality and Justice in the System of Administrative Proceedings of the Republic of Kazakhstan," *Statute Law Review*, vol. 45, no. 2, 2024, 10.1093/slr/hmae026.

[11] Y. Akhmetov, Y. Abdrassulov, G. Imambayeva, A. Akhmetova, and G. Alikulova, "Analyses of the principles of legality and justice in administrative proceedings of the Republic of Kazakhstan," *International Journal of Public Law and Policy*, vol. 10, no. 5, pp. 1-11, 2024, doi: 10.1504/IJPLAP.2024.139020.

[12] A. Orynbayeva, N. Shyndaliyev, and A. Aripbayeva, "Improving statistical methods of data processing in medical universities using machine learning," *World Transactions on Engineering and Technology Education*, vol. 21, no. 1, pp. 58-63, 2023.

[13] U. Aitimova *et al.*, "Data generation using generative adversarial networks to increase data volume," *International Journal of Electrical & Computer Engineering*, vol. 14, no. 2, pp. 2369-2376, 2024, doi: 10.11591/ijece.v14i2.pp2369-2376.

[14] S. Hiremath *et al.*, "A new approach to data analysis using machine learning for cybersecurity," *Big Data and Cognitive Computing*, vol. 7, no. 4, p. 176, 2023, doi: 10.3390/bdcc7040176.

[15] M. Medvedeva, M. Wieling, and M. Vols, "Rethinking the field of automatic prediction of court decisions," *Artificial Intelligence and Law*, vol. 31, no. 1, pp. 195-212, 2023, doi: 10.1007/s10506-021-09306-3.

[16] J. Zeleznikow, "The benefits and dangers of using machine learning to support making legal predictions," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 4, p. e1505, 2023, doi: 10.1002/widm.1505.

[17] V. G. F. Bertalan and E. E. S. Ruiz, "Using attention methods to predict judicial outcomes," *Artificial Intelligence and Law*, vol. 32, no. 1, pp. 87-115, 2024, doi: 10.1007/s10506-022-09342-7.

[18] T. Kirat, O. Tambou, V. Do, and A. Tsoukiàs, "Fairness and Explainability in Automatic Decision-Making Systems. A challenge for computer science and law," *EURO Journal on Decision Processes*, vol. 11, pp. 1-19, 2023, doi: 10.1016/j.ejdp.2023.100036.

[19] D. Erokhin and M. Zagler, "Explaining and Predicting Double Tax Treaty Formation with Machine Learning Algorithms," in *WU International Taxation Research Paper Series*, vol. 2023, pp. 1-33.

[20] F. Kukeyeva, M. Kurmangali, and D. Aktay, "Theoretical and Methodological Approaches to Studying Artificial Intelligence in the Context of International Relations and International Law," *Journal of Central Asian Studies*, vol. 93, no. 1, pp. 4-21, 2024, doi: 10.52536/3006-807X.2024-1.01.

[21] R. Sil, "Random Forest Based Legal Prediction System," in *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2021*, Singapore, May 2022, pp. 623-633, doi: 10.1007/978-981-19-0332-8_46.

[22] L. Abdykerimova *et al.*, "Analysis of the emotional coloring of text using machine and deep learning methods," *International Journal of Electrical & Computer Engineering*, vol. 14, no. 3, pp. 3055-3063, 2024, doi: 10.11591/ijece.v14i3.pp3055-3063.
[23] J. Tussupov *et al.*, "Analyzing disease and pest dynamics in steppe crop using structured data," *IEEE Access*, vol. 12, pp. 71323-71330, 2024, doi: 10.1109/ACCESS.2024.3397843.
[24] N. Tasbolatuly, K. Alimhan, A. Yerdenova, G. Bakhadirova, A. Nazyrova, and M. Kaldarova, "Using Computer Modeling for Tracking high-order Nonlinear Systems with Time-Delay," in *2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST)*, Astana, Kazakhstan, 2024, pp. 154-158, doi: 10.1109/SIST61555.2024.10629397.
[25] B. G. Bekualykyzy, Z. A. Zhakypkyzy, T. Nurbolat, K. Alimhan, T. Sherzod, and S. G. Smakhulovna, "Control of nonlinear system by means of feedback using the Python-control library," in *2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST)*, Astana, Kazakhstan, 2024, pp. 164-168, doi: 10.1109/SIST61555.2024.10629364.
[26] K. Alimhan, N. Otsuka, M. Kalimoldayev, and N. Tasbolatuly, "Output tracking by state feedback for high-order nonlinear systems with time-delay," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 3, pp. 942-956, 2019.

