Predictive Justice in Data Privacy: Explaining GDPR Fine Determinants Through Legal Feature Engineering and Ensemble Machine Learning Models

Ahlam Alshreef[1,*]

[1]Business Administration Department, Amman Arab University, Jordan

ABSTRACT

The General Data Protection Regulation (GDPR) establishes a harmonized legal framework for personal data protection across the European Union. However, the substantial variation in administrative fines imposed by different supervisory authorities raises questions regarding the consistency and predictability of GDPR enforcement. This study aims to develop an explainable predictive justice framework for estimating GDPR administrative fines using legal feature engineering and ensemble machine learning while identifying the legal and institutional factors that most strongly influence enforcement outcomes. The proposed framework utilizes a dataset of 437 GDPR enforcement cases and applies logarithmic transformation to address the highly skewed distribution of administrative fines. Legal and contextual information are transformed into predictive features through one hot encoding and feature engineering, including violated GDPR articles, supervisory authorities, countries, violation types, case complexity, and the number of violated articles. The predictive performance of a Random Forest ensemble model is compared with a Linear Regression baseline using the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Experimental results demonstrate that the proposed Random Forest model significantly outperforms the baseline model, achieving an R^2 value of 0.800, compared with 0.530 for Linear Regression, with an MAE of 0.8545 and an RMSE of 1.1750. Feature importance analysis reveals that Summary Length, Supervisory Authority, Country, and Violation Type are the most influential predictors, whereas individual GDPR articles contribute comparatively smaller importance values. These findings indicate that administrative fine determination is influenced not only by statutory provisions but also by institutional and contextual characteristics of enforcement cases. The proposed framework contributes to the field of cyberlaw by demonstrating that explainable ensemble machine learning can effectively support predictive justice analysis while providing interpretable empirical evidence regarding regulatory consistency and legal decision making under the GDPR.

Keywords Predictive Justice, General Data Protection Regulation (GDPR), Cyberlaw, Legal Feature Engineering, Random Forest

Submitted 2 June 2026
Accepted 15 July 2026
Published 2 November 2026

Corresponding author
Ahlam Alshreef,
a.alshreef@aau.edu.jo

Additional Information and Declarations can be found on page 255

DOI: 10.47738/jcl.v2i4.45

Copyright
2026 Alshreef

Distributed under
Creative Commons CC-BY 4.0

INTRODUCTION

The rapid growth of digital technologies has significantly increased the volume of personal data processed by both public and private organizations. Cloud computing, social media platforms, artificial intelligence, and Internet of Things (IoT) applications continuously generate and exchange personal information, creating new challenges for privacy protection and cybersecurity governance. In response to these developments, the European Union enacted the General

How to cite this article: A. Alshreef, "Predictive Justice in Data Privacy: Explaining GDPR Fine Determinants Through Legal Feature Engineering and Ensemble Machine Learning Models," *J. Cyber. Law.*, vol. 2, no. 4, pp. 239-257, 2026.

Data Protection Regulation (GDPR), which came into force in May 2018 as one of the world's most comprehensive legal frameworks for personal data protection [1]. The GDPR establishes a harmonized regulatory system intended to ensure consistent protection of individual privacy rights while imposing substantial administrative fines on organizations that violate its provisions. Through Article 83, supervisory authorities are empowered to impose penalties that are effective, proportionate, and dissuasive, with maximum fines reaching €20 million or 4% of an organization's annual global turnover.

Although the GDPR provides a unified legal framework for all European Union Member States, considerable differences in administrative fine decisions have been observed across supervisory authorities [2]. Organizations committing similar categories of violations may receive substantially different penalties depending on the jurisdiction, regulatory authority, and contextual circumstances of the case. Such variation raises important questions regarding legal certainty, regulatory consistency, and transparency in GDPR enforcement. From the perspective of cyberlaw, understanding the determinants of administrative fines is essential not only for regulatory compliance but also for evaluating whether enforcement decisions remain consistent with the harmonization objectives established by the GDPR.

Recent advances in artificial intelligence and machine learning have created new opportunities for computational legal analysis. Rather than replacing legal decision makers, machine learning techniques can assist researchers in identifying hidden patterns within historical legal decisions, estimating future outcomes, and evaluating consistency in regulatory enforcement. This interdisciplinary field, commonly referred to as predictive justice, applies computational models to analyze legal data while maintaining interpretability and transparency [3]. In the context of GDPR enforcement, predictive justice offers a data-driven approach for examining how statutory provisions, institutional characteristics, and case complexity collectively influence administrative fine determination.

Several previous studies have investigated GDPR enforcement from legal, economic, and statistical perspectives. Existing research has primarily examined trends in administrative fines, the interpretation of Article 83, compliance requirements, and the legal implications of data protection regulation. Other studies have employed machine learning techniques for legal prediction tasks, including judicial decision prediction, legal document classification, and legal text analytics. More recently, ensemble learning algorithms such as Random Forest have demonstrated strong predictive performance in complex regression problems because of their ability to capture nonlinear relationships and interactions among variables [4]. However, these studies generally focus on predictive accuracy and rarely incorporate legally interpretable features that explicitly represent statutory provisions and institutional characteristics.

Research Gap. Despite the growing body of literature on GDPR enforcement and legal analytics, several important gaps remain. First, most GDPR studies adopt descriptive or doctrinal approaches that explain legal provisions without quantitatively modeling administrative fine determination. Second, existing predictive studies frequently emphasize algorithmic performance while treating legal information as ordinary categorical variables, thereby limiting legal interpretability. Third, relatively few studies investigate how legal provisions,

supervisory authorities, geographical jurisdictions, and contextual characteristics jointly influence administrative fine decisions within a unified predictive framework. Consequently, empirical evidence regarding the relative contribution of legal and institutional factors to GDPR enforcement remains limited.

