Predicting Immigration Judge Decisions: A Machine Learning Approach*

INTRODUCTION TO MACHINE LEARNING - FINAL ASSIGNMENT

Francesco Colzi
Alma Mater - University of Bologna
MSc in Applied Economics and Markets
ID: 0001127782
francesco.colzi@studio.unibo.it

Ciro Spallitta
Alma Mater - University of Bologna
MSc in Applied Economics and Markets
ID: 0001140660
ciro.spallitta@studio.unibo.it

Roberto Vacante
Alma Mater - University of Bologna
MSc in Applied Economics and Markets
ID: 0001131429
roberto.vacante@studio.unibo.it

Abstract

Judicial decision-making in immigration courts exhibits substantial variability influenced by systemic factors and individual discretion. This study applies machine learning to examine judicial behavior in U.S. immigration courts, with a focus on cases in New York City. Leveraging a comprehensive dataset spanning 1997–2024, we develop a methodological framework integrating clustering and supervised learning models to analyze and predict immigration judge decisions. Clustering analysis identifies distinct judicial archetypes—lenient, stringent, and balanced—despite the random assignment of cases, highlighting the influence of judicial discretion. Using Random Forest and Neural Network models, we identify key predictors, such as attorney representation and decision-making duration, which significantly influence case outcomes. Although bias indicators derived from clustering did not enhance predictive accuracy, they provide valuable insights into systemic patterns and judicial tendencies. This research underscores the utility of machine learning not only for predicting judicial decisions but also for diagnosing structural inequities, offering pathways for data-driven reforms in immigration adjudication.

1 Introduction

Every year, thousands of immigrants navigate the complex U.S. immigration court system, often without legal representation or familiarity with immigration laws. This lack of representation profoundly impacts case outcomes, with represented individuals significantly more likely to receive favorable rulings. However, judicial decision-making in these cases exhibits moderate heterogeneity, influenced by the characteristics of

*This work builds on an extension of a broader and ongoing research project in which Roberto Vacante is involved, under the guidance of Nicola Persico (Northwestern-Kellogg), Decio Coviello (HEC Montréal), Lenni Benson (NYLS), and Petra Todd (UPenn) titled "Observers in the Courtroom: Experimental Evidence from Immigration Courts".

Immigration Judges (IJs) and their appointed contexts. These disparities in adjudications, often referred to as *decision roulette*, highlight the need to understand and address the factors driving such variations.

This paper investigates the application of machine learning tools to predict judicial decisions in immigration courts, focusing on proceedings within New York City (NYC). Using a rich dataset from the Executive Office for Immigration Review (EOIR) spanning 1997–2024, and biographical data on IJs from the Transactional Records Access Clearinghouse (TRAC), we employ a semi-supervised machine learning framework to explore patterns in judicial behavior and predict case outcomes. The dataset includes proceeding-specific information such as charges, custody status, and hearing schedules, as well as judge-specific attributes like gender, year of appointment, and the political affiliation of the appointing administration. This comprehensive dataset enables the modeling of case outcomes across three NYC courts—Varick, Broadway, and Federal Plaza—restricted to three types of decisions: removal, grant, and dismissal.

Our methodology combines clustering and supervised learning to address two interrelated goals: characterizing systemic patterns in judicial behavior and predicting case outcomes. Clustering, implemented using k-modes to handle the predominantly categorical nature of the data, groups proceedings based on shared attributes to identify judge-specific behavioral patterns, such as bias in decision-making. We hypothesized that explicit bias indicators derived from clustering could enhance the predictive performance of supervised models, such as Random Forest and Neural Networks. However, our results reveal that the inclusion of these explicit bias indicators did not improve the predictive accuracy of the supervised models. This finding suggests that the supervised models, through their rich feature set, were already implicitly capturing these systemic patterns, demonstrating their robustness in handling complex interactions between case-specific and judge-specific attributes.

Despite this, the clustering analysis provides valuable insights into systemic judicial behavior. It allows for the identification of latent patterns in how IJs handle cases, shedding light on potential sources of bias and the alignment between judges and specific clusters of decisions. For example, the clustering analysis highlights the extent to which certain judges dominate specific clusters, revealing tendencies that may align with perceived biases or systemic inconsistencies in adjudication.

