Journal of Statistical Research *https://doi.org/10.3329/jsr.v58i2.80626*
2024, Vol. 58, No. 2, pp. 385-395 *ISSN 0256 - 422 X*

USING MACHINE LEARNING TO PREDICT PERFORMANCE OF TRIAL COURT ADMINISTRATION: AN EMPIRICAL STUDY WITH IRANIAN PERFORMANCE INDICATORS OF TRIAL CASE PROCESSING

MOHADESEH ALSADAT FARZAMMEHR*

Judiciary Research Institute, Tehran, Iran

Email: m.farzammehr@jri.ac.ir

ELHAM TABRIZI

Department of Mathematics, Faculty of Mathematics and Computer Science, Kharazmi University, Tehran, Iran

Email: elham.tabrizi@khu.ac.ir

MEISAM MOGHIMBEYGI

Department of Mathematics, Faculty of Mathematics and Computer Science, Kharazmi University, Tehran, Iran

Email: M.moghimbeygi@khu.ac.ir

SUMMARY

This paper explores the significance of evaluating justice system performance to ensure effectiveness across diverse legal frameworks. Traditionally, methods like expert surveys and document analysis were used to generate empirical indicators. However, this study employs machine learning to predict trial court performance, using key processing indicators. Data from 21 branches of the General Court of Law in Tehran, Iran, comprising 119 case management records, is analyzed. logistic regression proves most effective among various models, achieving 98.5% AUC and 95.0% CA. Results indicate that resolved cases impact positively, while pending cases have minimal influence. Monitoring time and working days contribute insignificantly. Early detection of negative performance issues is crucial for maintaining public trust. Regular evaluations not only enhance court efficiency but also aid in developing decision support systems for improved performance.

Keywords and phrases: court performance prediction; data mining; judicial data; machine learning techniques

* Corresponding author
© Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh.

1 Introduction

The performance of court administration is a critical factor in ensuring the effectiveness of any justice system. Timely and efficient case management is essential to the delivery of justice, as delays or inefficiencies can undermine public trust and confidence in the legal system. In the Iranian judicial system, as in other jurisdictions, performance indicators such as case resolution times, clearance rates, and pending caseloads are commonly used to evaluate court efficiency. However, these traditional metrics often fail to capture the complex interactions between various factors influencing court performance, making it difficult to generate actionable insights for improvement (Ostrom and Hanson , 2000; Steelman et al., 2000).

To address this challenge, this study applies machine learning techniques to predict the performance of trial courts in Tehran, Iran, using key trial case processing indicators. By utilizing historical data from 21 branches of the General Court of Law, we develop a predictive model that classifies judicial unit performance as either ”positive” or ”negative.” A positive outcome indicates that the court has efficiently resolved cases, while a negative outcome reflects inefficiencies or delays in case processing. This binary classification serves as the primary target variable for our analysis and provides a clear metric for evaluating court administration effectiveness. Machine learning algorithms were chosen due to their ability to analyze large and complex datasets and uncover patterns that traditional statistical methods may overlook. Techniques such as logistic regression, gradient boosting, neural networks, and stochastic gradient descent (SGD) offer several advantages in this context. These models can detect intricate relationships between multiple input variables—such as the number of cases referred to judges, the number of resolved cases, and processing times—and provide accurate predictions based on these features.

Additionally, machine learning models provide the dual benefit of predictive power and interpretability, allowing court administrators and policymakers to understand which factors are most strongly associated with court performance. This interpretability is crucial for implementing data-driven decision-making, optimizing resource allocation, and improving judicial efficiency.

Beyond prediction, this study aims to identify key drivers of judicial performance. Specifically, we investigate how variables such as the number of resolved cases, pending caseloads, and processing times influence the likelihood of a court achieving positive performance. Understanding these relationships can help shape strategies for improving court operations, enhancing resource allocation, and ultimately strengthening public confidence in the judiciary.

We focus on the trial courts of Tehran due to the availability of detailed judicial data provided by the Statistical and Information Technology Center of the Judiciary of Iran. This dataset includes comprehensive information on case referrals, case resolutions, and processing durations, making it an ideal foundation for a data-driven assessment of court performance.