State of the Art. Recent developments in computational legal research increasingly emphasize the integration of explainable artificial intelligence with domain-specific feature engineering to improve both predictive performance and model interpretability [5]. Rather than relying solely on raw textual or categorical information, modern legal analytics transforms legal concepts into structured computational representations that can be interpreted alongside prediction results. Building upon this development, the present study proposes a legal feature engineering framework that converts statutory provisions, supervisory authorities, countries, violation types, case complexity, and the number of violated GDPR articles into meaningful predictive variables. These engineered legal features are combined with one hot encoding, logarithmic transformation to address the highly skewed distribution of administrative fines, and a Random Forest ensemble regression model capable of modeling nonlinear relationships while providing interpretable feature importance.

Accordingly, this study aims to develop an explainable predictive justice framework for estimating GDPR administrative fines using legal feature engineering and ensemble machine learning. Specifically, the study seeks to (1) construct a predictive model capable of handling the highly skewed distribution of GDPR administrative fines, (2) compare the predictive performance of Random Forest and Linear Regression models, and (3) identify the legal, institutional, and contextual factors that most strongly influence administrative fine determination through feature importance analysis. By integrating computational methods with legal interpretation, this research contributes to the fields of cyberlaw, data privacy, and computational legal studies by providing empirical evidence regarding the determinants of GDPR enforcement and the practical application of explainable artificial intelligence for predictive justice.

Literature Review and Related Works

The rapid advancement of digital technologies has transformed the way personal data are collected, processed, and exchanged across digital platforms. This development has simultaneously increased concerns regarding privacy protection, cybersecurity, and regulatory accountability. To address these challenges, the European Union introduced the General Data Protection Regulation (GDPR) as a comprehensive legal framework that establishes harmonized principles for the protection of personal data while strengthening the rights of data subjects and the responsibilities of organizations processing personal information [6]. The regulation also introduces administrative fines as an enforcement mechanism intended to ensure that data controllers and processors comply with the legal obligations defined by the GDPR [7].

Under Article 83 of the GDPR, supervisory authorities are authorized to impose administrative fines that are effective, proportionate, and dissuasive according to the circumstances of each infringement [8]. The amount of a fine is determined by considering multiple legal factors, including the nature and severity of the infringement, the duration of the violation, the degree of

responsibility of the organization, previous infringements, cooperation with supervisory authorities, and measures taken to mitigate potential harm [9]. Although these legal principles are intended to promote harmonized enforcement across the European Union, previous studies have reported noticeable differences in administrative sanctions imposed by different supervisory authorities, suggesting that institutional practices and national enforcement strategies may influence regulatory outcomes [10].

The increasing availability of legal datasets has encouraged the application of artificial intelligence and machine learning techniques within the field of computational legal studies. One emerging research direction is predictive justice, which utilizes computational models to analyze historical legal decisions and identify patterns that may improve the understanding of legal consistency and decision-making processes [11]. Rather than replacing legal professionals, predictive justice aims to support evidence-based legal analysis by providing quantitative insights into judicial and administrative decisions while maintaining transparency and accountability [12].

Supervised machine learning algorithms have been widely adopted for legal prediction tasks because of their ability to discover complex relationships within structured legal data. Previous studies have successfully applied machine learning to judicial decision prediction, legal document classification, legal information retrieval, sentencing analysis, and regulatory compliance assessment [13]. These studies demonstrate that legal outcomes often depend on interactions among multiple variables rather than a single statutory provision, making machine learning particularly suitable for modeling complex legal decision-making processes [14].

Among supervised learning algorithms, ensemble learning techniques have consistently demonstrated superior predictive performance for heterogeneous datasets containing nonlinear relationships. Random Forest is one of the most widely adopted ensemble algorithms because it combines the predictions of multiple decision trees to improve robustness while reducing prediction variance [15]. In addition to achieving high predictive accuracy, Random Forest naturally provides feature importance measures that allow researchers to identify the relative contribution of each predictor, thereby improving model transparency and supporting explainable artificial intelligence [16].

Recent developments in explainable artificial intelligence have emphasized the importance of feature engineering for transforming domain knowledge into computationally meaningful variables. Within legal analytics, feature engineering enables statutory provisions, institutional characteristics, geographical information, and contextual case attributes to be represented as structured predictors while preserving their legal interpretation [17]. This approach not only improves predictive performance but also allows legal researchers to interpret model outputs in accordance with established legal concepts, making computational models more transparent and practically applicable in legal research [18].

Despite these developments, important research challenges remain. Existing GDPR studies predominantly focus on legal interpretation, compliance evaluation, or descriptive analyses of enforcement activities, whereas relatively few studies attempt to predict administrative fine amounts using explainable machine learning techniques [19]. Furthermore, studies that employ machine learning frequently prioritize predictive accuracy while providing limited

discussion regarding the legal interpretation of predictive variables and the institutional factors influencing enforcement decisions. Consequently, there remains a need for an interpretable computational framework capable of integrating legal provisions, institutional characteristics, and contextual information into predictive models that can simultaneously achieve high predictive performance and meaningful legal interpretation [20].

The present study addresses these limitations by proposing a predictive justice framework that integrates legal feature engineering, one hot encoding, logarithmic transformation, and Random Forest ensemble regression for predicting GDPR administrative fines. Unlike previous studies that primarily emphasize either doctrinal legal analysis or predictive modeling alone, this research combines explainable machine learning with legally meaningful feature engineering to investigate the determinants of GDPR enforcement. Through this integrated approach, the study contributes to the growing field of computational cyberlaw by providing empirical evidence regarding the relative influence of statutory provisions, institutional characteristics, and case complexity on administrative fine determination.