The use of machine learning in the context of judicial decision-making has been explored extensively in recent years. Kleinberg et al. [2018] demonstrate the predictive power of machine learning in human decision-making contexts, such as criminal justice, while emphasizing its potential to mitigate biases. Similarly, Dunn and Sagun [2018] compare the effectiveness of human forecasting and machine learning models, highlighting the advantages of algorithmic approaches in addressing complex decision problems. Chen and Eagly [2020] analyze judicial decision-making under increased immigration caseloads, showing how reforms and external pressures influence outcomes in immigration courts.

In the context of immigration courts, Ramji-Nogales et al. [2007] document *refugee roulette*, highlighting stark disparities in asylum outcomes depending on the assigned judge and the immigrant's country of origin. Hausman [2016] underscore the significant variation in judicial behavior even within the same immigration court, illustrating that some judges are substantially harsher than others in their rulings. He also emphasizes how harsher judges tend to resolve cases early, reducing the opportunity for immigrants to secure legal representation or file for relief. Ryo and Peacock [2021] provide complementary evidence, showing that legal representation and judges' caseloads significantly impact decision outcomes. They also reveal that female judges and those with more experience exhibit distinct decision-making patterns. Haire and Moyer [2015] expand on the broader implications of judicial politics and decision-making patterns, offering a theoretical foundation for understanding systemic heterogeneity in judicial outcomes.

These studies underscore the variability in judicial decisions, often shaped by the characteristics of the adjudicating judge and external factors. Additionally, Zhou [2009] highlight the advantages of semi-supervised

learning frameworks for combining structured and unstructured data, which aligns well with our methodological approach.

Our study builds on this body of work by introducing a semi-supervised framework that explicitly incorporates judge-level bias indicators derived from clustering to assess their contribution to predictive models. While the explicit inclusion of these indicators did not enhance model accuracy, the clustering analysis remains a crucial step in understanding judicial behavior, providing an exploratory lens through which patterns of systemic bias can be identified and analyzed. By identifying heterogeneity in how judges handle cases and quantifying the extent to which these patterns influence outcomes, our findings provide a data-driven perspective on the systemic issues in immigration adjudication.

2 Data

This study relies on two primary sources of data to investigate the decision-making processes of immigration judges in New York City (NYC) courts. The first source, administrative records from the Executive Office for Immigration Review (EOIR), spans the period from April 1997 to August 2024 and provides detailed case-level information, including data on case type, hearing duration, judge decisions, and immigrant characteristics, such as gender and country of origin. The second source includes biographical information about immigration judges, obtained from the TRAC (Transactional Records Access Clearinghouse) database. These details encompass the judges' education, gender, year of appointment, and the broader political context of their appointment. The data from these two sources were processed independently and underwent cleaning and preprocessing specific to each dataset. Following this step, the data were merged using judges as the linking key, resulting in a unified dataset that includes case-specific information and immigrant attributes from the EOIR data alongside judge-specific characteristics from the TRAC data. This integration enables a comprehensive analysis of how case-specific factors, immigrant characteristics, and judge-level attributes interact to shape judicial decision-making in immigration courts.

2.1 Data Description

The dataset consists of 14 variables, which can be categorized into three main groups: case-specific, judge-specific, and immigrant-specific variables, all contributing to the prediction of the proceeding outcome. Case-specific variables include the decision duration, measured as the time difference between the start and end of the case, the number of charges filed against the applicant, the immigration court jurisdiction where the case was processed, the number of applications for relief filed, the number of scheduled hearings, and the applicant's custody status at the time of the case. These variables provide a procedural overview of each proceeding, capturing the complexity of the case and the administrative factors influencing its trajectory.

Judge-specific variables include the gender of the judge and the year of their appointment. Additionally, a noteworthy feature of this dataset is the inclusion of a variable identifying the political party associated with the appointing administration of each judge. Immigration Judges in the United States are appointed by the Attorney General, who is selected by the President. Therefore, the political party of the President at the time of a judge's appointment serves as a proxy for the broader policy priorities of the administration. This variable reflects the assumption that the Attorney General is likely to select judges whose views align with the administration's policy goals. While this variable provides useful contextual insights into the possible influence of political dynamics, it is not treated as a key feature in this study. Instead, the primary focus remains on judicial decisions and their associations with other case-specific and judge-level attributes.

Immigrant-specific variables include whether the case involved criminal charges, the number of appeals filed to the Board of Immigration Appeals (BIA), whether the applicant was represented by an attorney, and the native region or continent of the applicant. These variables encapsulate the aspects of the applicant's background, such as legal representation and regional diversity, which may influence the proceeding's outcome.