The following sections describe the dataset, machine learning algorithms used, and evaluation methods applied to assess model performance. We also present the results and discuss their implications for improving the efficiency of trial courts in Tehran. The findings of this study contribute to the growing body of research on data-driven judicial performance analysis and demonstrate the potential of machine learning to enhance court administration worldwide. In the following section, we review the existing literature on court administration performance and the application of machine

learning techniques.

1.1 Literature Review

In recent years, there has been increased attention to the importance of measuring court administration performance, and techniques from machine learning have been employed to analyze performance indicators. Previous research on court administration performance and predictive analytics has focused on two key areas: predictive modeling of court case outcomes and case management optimization.

One stream of research has applied machine learning algorithms to predict case outcomes, such as the likelihood of a case being granted, dismissed, or settled. Studies by Medvedeva et al. (2020) and Corbett-Davies and Goel (2018) have demonstrated the effectiveness of supervised machine learning techniques in modeling court decisions. Similarly, predictive analytics has been employed to evaluate factors influencing judicial efficiency and fairness (Fu et al., 2023).

Another area of research has explored the application of machine learning in case management. Oliveira de et al. (2022) used machine learning to estimate trial durations, while Chhatwal et al. (2017) demonstrated how predictive models can enhance document discovery in litigation. These studies highlight the potential of machine learning in improving resource allocation and reducing case backlogs.

While these studies provide valuable insights, they do not specifically address the performance evaluation of judicial units using a comprehensive set of trial court processing indicators. Moreover, limited research has been conducted on applying machine learning to Iranian judicial data. This study extends the existing body of knowledge by utilizing a structured dataset from the Iranian judiciary to build predictive models that assess court administration performance.

By bridging this gap, we contribute to the growing field of data-driven decision-making in the legal sector. Our study applies machine learning not only to predict performance outcomes but also to identify the most significant factors influencing judicial efficiency, thereby providing a practical framework for court performance monitoring and improvement.

2 Materials and Methods

The dataset used in this study was obtained from the Statistical and Information Technology Center of the Judiciary of Iran, which systematically collects and maintains case management records from trial courts. These records are administrative reports compiled monthly by court staff and include key performance indicators such as case referrals, resolutions, pending caseloads, and processing times. The dataset covers 21 branches of the General Court of Law in Tehran from April 2022 to September 2022, though some branches have missing data for certain months. The final dataset consists of 119 case management records, providing a structured basis for analyzing court performance. The number of working days in a month reflects the operational capacity of courts, typically ranging between 18 and 21 days per month. Pending cases at the beginning of a period represent the number of cases awaiting resolution at the start of a given period, which is measured in months. The number

of cases referred to a judge indicates the caseload assigned to judges during a given month. The number of resolved cases measures judicial efficiency by counting the number of cases concluded within a month. The pending trial caseload represents the number of cases committed for trial but not yet finalized at the end of the reporting period. Precautionary or monitoring time refers to the duration for which a case is temporarily put on hold for additional review or investigation, measured in days. Processing time captures the time elapsed between case readiness for trial and the actual hearing date, also measured in days. Precautionary or monitoring time period is the total period during which a case remains suspended before proceeding, measured in days. The final decision number is the total number of judgments, verdicts, or rulings issued by the court within the given month. The average entry processing time represents the duration taken from the last case registered for trial to its finalization, measured in days.

The dataset was obtained from the Statistical and Information Technology Center of the Judiciary of Iran, ensuring reliability and authenticity. The unit of time used in all processing and monitoring indicators is days, while periods referenced in caseload assessments reflect a monthly reporting cycle.