Methodology

This study proposes a predictive justice framework for estimating General Data Protection Regulation (GDPR) administrative fines by integrating legal feature engineering with ensemble machine learning. As illustrated in figure 1, the proposed methodology consists of six sequential stages: (1) data collection, (2) data preprocessing, (3) legal feature engineering and one hot encoding, (4) model training, (5) performance evaluation, and (6) feature importance analysis. The workflow begins with the collection of historical GDPR enforcement cases, followed by preprocessing procedures to address the highly skewed distribution of administrative fines. Next, legally meaningful variables are extracted through feature engineering and combined with encoded categorical attributes to construct the predictive feature matrix. The processed data are subsequently used to train both Linear Regression and Random Forest models, whose predictive performances are evaluated using standard regression metrics. Finally, feature importance analysis is conducted to identify the legal, institutional, and contextual factors that most strongly influence GDPR administrative fine determination.

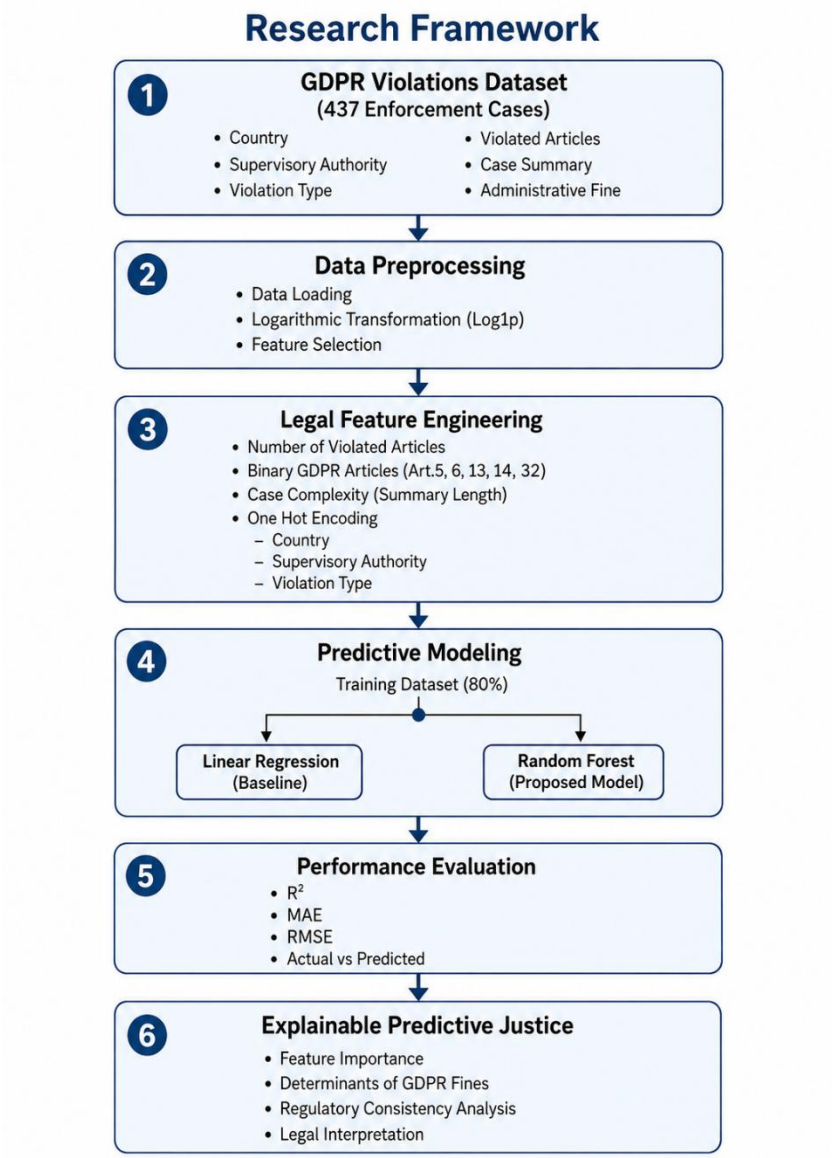


Figure 1 Research Methodology of the Proposed Predictive Justice Framework

Dataset Description

The dataset employed in this study consists of 437 historical GDPR administrative fine cases obtained from the publicly available GDPR Violations Dataset. Each observation represents a completed enforcement decision issued by a supervisory authority within the European Union. The dataset contains information describing the country, supervisory authority, violation type, violated GDPR articles, textual case summary, and the imposed administrative fine. The administrative fine (price) serves as the dependent variable, whereas the remaining attributes are utilized as predictor variables.

Because the administrative fines range from zero-value penalties to several million euros, the target variable exhibits a highly right-skewed distribution. Such characteristics require preprocessing before predictive modeling can be effectively performed.

Data Preprocessing

To reduce the influence of extreme observations, the administrative fine was transformed using the Log1p function:

$$y = \log(1 + \text{Fine}) \tag{1}$$

y denotes the transformed target variable,

Fine represents the original administrative fine.

The logarithmic transformation compresses extremely large fine values while preserving the relative ordering of observations, thereby producing a more balanced distribution suitable for regression analysis.

Three categorical variables were selected as institutional predictors:

Country

Violation Type

Supervisory Authority

These variables represent the geographical and regulatory characteristics of each GDPR enforcement case.

Legal Feature Engineering

The principal contribution of this research lies in transforming legal information into computationally meaningful predictors through legal feature engineering.

Number of Violated Articles

The number of violated GDPR provisions is calculated as:

$$\text{NumArticles} = \sum_{i=1}^{m} I_i \tag{2}$$

$I_i = 1$ if Article i appears in the violation record,

m represents the total number of violated articles.

This variable represents the legal complexity of each infringement.