To focus on judicial outcomes where discretion is most apparent, the analysis restricts its scope to three specific decision types: grant, dismissed, and removal. Table 1 provides a detailed description of these decision types. Grant decisions indicate that relief has been awarded to the applicant, allowing them to remain in the United States under certain protections or conditions. Dismissed cases involve the judge closing the case due to procedural reasons or lack of sufficient evidence for removal. Removal decisions represent a formal order for the applicant to leave the United States, typically following a denial of relief or defense against deportation. By concentrating on these three outcomes, the study emphasizes decision types where judicial discretion plays a critical role, as these are the most interpretable outcomes for understanding judicial behavior. Additionally,

Decision Code	**Description**
Relief Granted	Relief is granted to the applicant, allowing them to remain in the country under certain conditions or protections.
Dismissed by IJ	The case is dismissed by the Immigration Judge, often due to procedural reasons or lack of sufficient evidence for removal.
Removal	A removal order is issued, requiring the applicant to leave the country, typically after unsuccessful relief or defense against deportation.

Table 1: Description of Immigration Case Decision Types

immigration case decisions reveal notable regional disparities, as summarized in Table 2. For instance, Europe and Asia exhibit higher proportions of grant decisions, suggesting more lenient legal frameworks or procedural benefits in those regions. Conversely, the Americas and Middle-East regions are characterized by a predominance of removal decisions, reflecting stricter enforcement practices or significant barriers to securing favorable rulings. The regional distribution of outcomes highlights the influence of geographical and procedural contexts on immigration proceedings.

	Grant	Dismissal	Removal
Africa	0.492	0.045	0.464
Americas	0.167	0.192	0.641
Asia	0.497	0.075	0.427
Europe	0.524	0.081	0.395
Middle East	0.331	0.076	0.594
Oceania	0.216	0.146	0.637
Others	0.421	0.058	0.521

Table 2: Regional distribution of outcomes

2.2 Data Preprocessing

Before proceeding with the analysis, a preliminary exploration of missing values was conducted to ensure the integrity of the data. The proportion of missing data is minimal, with only 0.5% of values missing overall. Most variables exhibit complete data or a negligible proportion of missingness, except the variables ij_gender (1%) and party (3%), which correspond to immigration judges' characteristics. This minimal level of missingness, coupled with its distribution across variables, suggests that the missing data are consistent with the assumption of missing at random (MAR). This assumption justifies the use of imputation techniques to handle missing entries.

To address missing values while preserving the statistical validity of the dataset, we implemented Multivariate Imputation by Chained Equations (MICE). The method operates iteratively, imputing missing values for one variable at a time while accounting for the uncertainty of predictions in subsequent imputations. By iterating through this process until convergence, MICE produces multiple imputed datasets that reflect the variability in the imputation process. The resulting imputations preserve the joint distribution of the data and ensure consistency across variables, minimizing the risk of introducing systematic errors.

Additional preprocessing steps were conducted to prepare the dataset for analysis. The dataset was filtered to include only proceedings classified as dismissed, grant, or removal, which are the primary decision outcomes of interest. Categorical variables were encoded for compatibility with the clustering and supervised learning models, while continuous variables, such as proceeding duration, were normalized to ensure comparability across features. Judge-level attributes, including gender and political affiliation, were integrated into the main dataset to allow for an analysis of decision-making patterns and potential biases.

3 Methodology

The methodological framework of this study integrates clustering and supervised machine learning techniques in a multi-step approach to analyze the decision-making processes of immigration judges. This approach draws on prior work exploring judicial heterogeneity, clustering methods, and supervised machine learning frameworks [Aggarwal and Reddy, 2013, Chapelle et al., 2010, Aletras et al., 2016].

The primary objective of the first step is to identify potential patterns and biases in judicial behavior through an unsupervised clustering analysis. Clustering, particularly suited for categorical data [Aggarwal and Reddy, 2013], allows us to group proceedings based on shared characteristics. This analysis reveals the distribution of judges across clusters, providing insights into whether specific clusters are dominated by individual judges. By identifying judge-specific dominance in clusters, we quantify potential bias levels, aligning with previous efforts to measure judicial disparities and tendencies [Epstein et al., 2013, Berdejó and Chen, 2017]. These bias levels are then incorporated as an additional feature for subsequent supervised learning analysis.