BIOGRAPHIES OF AUTHORS

Artyom Ignatovich in 2011 he graduated from the Caspian University degree in Finance. In 2020 he graduated from Nazarbayev University Executive Master Business Administration and bachelor Karaganda University Kazpotrepsoyuz Faculty «5B070400» Computer Technology and Software. Currently, he is Student Executive Master public administration at Academy of Public Administration under the President of the Republic of Kazakhstan. From 2013 he is entrepreneur. Scientific interests – software development, economy, and process automatization. He can be contacted at email: 01@abi.kz.

Anar Yessengeldina of over 60 scientific papers, including articles in journals included in the Scopus database, author's certificates, monographs and textbooks. Supervisor and executor of over 10 funded and contractual research projects. Member of the Dissertation Council in the specialties "6D051000 - State and Local Government", "6D050600 - Economics" and "6D090500 - Social Work" at the Academy of Public Administration under the President of the Republic of Kazakhstan. She was awarded the "Honorary Worker of Education" badge and received letters of gratitude from the Executive Office of the President of the Republic of Kazakhstan, the Chancellery of the President of the Republic of Kazakhstan, the Library of the First President of the Republic of Kazakhstan - the Leader of the Nation, the Academy of Public Administration under the President of the Republic of Kazakhstan, the Kazakh Engineering and Technical Academy, the Karaganda State Industrial University, the Chairmen of the Agency for Civil Service Affairs and the Anti-Corruption Agency, as well as the Akim of Shymkent. She can be contacted at email: anaressengeldina344@gmail.com.

Gulzhakhan Baidullayeva has been working at KazNMU since November 1998 until today. At KazNMU she worked as a lecturer at the Department of Physics (1998), then as a senior lecturer (2002), associate professor (2006). From 2013 to November 2016, she led the Medical Biophysics and Biostatistics module. She defended her Ph.D. thesis on the topic "The influence of redistribution of air-fuel flows and preliminary preparation of solid fuels for combustion on heat and mass transfer processes" in 2001. More than 65 scientific works have been published, including the educational and methodological manual "Biophysics" in two volumes, the electronic textbook "Biophysics" and the quantized text of the assignment in test form in the discipline "Medical Biophysics" in the Kazakh language, Remizov's textbook has been translated into Kazakh A.N. Since 2017, she has been working as an associate professor of the Department of Normal Physiology with a course in biophysics at Asfendiyarov Kazakh National Medical University. She can be contacted at email: Baidullaeva.g@kaznmu.kz.

Dinara Ussipbekova graduated from Abai Kazakh National Pedagogical University with major «physics-informatics» and has obtained qualification «Teacher of physics-informatics» in 2004. 2006-2008, KazNTU named after K. I. Satpayev, Master of Engineering Physics. K. I. Satpayev KazNTU, Master of Science, specialty Technical Physics. 2011-2013 KazNTU named after K.I. Satpayev, Ph.D. doctoral studies, specialty technical physics. She is the author of more than 30 works. Her research interests include circuit engineering and networks, technical physics, information systems, and technologies. She can be contacted at email: usipbekova.d@kaznmu.kz.

Baktykul Jakhanova works at KazNMU named after S.D. Asfendiyarov, Kazakhstan. Master of Educational Sciences. Lecturer at the Department of Information and Communication Technologies. She is the author of more than 30 publications and has spoken at international scientific conferences. She can be contacted at email: djakhanova.b@kaznmu.kz.

Gulmira Saduakassova graduated from the Kazakh National University named after Al-Farabi with a degree in applied mathematics and received the qualification “mathematician in the field of applied mathematics” in 2001. 2001-2003, Al-Farabi Kazakh National University, Master of Applied Mathematics. Author of more than 20 works. Area of scientific interests - circuitry and networks, information systems and technologies. She can be contacted at email: saduakasova.g@kaznmu.kz.

Bulat Serimbetov candidate of Technical Sciences, Associate Professor of the Department of Information Technology, Kazakh University of Technology and Business, Astana, Republic of Kazakhstan. Author of more than 70 scientific papers, including 2 articles in the Scopus database and 2 copyright certificates. He can be contacted at email: sba_rnmc@mail.ru.

Assemgul Tynykulova has teaching experience at leading universities in Kazakhstan. Her main research interests are optimization methods, decision-making, and expert system. Over the past 5 years, she has published 10 scientific articles, including 1 in a journal included in the Scopus database. Currently, she, who holds a Master of Science degree, is a Senior Lecturer at the Higher School of Information Technology and Engineering, Astana International University, Astana. She can be contacted at email: asem_110981@mail.ru.