Binary Representation of Key GDPR Articles

Five frequently violated GDPR articles were encoded as binary variables:

Article 5

Article 6

Article 13

Article 14

Article 32

Each feature is represented as:

$$A_j = \begin{cases} 1, & \text{if Article } j \text{ is violated} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

A_j denotes the binary representation of GDPR Article j.

Case Complexity

The complexity of an enforcement case is approximated by calculating the number of words contained in the summary:

$$\text{SummaryLength} = \sum_{k=1}^{n} 1 \quad (4)$$

n denotes the total number of words in the case summary.

Longer summaries generally indicate more comprehensive investigations involving multiple legal and factual considerations.

One Hot Encoding

Since machine learning algorithms require numerical input, categorical variables were transformed using One Hot Encoding.

For each categorical variable:

$$X = [x_1, x_2, \ldots, x_c] \quad (5)$$

c denotes the number of categories; each encoded feature satisfies:

$$x_i \in \{0,1\} \quad (6)$$

Only one category is assigned the value 1, while all remaining categories are assigned 0.

The encoded categorical variables were combined with the engineered legal features to form the final predictor matrix.

Prediction Models

Linear Regression serves as the baseline model and assumes a linear relationship between predictor variables and the target variable:

$$\hat{y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon \quad (7)$$

$\hat{y}$ denotes the predicted logarithmic fine,

x_i represents predictor variables,

β_i denotes regression coefficients,

ε is the residual error.

The proposed model utilizes Random Forest Regression, an ensemble learning algorithm that combines multiple decision trees.

The final prediction is computed as:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \quad (8)$$

T is the number of decision trees,

$f_t(x)$ is the prediction generated by tree t.

The Random Forest model was configured using:

Number of trees = 300

Maximum tree depth = 20

Random state = 42

Parallel processing enabled

Model Training

The complete dataset was divided into training and testing subsets using an 80:20 split.

The partitioning process is expressed as:

$$D = D_{train} \cup D_{test} \tag{9}$$

with

$$|D_{train}| = 0.8|D|, \qquad |D_{test}| = 0.2|D| \tag{10}$$

D denotes the complete dataset.

Both prediction models were trained using the training dataset and evaluated on the testing dataset.

Performance Evaluation

The predictive performance was evaluated using three regression metrics.

Coefficient of Determination (R^2)

$$xR^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

y_i is the observed value,

$\hat{y}_i$ is the predicted value,

$\bar{y}$ is the mean of the observed values.

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - \hat{y}_i| \tag{12}$$

Lower MAE values indicate higher predictive accuracy.

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

RMSE penalizes larger prediction errors more heavily than MAE.

Feature Importance Analysis

To improve model interpretability, feature importance scores were extracted from the trained Random Forest model.

The importance of feature j is computed as:

$$FI_j = \frac{1}{T}\sum_{t=1}^{T} I_j^{(t)} \quad (14)$$

FI_j denotes the importance score of feature j,

$I_j^{(t)}$ represents the importance contribution of feature j in decision tree t,

T is the total number of trees.

The resulting feature importance values were subsequently grouped into legally meaningful categories, including Country, Supervisory Authority, Violation Type, Number of Violated Articles, Article 5, Article 6, Article 13, Article 14, Article 32, and Summary Length. This analysis enables the identification of the legal, institutional, and contextual variables that contribute most significantly to GDPR administrative fine prediction and supports the interpretability of the proposed predictive justice framework.

Results

Distribution of GDPR Administrative Fines

Figure 2 illustrates the distribution of GDPR administrative fines after applying the logarithmic transformation (Log1p). The original GDPR fine data exhibit a highly right skewed distribution, characterized by a large concentration of relatively small penalties alongside a limited number of exceptionally large fines reaching several million euros. In addition, several observations correspond to zero-value fines, further increasing the imbalance of the dataset. Such a distribution violates the assumptions of normality and homoscedasticity that underlie conventional linear regression models, making the prediction task considerably more challenging. Without appropriate preprocessing, extreme observations would dominate the learning process and potentially bias the estimation of model parameters, resulting in reduced predictive performance.

To address this issue, a logarithmic transformation using the Log1p function was applied to the target variable before model training. This transformation compresses the scale of very large fines while preserving the relative ordering of all observations, including zero-value penalties. As shown in figure 2, the transformed data exhibit a substantially more balanced and approximately symmetric distribution, allowing the prediction models to learn more stable relationships between the explanatory variables and the target variable. The improved distribution reduces the influence of outliers, stabilizes the variance of the response variable, and provides a more suitable foundation for supervised machine learning algorithms, particularly the proposed Random Forest ensemble model.