The second step involves two supervised learning tasks designed to predict the outcome of the proceedings (i.e., the decision code). Initially, the supervised models are trained and evaluated without incorporating the bias levels identified in the clustering analysis. This step establishes a baseline for predictive performance and ensures that the supervised models are robust in capturing patterns directly from the case-, judge-, and immigrant-specific variables [Aletras et al., 2016]. In the final step, the bias levels derived from the clustering analysis are introduced as an additional feature in the supervised models. This stage assesses whether the incorporation of bias information enhances the model's predictive capabilities.

Crucially, this integration reflects a key principle of multi-step methodologies, wherein the uncertainty of the initial unsupervised stage propagates into the subsequent supervised stage [Chapelle et al., 2010]. By accounting for this uncertainty, the framework provides a richer understanding of how judicial behavior might influence case outcomes. While this approach shares similarities with common two-stage methodologies, it deviates in structure, as the initial supervised learning step is performed independently of the clustering-derived bias information. The final step serves to bridge the unsupervised and supervised components, explicitly testing the added value of the bias feature in improving predictions.

3.1 Clustering Analysis

To analyze the dataset, which predominantly contains categorical variables, we employed the k-modes clustering algorithm. This method is particularly suited to datasets with categorical data, as it minimizes dissimilarity by counting mismatches between data points and centroids, effectively grouping observations with similar characteristics. The optimal number of clusters was determined using the Elbow Method, which evaluates clustering quality across a range of cluster numbers (k). In this approach, the k-modes algorithm was applied iteratively with values of k from 1 to 10, and the within-cluster cost was plotted for each k. The

optimal number of clusters is identified at the *elbow point*, where the marginal gain in clustering quality begins to diminish. Based on this method, the optimal number of clusters was found to be three.

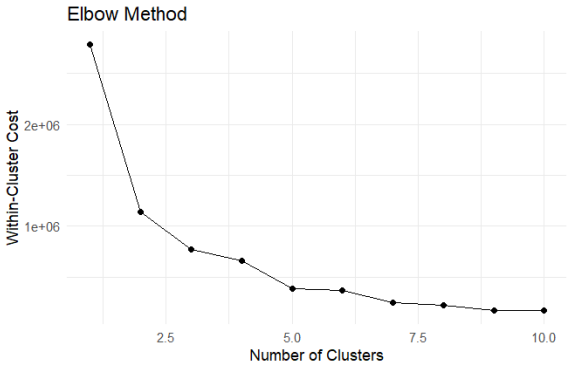

Figure 1

Prior to clustering, the dataset was pre-processed to ensure compatibility with the k-modes algorithm. Specifically, all variables were converted into factors to treat them as categorical, ensuring the algorithm's correct handling of data types. This pre-processing step was essential to avoid potential issues arising from mixed data types. The resulting clusters provided a structured framework for exploring patterns and relationships in the data, serving as the foundation for subsequent analyses.

Once the clusters were generated, a cluster identifier was assigned to each observation and appended as a new column in the original dataset. This integration facilitated the analysis of the distribution of judges across clusters and the investigation of their decision patterns within each cluster. Specifically, the analysis focused on two aspects: (1) how judges were distributed across clusters, and (2) whether the predominance of cases associated with a single judge in a specific cluster could indicate potential bias or recurring behavior.

Identification of Judges' Bias

A key focus of the analysis was to assess potential biases in judicial decision-making. The identification of bias in this context is meaningful only under the assumption that cases are assigned randomly to immigration judges. The Executive Office for Immigration Review (EOIR) has confirmed that immigration court cases follow a random allocation process. This confirmation provides an important foundation for the validity of our analysis, allowing us to attribute patterns in judicial behavior to individual decision-making rather than external case assignment factors.

In order to quantify potential bias, we defined different "bias levels" based on the proportion of a judge's cases concentrated within a single cluster. Judges with fewer than three cases were excluded from the analysis to ensure sufficient data for evaluating recurring behavior. Three bias levels were established, with thresholds set at 60%, 75%, and 90% of cases within a single cluster. Judges with over 90% of their cases in one cluster were assigned the highest bias level (3), while those with 75–90% and 60–75% were assigned bias levels of 2 and 1, respectively. This framework aimed to quantify the extent to which judges exhibited consistent decision patterns within specific clusters. These thresholds were selected to balance sensitivity in detecting bias with the need to exclude noise from minor variations in case distribution. However, alternative thresholds could be explored to refine the measure further, particularly in cases where cluster assignments may reflect external factors rather than inherent judicial tendencies.