Table 1: Summary Statistics of Court Performance Indicators

Variable	Mean	Min	Max	Standard Deviation
Number of working days in a month (days)	19.89	18	21	1.21
Pending cases at beginning of period (cases)	599.59	0	1154	282.46
Number of cases referred to a judge (cases)	180.25	8	246	42.93
Number of resolved cases (cases)	171.15	14	277	61.10
Pending trial caseload (cases)	135.10	3	251	43.62
Precautionary/monitoring time (days)	158.97	0	778	101.15
Processing time (in days)	87.80	4	1589	199.99
Precautionary/monitoring time period (days)	87.53	7	2624	255.25
Final decision number (cases)	168.59	14	277	60.43
Average entry processing time (days)	66.72	8	203	34.86

The descriptive statistics in Table 1 provide valuable insights into the workload and efficiency of trial courts in Tehran. The number of working days per month remains relatively stable across courts, suggesting consistency in operational schedules. The large variation in pending cases, ranging from zero to over 1,150, highlights disparities in case backlogs among courts. Similarly, the number of cases referred to judges varies significantly, reflecting differences in workload distribution. The high standard deviation in precautionary and processing times suggests that some cases experience substantial delays, indicating potential inefficiencies in case management. The number of resolved cases per period is close to the number of referred cases, indicating that courts are keeping up

with newly received cases. However, the presence of a substantial pending trial caseload suggests that backlog reduction remains a challenge. These findings underscore the importance of data-driven decision-making in judicial resource allocation and performance improvement. The dataset reveals that only 0.57% of courts demonstrate positive performance, indicating that the majority of judicial units struggle to meet efficiency benchmarks. This finding highlights potential systemic inefficiencies in case resolution and resource allocation.

To analyze this data, we implemented supervised machine learning techniques to classify court performance into two categories: "positive" (efficient case resolution) and "negative" (inefficiencies or delays). The supervised learning approach was chosen due to its ability to learn patterns from labeled historical data and make accurate predictions on new cases.

The machine learning models applied in this study include logistic regression, gradient boosting, neural networks, stochastic gradient descent (SGD), Support Vector Machines (SVM), Decision trees, random forest, k-Nearest Neighbors (kNN), Naïve Bayes, and Adaptive Boosting (AdaBoost). These models were selected to ensure a diverse range of classification approaches, balancing computational efficiency, interpretability, and predictive accuracy.

The training process involved splitting the dataset into training (80%) and testing (20%) sets. Each model was trained on the training set using five-fold cross-validation to optimize hyperparameters and prevent overfitting. The trained models were then evaluated on the test set using classification accuracy, F1-score, precision, recall, and area under the curve (AUC) as performance metrics. As we said, in our study, we used Orange version 3.34.0 software to calculate various performance metrics for each of the classification models.

These metrics were computed based on three different approaches: Positive Class (Metrics were calculated for instances where the target variable was classified as "positive" (indicating efficient court performance)), Negative Class (Metrics were also calculated for instances classified as "negative" (indicating inefficiency or delays in court performance)), and Average Over Classes (In addition to the separate calculations for each class, we also computed the average of the performance metrics across both the positive and negative classes. This provides a more balanced overview of the model's overall performance, particularly in cases where the dataset may have an unequal distribution of instances across the two classes). This approach allows for a more detailed evaluation of the models, ensuring that we account for both the efficiency and inefficiency of court performance, while also considering the overall accuracy of the predictions.

After training, the best-performing model was deployed to predict court performance on unseen data. logistic regression emerged as the most effective model, achieving the highest AUC and classification accuracy. Feature importance analysis further revealed that the number of resolved cases and referred cases were the strongest predictors of court performance.

This structured methodology ensures transparency in how machine learning models were developed, trained, and applied to real-world judicial data, making the results both interpretable and actionable for court administrators and policymakers.