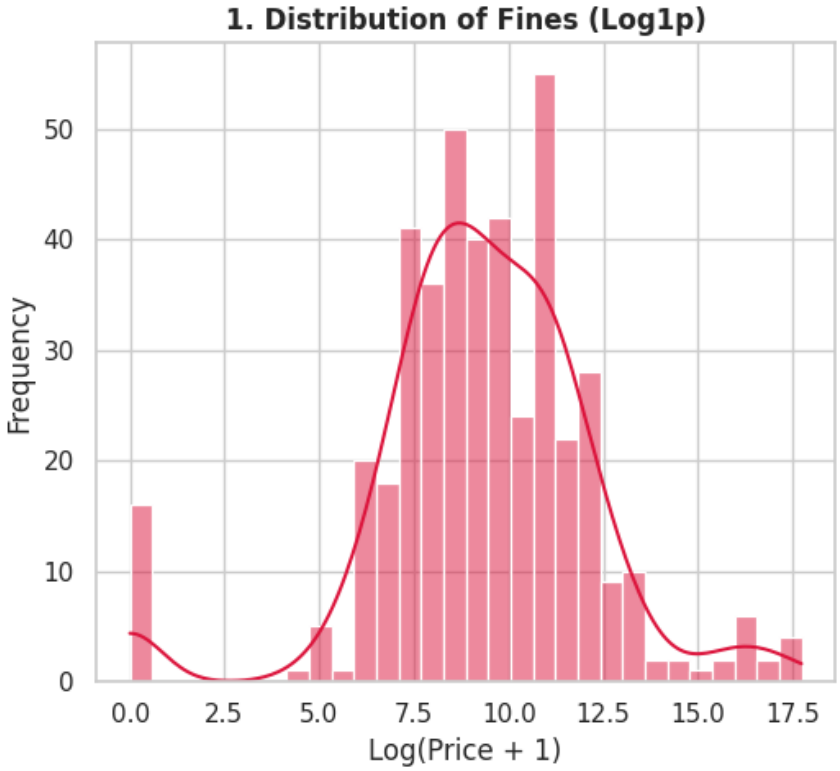


Figure 2 Distribution of GDPR Administrative Fines after Logarithmic Transformation.

Disparity of Administrative Fines Across Countries

Figure 3 presents the distribution of GDPR administrative fines across the ten countries with the highest number of recorded enforcement cases. The boxplots reveal noticeable differences in both the median values and the spread of administrative fines among jurisdictions. Italy and the United Kingdom exhibit the widest interquartile ranges and the greatest number of extreme observations, indicating that enforcement decisions in these countries vary substantially from one case to another. Germany and Spain also display relatively broad distributions, although with less pronounced variability. In contrast, countries such as the Czech Republic and Hungary show narrower interquartile ranges and fewer outliers, suggesting that the imposed fines are comparatively more concentrated within a limited range. These patterns demonstrate that the magnitude of GDPR administrative fines differs considerably across Member States despite being governed by the same regulatory framework.

The observed variation suggests that the practical implementation of GDPR enforcement is influenced by more than the statutory provisions specified in the regulation. While the GDPR establishes harmonized principles for determining administrative fines, supervisory authorities retain a degree of discretion when evaluating the severity, circumstances, and impact of individual cases. This institutional discretion may contribute to the differences observed across jurisdictions, resulting in varying enforcement practices even for similar categories of violations. Consequently, figure 3 provides preliminary empirical evidence that geographical and institutional characteristics may influence administrative fine determination, supporting the inclusion of country and

supervisory authority variables in the proposed predictive justice framework.

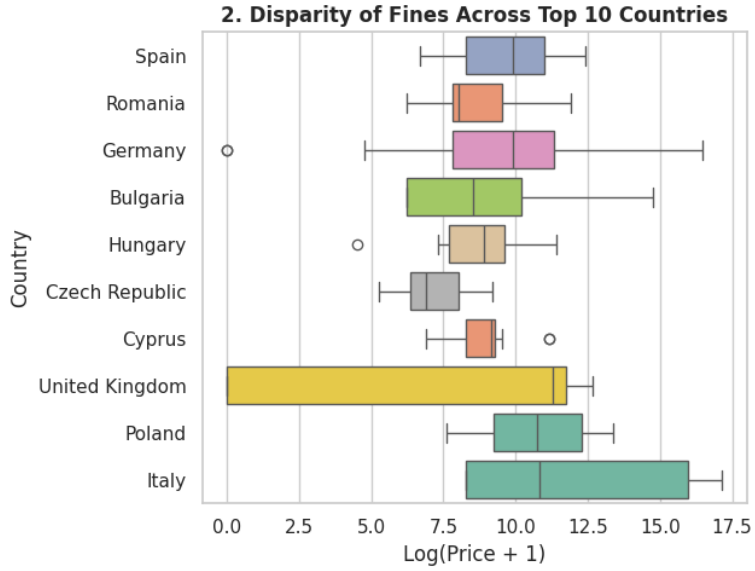


Figure 3 Disparity of GDPR Administrative Fines Across the Top Ten Countries.

Performance Comparison of Prediction Models

To evaluate the effectiveness of the proposed legal feature engineering framework, the Random Forest ensemble model was compared with a conventional Linear Regression model that served as the baseline. Linear Regression was selected because it is one of the most widely used statistical techniques for predictive modeling and provides an appropriate benchmark for assessing the benefits of nonlinear ensemble learning. Both models were trained and evaluated using the same training and testing datasets to ensure a fair comparison. Model performance was assessed using the coefficient of determination (R^2), while the Random Forest model was further evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to quantify prediction accuracy.

As presented in table 1, the proposed Random Forest model achieved an R^2 value of 0.800, substantially outperforming the Linear Regression baseline, which obtained an R^2 value of 0.530. This improvement indicates that the ensemble model explains approximately 80% of the variance in the logarithmically transformed GDPR fine values, whereas the baseline model explains only 53%. In addition, the Random Forest model produced a Mean Absolute Error (MAE) of 0.8545 and a Root Mean Squared Error (RMSE) of 1.1750, demonstrating relatively low prediction errors on the transformed target variable. These findings suggest that the proposed model is more effective in capturing the complex nonlinear relationships among legal, institutional, and contextual features, resulting in substantially improved predictive performance compared with the conventional linear approach.

Table 1 Performance Comparison Between Prediction Models.

Model	R^2	MAE	RMSE
Linear Regression	0.530	1.3386	1.8010
Random Forest	0.800	0.8545	1.1750

The comparative performance of both prediction models is further illustrated in figure 4. The figure clearly shows that the Random Forest model achieved a substantially higher coefficient of determination than the Linear Regression model, indicating a marked improvement in predictive capability. While the Linear Regression model obtained an R^2 value of 0.530, the proposed Random Forest model increased this value to 0.800, representing a considerable enhancement in the model's ability to explain the variability of GDPR administrative fines. The superior performance of the ensemble model demonstrates its effectiveness in learning complex patterns from the engineered legal features and categorical variables derived through one hot encoding.