The calculated bias levels were subsequently incorporated into the dataset, enabling further analyses to evaluate whether a judge's bias influenced their decision-making processes. In particular, the random forest model was re-run, incorporating the bias level as an additional feature to assess its predictive power in improving case outcome predictions. This second step allowed us to account for whether bias levels have predictive power on immigration court decisions.

3.2 Supervised Learning Models

Random Forest

In our study, we developed and optimized a Random Forest classifier to predict the decision variable using all available predictors. To determine the optimal number of trees, we evaluated the out-of-bag (OOB) error for tree counts ranging from 25 to 200 in increments of 25. This OOB error provides an unbiased estimate of model accuracy, as it is calculated using data excluded from the bootstrap sample for each tree, effectively simulating an independent test set. As shown in Figure 2, increasing the number of trees beyond 50 resulted in minimal improvement in the OOB error, indicating that 50 trees were sufficient to balance performance and computational efficiency.

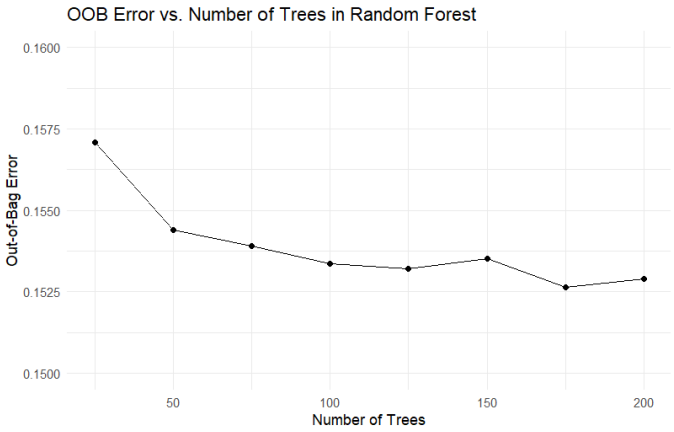

Figure 2

Our final model was configured with 50 trees, a choice supported by the observation that the most significant reduction in OOB error occurred up to this threshold, with diminishing returns thereafter. This setup reflects the principle of parsimony, favoring simplicity while maintaining optimal performance. Furthermore, the model included variable importance measures to evaluate the predictive contribution of each feature, enhancing its interpretability. Validation through a confusion matrix on the test data confirmed the model's robustness and accuracy in classifying decision codes. This thorough tuning and validation process underscored the model's ability to generalize effectively to new data while avoiding overfitting. Secondly, the random forest model was re-run, incorporating the bias level assessed in the clustering analysis, as an additional feature to measure its predictive power in improving case outcome predictions. This second step allowed us to account for whether bias levels have predictive power on immigration court decisions.

Neural Network

The neural network methodology was designed to evaluate the impact of activation functions, regularization techniques, and hyperparameter configurations to achieve optimal performance. Activation functions, including ReLU, sigmoid, and tanh, were tested to determine their influence on accuracy over 15 epochs. As shown in Figure 3, both tanh and ReLU performed comparably; however, tanh demonstrated a slight advantage due to its ability to center outputs within the range $[-1, 1]$, facilitating efficient gradient propagation. Based on these findings, the tanh activation function was chosen for the final model.

The role of batch normalization was examined within networks employing the tanh activation function. Figure 4 shows that the network performed better without batch normalization. This behavior is attributed to the implicit normalization properties of the tanh function, which mitigate internal covariate shifts, reducing the benefits of batch normalization. As a result, batch normalization was excluded from the final architecture.

The selection of the learning rate was critical for balancing training efficiency and stability. Learning rates of 0.1, 0.01, and 0.001 were evaluated, as shown in Figure 5. Higher rates, such as 0.1, resulted in unstable

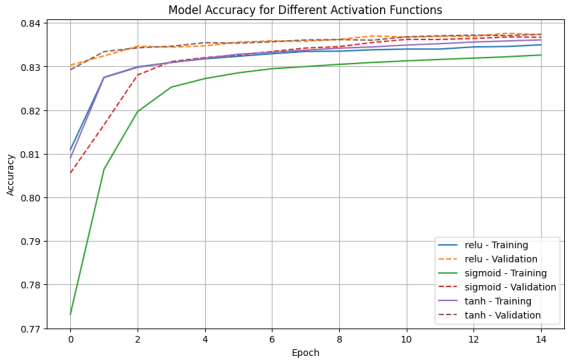

Figure 3: Comparison of activation functions.