Table 2: Performance Metrics of the Ten Data Mining Models

Model (average over classes)	AUC	CA	F1	Precision	Recall
logistic regression	0.985	0.950	0.950	0.950	0.950
gradient boosting	0.955	0.899	0.899	0.899	0.899
neural network	0.933	0.882	0.881	0.883	0.882
SGD	0.932	0.941	0.941	0.944	0.941
SVM	0.929	0.824	0.820	0.827	0.824
random forest	0.848	0.748	0.747	0.747	0.748
tree	0.834	0.849	0.849	0.850	0.849
kNN	0.825	0.765	0.763	0.763	0.765
Naïve Bayes	0.816	0.731	0.732	0.733	0.731
AdaBoost	0.780	0.773	0.775	0.785	0.773
Model (positive class)	AUC	CA	F1	Precision	Recall
logistic regression	0.985	0.950	0.957	0.957	0.957
gradient boosting	0.959	0.899	0.914	0.914	0.914
SGD	0.928	0.941	0.952	0.920	0.986
SVM	0.925	0.824	0.859	0.810	0.914
neural network	0.920	0.882	0.903	0.878	0.929
random forest	0.853	0.748	0.789	0.778	0.800
kNN	0.825	0.765	0.806	0.784	0.829
tree	0.819	0.849	0.870	0.882	0.857
Naïve Bayes	0.814	0.731	0.768	0.779	0.757
AdaBoost	0.777	0.773	0.794	0.852	0.743
Model (negative class)	AUC	CA	F1	Precision	Recall
logistic regression	0.985	0.950	0.939	0.939	0.939
gradient boosting	0.959	0.899	0.878	0.878	0.878
SGD	0.928	0.941	0.925	0.977	0.878
SVM	0.925	0.824	0.764	0.850	0.694
neural network	0.920	0.882	0.851	0.889	0.816
random forest	0.854	0.748	0.688	0.702	0.673
kNN	0.825	0.765	0.702	0.733	0.673
tree	0.819	0.849	0.820	0.804	0.837
Naïve Bayes	0.814	0.731	0.680	0.667	0.694
AdaBoost	0.777	0.773	0.748	0.690	0.816

3 Results

As mentioned, ten different data mining models were employed to classify the outcome into positive or negative using ten independent variables (detailed in Table 1). Table 2 presents the model fitting performance, which assesses how well each algorithm captured relationships within the training data. The classification accuracy (CA), area under the curve (AUC), and F1-score in this table indicate the effectiveness of each model in identifying meaningful patterns from past judicial data.

The analysis indicated that logistic regression was the best-performing model across all three cases (Positive Class, Negative Class, and Average Over Classes), achieving an AUC and CA of 98.5% and 95.0%, respectively. However, when the target classification is 'negative', the sensitivity is lower when compared to classifying 'positive'. Notably, the model performed better for the 'positive' target class and worse for the 'negative,' possibly due to unequal class sizes. While the AUC and CA values were uniform across all three cases, there were notable differences in F1-score, precision, and recall values. Similar differences were observed with the other nine models evaluated in this study. Overall, the results suggest that logistic regression is the best model for the classification task, with consistent AUC and CA performance across all three target classes. However, there were variations in the model's ability to predict 'positive' and 'negative' target classes, indicating the need to further explore the class imbalance.

Table 3: Classification Instances for Machine Learning Models

Model	TP	FP	FN	TN	Correct	Incorrect
SGD	69	5	1	44	113	6
logistic regression	67	3	3	46	113	6
SVM	64	15	6	34	98	21
Naïve Bayes	53	15	17	34	87	32
neural network	65	9	5	40	105	14
kNN	58	16	12	33	91	28
gradient boosting	64	6	6	43	107	12
tree	60	8	10	41	101	18
random forest	58	12	12	37	95	24
AdaBoost	52	9	18	40	92	27

Table 3 reports the models' classification performance on the testing dataset, measuring their ability to generalize to new instances. This table demonstrates how well each trained model predicts court performance outcomes when applied to previously unseen court records. The results in Table 3 provide insight into each model's effectiveness in distinguishing between "positive" and "negative" judicial unit performance in real-world scenarios. The matrix comprises True Positives (TP), False

Positives (FP), False Negatives (FN), and True Negatives (TN). The logistic regression and SGD models correctly classified 113 out of 119 instances, with only 6 out of 119 misclassified.

Generally, the number of models in which the number of false positives is lower than the number of false negatives is equal to the number of models in which the number of false positives is more or equal to the number of false negatives. It means that Type I errors and Type II errors are almost equal.

Regarding the best-performing model, logistic regression's success can be attributed to its utilization of the new feature introduced in Orange Software, as demonstrated in Fig 1.