The observed improvement suggests that the relationship between GDPR administrative fines and their explanatory variables cannot be adequately represented by a simple linear function. Instead, fine determination appears to be influenced by intricate interactions among statutory provisions, institutional characteristics, geographical factors, and case complexity. Random Forest is capable of modeling these nonlinear relationships by constructing multiple decision trees and aggregating their predictions, thereby reducing prediction variance and improving generalization performance. Consequently, the results shown in figure 4 provide empirical evidence that ensemble learning offers a more appropriate computational approach for predictive justice analysis than conventional linear regression when dealing with heterogeneous and highly skewed legal enforcement data.



Figure 4 Comparison of Prediction Performance Between Linear Regression and Random Forest Models.

Prediction Accuracy of the Proposed Model

Figure 5 compares the predicted GDPR administrative fines generated by the proposed Random Forest model with the corresponding observed values in the testing dataset. The scatter plot demonstrates that most observations are

concentrated near the 45-degree reference line, indicating strong agreement between the predicted and actual fine values. This distribution suggests that the model successfully captures the underlying relationship between the engineered legal features and the target variable across a broad range of fine values. The relatively dense clustering of points around the diagonal line further indicates that prediction errors remain small for the majority of cases, reflecting the model's ability to produce reliable estimates for typical GDPR enforcement decisions.

Although the overall prediction performance is strong, a limited number of observations corresponding to exceptionally large administrative fines deviate from the reference line, indicating slight underestimation by the model. These cases represent rare enforcement events that occur infrequently within the dataset and therefore provide fewer training examples for the learning algorithm. Such behavior is commonly observed in ensemble tree-based regression models because the final prediction is obtained by averaging the outputs of multiple decision trees, which naturally reduces the influence of extreme values. Despite these minor deviations, the overall distribution of predicted values closely follows the observed values, demonstrating that the proposed Random Forest model achieves stable predictive performance and satisfactory generalization ability for estimating GDPR administrative fines.

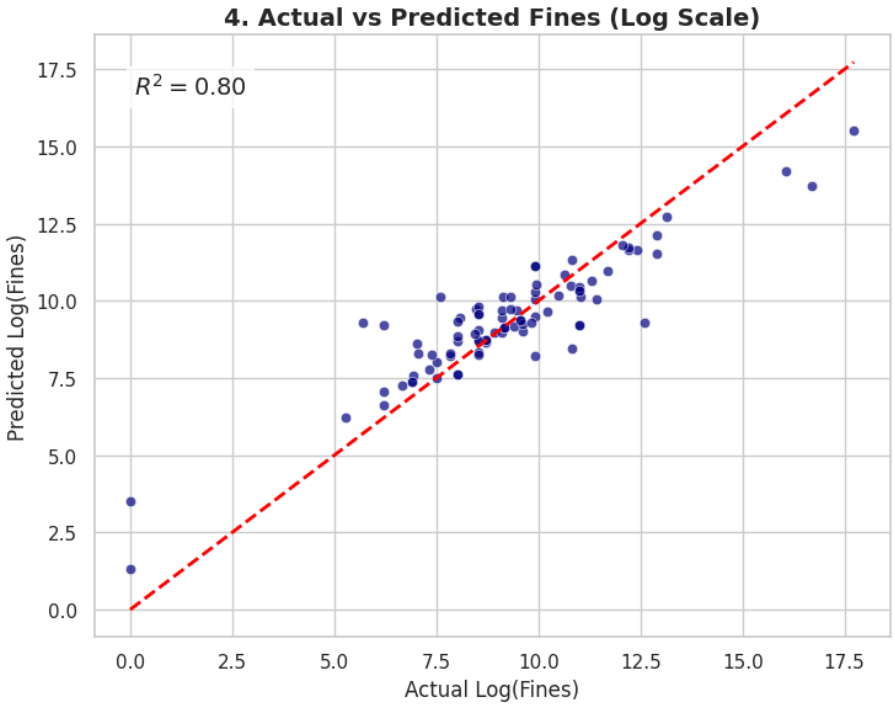


Figure 5 Actual versus Predicted GDPR Administrative Fines Using the Random Forest Model.

Feature Importance Analysis

To enhance the interpretability of the proposed predictive framework, feature importance was extracted from the trained Random Forest model and aggregated into legally meaningful categories, as presented in figure 6. Grouping the engineered variables into broader legal and institutional dimensions provides a clearer understanding of the factors that contribute to

GDPR administrative fine prediction. The results indicate that Summary Length is the most influential predictor, accounting for 35.1% of the total feature importance. This finding suggests that the complexity of a case, represented by the length of the enforcement summary, serves as a strong proxy for the severity and comprehensiveness of the underlying investigation. The second most influential factor is Supervisory Authority, contributing 26.8%, followed by Country with 17.2% and Violation Type with 12.3%. Together, these four categories explain the majority of the predictive capability of the proposed model.

In contrast, the individual statutory provisions represented by Article 5, Article 6, Article 13, Article 14, and Article 32 each contribute relatively small importance values, ranging from 0.6% to 2.2%. Although these articles constitute the legal basis for GDPR enforcement, the results indicate that their isolated presence is less informative than broader institutional and contextual characteristics when predicting the magnitude of administrative fines. This observation suggests that fine determination is influenced not only by the specific legal provisions that have been violated but also by the overall circumstances of the case and the supervisory authority responsible for enforcement. Consequently, the findings provide empirical support for the concept of predictive justice by demonstrating that explainable machine learning can reveal the combined influence of statutory, institutional, and contextual factors on regulatory decision making. The feature importance analysis therefore offers both predictive insight and legal interpretability, enabling a more comprehensive understanding of the determinants of GDPR administrative fines.