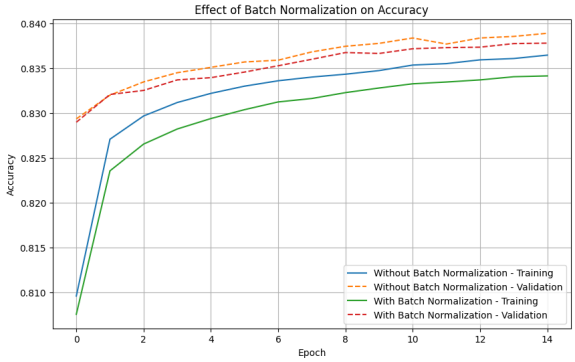

Figure 4: Effect of batch normalization.

training and failed to converge, while 0.01 achieved faster convergence but showed instability in later epochs. A learning rate of 0.001 ensured slower but stable convergence and was selected to maintain consistency during training.

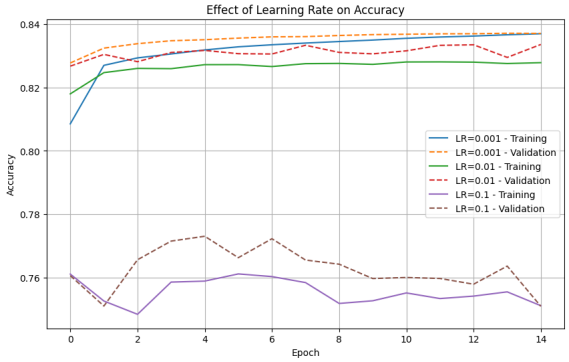

Figure 5: Comparison of learning rates.

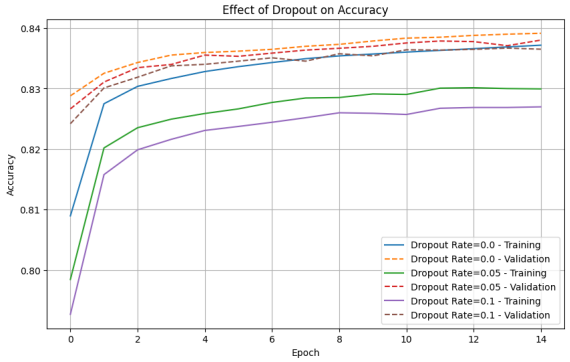

Figure 6: Effect of dropout regularization.

Dropout regularization was tested to analyze its impact on performance. As depicted in Figure 6, introducing dropout at rates of 0.05 and 0.1 led to a decline in accuracy, which can be attributed to the simplicity of the network and the normalization properties of the tanh function. Dropout was excluded from the final architecture to preserve the model's ability to capture patterns effectively.

Training and validation metrics were monitored to ensure generalization. As shown in Figure 7, the training accuracy reached 0.8376, closely aligning with the validation accuracy of 0.8388. This strong alignment indicates robust generalization with minimal overfitting. The steady decline of both training and validation loss further supports the stability of the training process.

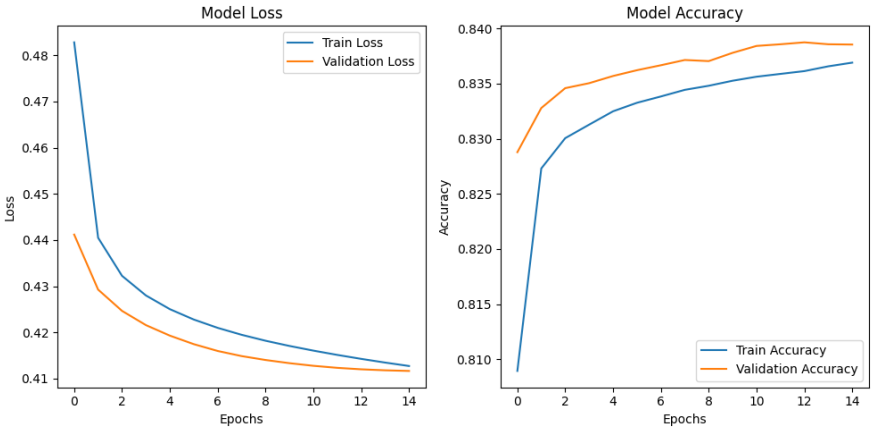

Figure 7: Training and validation loss and accuracy dynamics.