In summary, the confusion matrix assessed the model's effectiveness and classified instances into TP, FP, FN, and TN. logistic regression outperformed other models, achieving high classification accuracy and minimal misclassification. All models demonstrated lower false positive rates than false negative rates, correlating with fewer Type I and more Type II errors.

As depicted in Fig 1, the key predictor of positive court performance was the number of resolved cases during a specified period. The number of cases referred to a court judge also played a significant role in predicting positivity. Since, red colour represents higher feature value, while blue colour is a lower value and the positive points (points right from the centre) in Fig 1 are feature values with the impact toward the prediction for the selected class, Obviously, Increasing the number of resolved cases and reducing the number of referred cases leads to an increase in the performance of the courts.

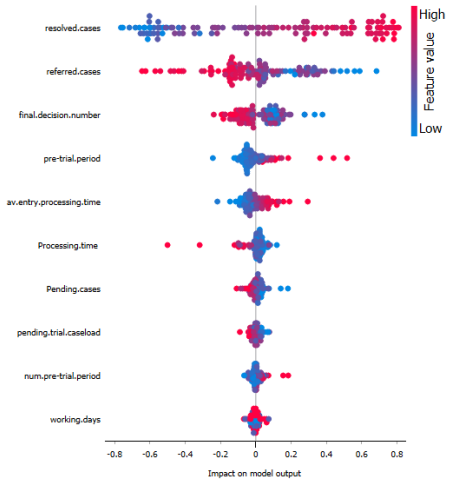

Figure 1: The ranking of the impact of the variables obtained using logistic regression model

Fig 2 can help us to determine which features most contributed to the prediction (features with longer tape length) and how they affect it. So, the number of resolved cases during the specified period emerged as the key contributor to increase the probability of positive court performance. In other word, as the number of resolved cases increases, the probability of positive court performance

also tends to increase. Also, the probability of positive court performance tends to decrease with increasing the number of referred cases. The average probability of positive court performance in this dataset (baseline probability) is 0.52.

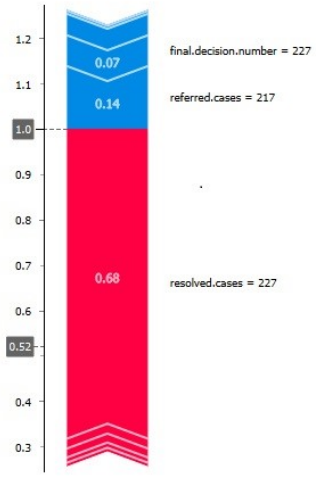

Figure 2: Features contribute the most to the prediction of performance of trial court administration for a single instance based on logistic regression model

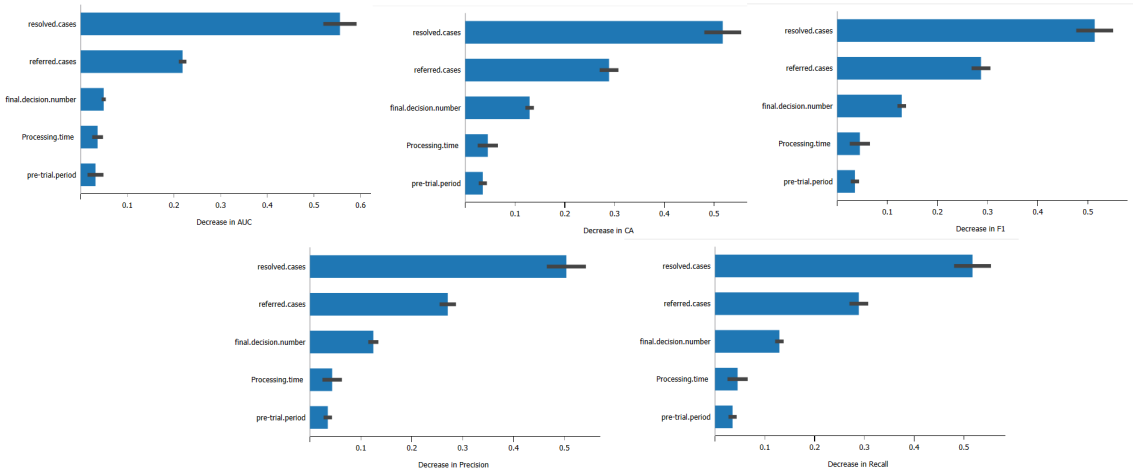

Figure 3: Features Importance based on all AUC, CA, F1-score, precision, and recall scores in logistic regression model

Additional analysis based on all AUC, CA, F1-score, precision, and recall scores confiremd that the number of resolved and reffered cases during a specified period are the most important features , as demonstrated in Fig. 3.