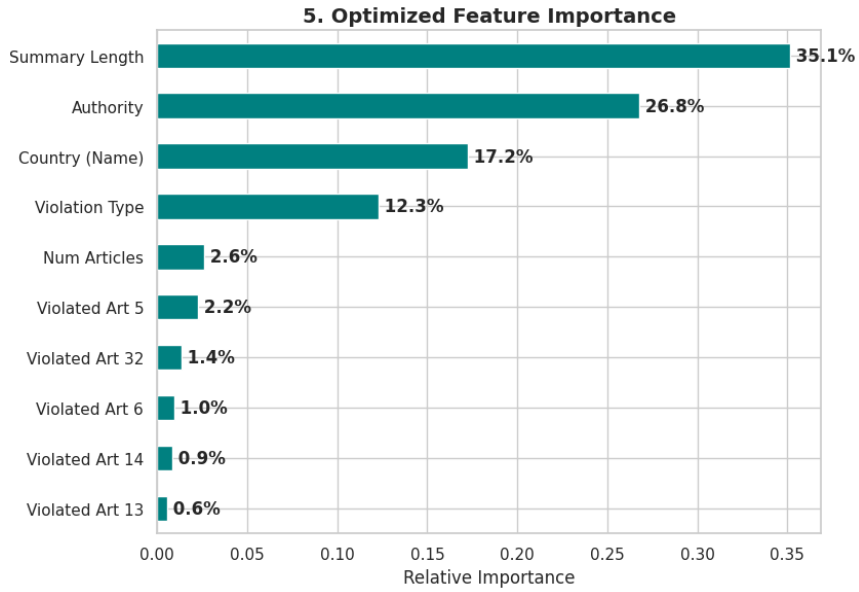


Figure 6 Grouped Feature Importance Derived from the Proposed Legal Feature Engineering Framework.

Discussion

The findings of this study demonstrate that the proposed legal feature engineering framework combined with the Random Forest ensemble algorithm provides an effective approach for predicting GDPR administrative fines. As presented in the Results section, the Random Forest model achieved an R^2

value of 0.800, substantially outperforming the baseline Linear Regression model, which achieved an R^2 value of 0.530. This considerable improvement indicates that the proposed model is capable of explaining a much larger proportion of the variation in GDPR fine amounts while maintaining relatively low prediction errors, as reflected by the obtained MAE and RMSE values. The superior performance of the ensemble model suggests that the relationships among legal provisions, institutional characteristics, geographical information, and case complexity are highly complex and cannot be adequately represented using a conventional linear regression model. Unlike Linear Regression, which assumes a linear relationship between predictor variables and the target variable, Random Forest is capable of identifying nonlinear interactions and hidden patterns by combining the predictions of multiple decision trees. Consequently, the proposed framework is better suited for modeling the heterogeneous characteristics of GDPR enforcement data [21].

From the perspective of computational legal studies, these findings demonstrate that machine learning can provide meaningful analytical support for understanding regulatory enforcement while maintaining model interpretability through legal feature engineering [22]. The improvement in predictive performance is not merely a technical achievement but also indicates that GDPR administrative fine determination follows identifiable patterns that can be learned from historical enforcement data [23]. The engineered legal variables, including statutory provisions, supervisory authorities, countries, violation types, and indicators of case complexity, collectively enable the model to capture important aspects of regulatory decision making [24]. This suggests that administrative fines are determined through the interaction of multiple legal and institutional factors rather than by any single variable in isolation. Therefore, the proposed approach contributes to the growing field of predictive justice by demonstrating that explainable ensemble machine learning can be employed as a practical analytical tool to examine consistency, transparency, and decision-making patterns within the enforcement of European data protection law [25].

Conclusion

This study proposed a predictive justice framework for estimating GDPR administrative fines by integrating legal feature engineering with a Random Forest ensemble learning model. The proposed framework transformed legal and institutional information into machine learning features through logarithmic transformation, one hot encoding, and engineered legal variables representing violated GDPR articles, supervisory authorities, countries, violation types, and case complexity. Using a dataset of 437 GDPR enforcement cases, the Random Forest model demonstrated superior predictive performance compared with the baseline Linear Regression model. The proposed model achieved an R^2 value of 0.800, while the baseline model achieved an R^2 value of 0.530, indicating that ensemble learning is substantially more effective for modeling the complex and highly skewed distribution of GDPR administrative fines. The model also produced relatively low prediction errors, with an MAE of 0.8545 and an RMSE of 1.1750, demonstrating its ability to generate reliable predictions on logarithmically transformed fine values.

Feature importance analysis further revealed that Summary Length, Supervisory Authority, Country, and Violation Type were the most influential predictors of administrative fines, whereas individual statutory provisions contributed comparatively smaller importance values. These findings indicate

that GDPR fine determination is influenced not only by the legal provisions that have been violated but also by institutional and contextual characteristics surrounding each enforcement case. The results therefore provide empirical evidence that legal outcomes can be effectively analyzed using explainable machine learning while maintaining meaningful legal interpretation. From the perspective of cyberlaw, the findings contribute to the understanding of predictive justice by demonstrating that historical enforcement decisions contain identifiable patterns that can be modeled computationally. At the same time, the substantial influence of supervisory authorities and geographical factors suggests that institutional practices remain an important component of GDPR enforcement despite the existence of a harmonized regulatory framework.