Feature importance was analyzed using the first-layer weights of the network to identify the most influential variables. Additionally, the network's ability to classify outcomes was assessed using ROC-AUC metrics for each decision class, as detailed in the next section.

4 Results

4.1 Findings from Clustering Analysis

The clustering analysis revealed three distinct groups characterized by varying compositions of cases, judges, and decision patterns. The graphs below summarize the primary attributes associated with each cluster:

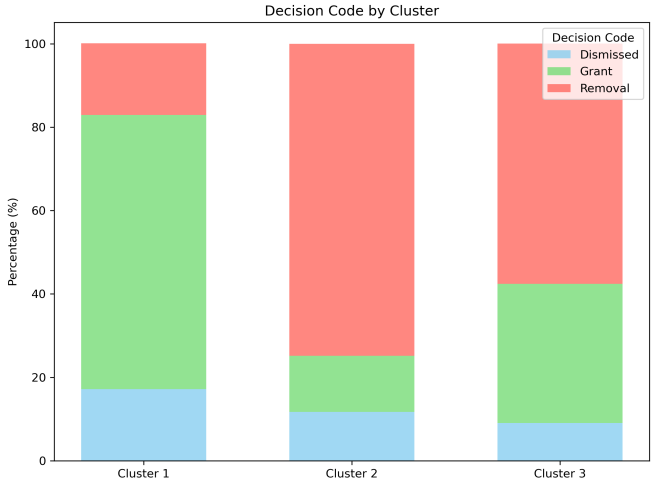

Figure 8

→ *The Lenient Deliberators:* Cluster 1 includes cases with a higher likelihood of **grant** decisions, reflecting a more lenient adjudication process. Judges in this cluster are predominantly female and Republican-appointed, which may suggest demographic or political influences, although this remains speculative. The relatively long decision durations and low appeal rates indicate that the cases handled in this cluster are likely complex, requiring detailed examination and thorough deliberation. The high level of confidence in the outcomes may further underscore the deliberative nature of the decisions in this cluster.

→ *The Strict Judges:* Cluster 2 is largely associated with **removal** decisions, indicating a stricter and more resolute approach to adjudication. The gender distribution remains predominantly female, with most judges Democrat-appointed. The shorter decision durations suggest that cases in this cluster are less complex or involve more straightforward legal issues. Additionally, the relatively low number of appeals highlights the decisive nature of the judgments, with outcomes that are generally less likely to be contested.

→ *The Complex Case Handlers:* Cluster 3 demonstrates a mix of **removal** and **grant** outcomes, reflecting a more nuanced and varied adjudication process. This cluster is predominantly male and features judges with the highest average years on the bench, suggesting that more seasoned and experienced adjudicators are assigned to these cases. The intermediate decision durations and the highest appeal rates indicate that cases in this cluster are the most contentious or legally complex, requiring deeper deliberation and often leading to challenges. This complexity highlights the expertise needed to navigate these multifaceted cases.