4 Conclusion

This study applied machine learning techniques to classify court performance based on past judicial behavior. While automatic legal analysis has a long history, our focus was specifically on machine learning approaches. The results demonstrate that machine learning models, particularly logistic regression, gradient boosting, neural networks, and stochastic gradient descent (SGD), can effectively predict court performance using ten key indicators from trial courts in Tehran, Iran. Among these models, logistic regression exhibited the highest classification accuracy.

Our findings contribute to the growing body of research supporting the use of data mining techniques in judicial performance assessment. By leveraging these models, decision support systems can be developed to enhance court monitoring and evaluation, ultimately strengthening public confidence in the judicial system. However, we acknowledge that disparities in data, including variations in features and model selection, may influence prediction accuracy. Expanding the dataset and improving data management practices in judicial institutions will be essential for enhancing the reliability of future predictions.

This study underscores the potential of machine learning in judicial performance assessment and provides a foundation for further research. Future studies should explore additional machine learning techniques and broader datasets to refine predictive accuracy. By advancing data-driven decision-making, these efforts can significantly improve court administration and the overall effectiveness of the justice system.

Acknowledgments

The authors would like to express their sincere gratitude to the Statistical and Information Technology Center of the Judiciary of Iran for their invaluable assistance and provision of data for the completion of this research. Without access to these essential resources, the present study would not have been feasible in its current form and quality.

Declarations

The authors declare that they have no conflicts of interest that could influence the outcomes or interpretations presented in this paper. The dataset utilized in this study is made accessible to the authors through the Statistical and Information Technology Center of the Judiciary of Iran. As per the policies of this organization, the raw data cannot be shared publicly. However, interested parties can initiate a formal request for data access by following the appropriate legal procedures. Requests for data access can be directed to info@eadl.ir or via the website of the Statistical and Information Technology Center at https://mafa.eadl.ir/. While the dataset's raw data cannot be publicly shared due to legal and organizational constraints, the authors acknowledge the importance of accessibility

through proper channels for interested parties to conduct valid and ethical analyses. The authors did not receive specific financial support or funding for working with the dataset.

References

Chhatwal, R., Huber-Fliflet, N., Keeling, R., Zhang, J., and Zhao, H. (2017, December), "Empirical evaluations of active learning strategies in legal document review," In *2017 IEEE International Conference on Big Data (Big Data)*, 1428–1437.

Corbett-Davies, S., and Goel, S. (2018), "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*.

Fu, C., Pang, H., Zhou, S., and Zhu, J. (2023), "Covariate handling approaches in combination with dynamic borrowing for hybrid control studies," *Pharmaceutical Statistics*.

Medvedeva, M., Vols, M., & Wieling, M. (2020), "Using machine learning to predict decisions of the European Court of Human Rights," *Artificial Intelligence and Law*, 28, 237–266.

Oliveira de, R. S., Reis Jr, A. S., and Sperandio Nascimento, E. G. (2022), "Predicting the number of days in court cases using artificial intelligence," *PloS One*, 17(5), e0269008.

Ostrom, B. J., & Hanson, R. A. (2000). "Efficiency, timeliness, and quality: A new perspective from nine state criminal trial courts". *US Department of Justice, Office of Justice Programs, National Institute of Justice.*

Steelman, D. C., Goerdt, J., & McMillan, J. E. (2000). "Caseflow management: The heart of court management in the new millennium". *Williamsburg, VA: National Center for State Courts*, 43–137.

Received: August 4, 2024

Accepted: March 7, 2025