Although the proposed framework achieved promising predictive performance, several limitations should be acknowledged. The study was conducted using a relatively limited dataset of 437 historical GDPR enforcement cases and employed a single train test split for model evaluation. In addition, the predictive model considered only a selected set of legal and contextual variables and did not incorporate other potentially relevant factors such as organizational size, industry sector, duration of the infringement, previous compliance history, or intentional misconduct. Future research should expand the dataset, employ cross validation techniques to improve model robustness, compare additional ensemble learning algorithms such as Extreme Gradient Boosting and Light Gradient Boosting Machine, and incorporate explainable artificial intelligence methods such as SHAP to provide a more comprehensive interpretation of individual prediction outcomes. These improvements would further strengthen the applicability of predictive justice frameworks for supporting evidence-based research and decision analysis in cyberlaw and data protection regulation.

Declarations

Author Contributions

Conceptualization: A.A.; Methodology: A.A.; Software: A.A.; Validation: A.A.; Formal Analysis: A.A.; Investigation: A.A.; Resources: A.A.; Data Curation: A.A.; Writing Original Draft Preparation: A.A.; Writing Review and Editing: A.A.; Visualization: A.A.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

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.

References

[1] C. J. Hoofnagle, B. van der Sloot, and F. J. Zuiderveen Borgesius, “The European Union general data protection regulation: What it is and what it means,” *Information & Communications Technology Law*, vol. 28, no. 1, pp. 65–98, 2019, doi: 10.1080/13600834.2019.1573501.

[2] G. Gentile and O. Lynskey, “Deficient by design? The transnational enforcement of the GDPR,” *International & Comparative Law Quarterly*, vol. 71, no. 4, pp. 799–830, 2022, doi: 10.1017/S0020589322000355.

[3] M. Medvedeva, M. Vols, and M. Wieling, “Using machine learning to predict decisions of the European Court of Human Rights,” *Artificial Intelligence and Law*, vol. 28, no. 2, pp. 237–266, 2020, doi: 10.1007/s10506-019-09255-y.

[4] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[5] K. Atkinson, T. Bench-Capon, and D. Bollegala, “Explanation in AI and law: Past, present and future,” *Artificial Intelligence*, vol. 289, no. December, Art. no. 103387, 2020, doi: 10.1016/j.artint.2020.103387.

[6] P. Tikkinen-Piri, A. Rohunen, and J. Markkula, “EU General Data Protection Regulation: Changes and implications for personal data collecting companies,” *Computer Law & Security Review*, vol. 34, no. 1, pp. 134–153, 2018, doi: 10.1016/j.clsr.2017.05.015.

[7] M. N. Lintvedt, “Putting a price on data protection infringement,” *International Data Privacy Law*, vol. 12, no. 1, pp. 1–15, 2022, doi: 10.1093/idpl/ipab024.

[8] O. Lynskey, “Complete and effective data protection,” *Current Legal Problems*, vol. 76, no. 1, pp. 297–344, 2023, doi: 10.1093/clp/cuad009.

[9] M. Saemann, D. Theis, T. Urban, and M. Degeling, “Investigating GDPR fines in the light of data flows,” *Proceedings on Privacy Enhancing Technologies*, vol. 2022, no. 4, pp. 314–331, 2022, doi: 10.56553/popets-2022-0111.

[10] I. Sivan-Sevilla, “Varieties of enforcement strategies post-GDPR: A fuzzy-set qualitative comparative analysis (fsQCA) across data protection authorities,” *Journal of European Public Policy*, vol. 31, no. 2, pp. 552–585, 2024, doi: 10.1080/13501763.2022.2147578.

[11] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, “Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective,” *PeerJ Computer Science*, vol. 2, no. October, Art. no. e93, 2016, doi: 10.7717/peerj-cs.93.

[12] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[13] D. M. Katz, M. J. Bommarito II, and J. Blackman, “A general approach for predicting the behavior of the Supreme Court of the United States,” *PLOS ONE*, vol. 12, no. 4, Art. no. e0174698, 2017, doi: 10.1371/journal.pone.0174698.

[14] E. Jacob de Menezes-Neto and M. B. M. Clementino, “Using deep learning to

predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts," *PLOS ONE*, vol. 17, no. 7, Art. no. e0272287, 2022, doi: 10.1371/journal.pone.0272287.

[15] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000, doi: 10.1023/A:1007607513941.

[16] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the Gini importance?," *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, 2018, doi: 10.1093/bioinformatics/bty373.

[17] J. F. Muñoz Soro and C. Serrano-Cinca, "A model for predicting court decisions on child custody," *PLOS ONE*, vol. 16, no. 10, Art. no. e0258993, 2021, doi: 10.1371/journal.pone.0258993.

[18] S. M. Lundberg, G. Erion, H. Chen, et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. January, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[19] J. Ruohonen and K. Hjerppe, "The GDPR enforcement fines at glance," *Information Systems*, vol. 106, no. May, Art. no. 101876, 2022, doi: 10.1016/j.is.2021.101876.

[20] 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.

[21] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[22] B. Verheij, "Artificial intelligence as law," *Artificial Intelligence and Law*, vol. 28, no. 2, pp. 181–206, 2020, doi: 10.1007/s10506-020-09266-0.

[23] N. Mike, "Data protection has entered the chat: Analysis of GDPR fines," *Masaryk University Journal of Law and Technology*, vol. 16, no. 2, pp. 163–213, 2022, doi: 10.5817/MUJLT2022-2-3.

[24] S. Li, H. Zhang, L. Ye, X. Guo, and B. Fang, "MANN: A multichannel attentive neural network for legal judgment prediction," *IEEE Access*, vol. 7, no. October, pp. 151144–151155, 2019, doi: 10.1109/ACCESS.2019.2945771.

[25] P. M. Baquero, A. Wang, and D. Restrepo Amariles, "Unveiling transparency in data protection enforcement across the EU: Assessing the level and quality of disclosure of GDPR fines by data protection authorities," *European Law Journal*, vol. 31, no. 3, pp. 177–202, 2025, doi: 10.1111/eulj.70008.