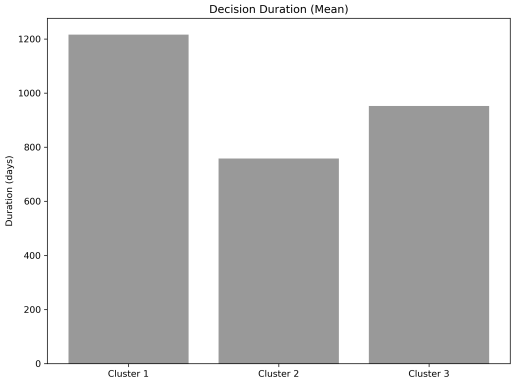

Figure 9

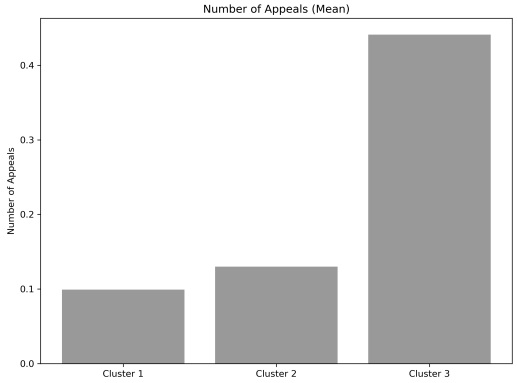

Figure 10

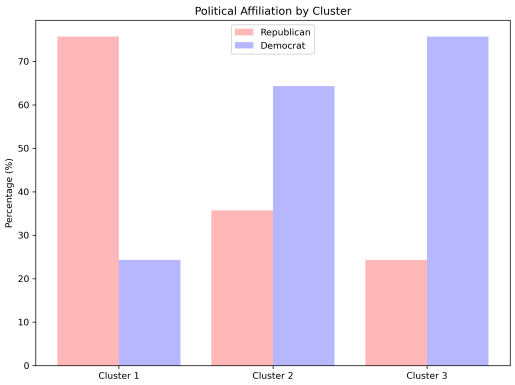

Figure 11

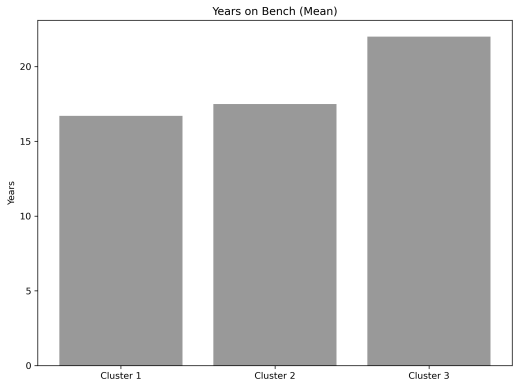

Figure 12

Findings from Judges' Distribution and Bias

The analysis of bias levels, which reflects the proportion of cases handled within a single cluster, reveals a mixed distribution. Most cases in Cluster 1 (56.6%), Cluster 2 (74.2%), and Cluster 3 (41.9%) involve judges with no significant bias (Bias Level 0). However, as seen in Figure 13, a notable portion of cases (particularly in Cluster 1 and 3) are associated with higher bias levels (2 or 3). These patterns warrant further investigation to determine whether they stem from structural factors or inherent judicial preferences. The clusters provide a useful framework for exploring patterns in judicial decisions, though the findings should be interpreted cautiously. The presence of judges across multiple clusters and varying bias levels underscores

the complexity of judicial practices and the need for further study to assess how structural, demographic, or political factors influence case outcomes. These insights can inform future research aimed at understanding the dynamics of judicial decision-making.

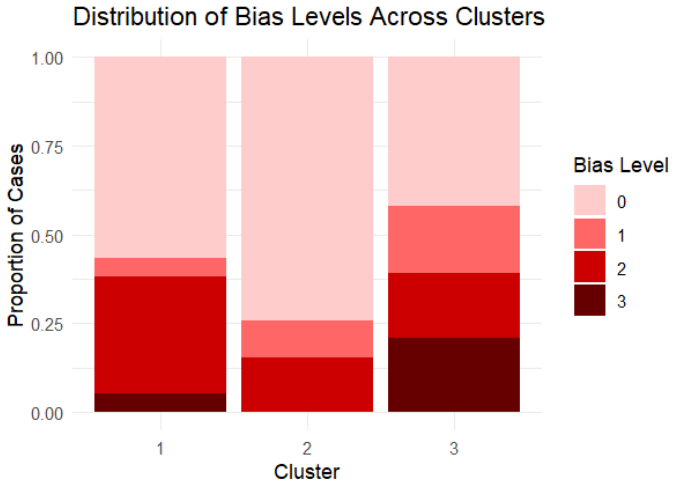

Figure 13

4.2 Outcomes of Supervised Learning Models

The application of the Random Forest model demonstrated substantial predictive capabilities, achieving an overall accuracy of 84.81%. This underscores its robust performance in classifying the three categories of outcomes. Our parallel analysis using Neural Networks yielded comparable results in both accuracy and the importance of predictive variables. However, to maintain focus and avoid redundancy, our comments will mainly focus on the Random Forest results due to its faster computational performance and clearer interpretability.

Findings from Random Forest Model

The model proved particularly adept at identifying Grant decisions with a sensitivity of 90.33%, followed by Removal at 84.90%, and Dismissed at 68.41%, as shown below. The lower sensitivity for Dismissed cases

	Prediction outcome			
	Dismissed	**Grant**	**Removal**	**Total**
Dismissed	**7617**	1559	1232	10408
Grant	2168	**29165**	5618	36951
Removal	1349	1562	**38507**	41418
Total	11134	31286	44357	

Table 3: Confusion Matrix

could be attributed to unobserved factors that influence case dismissals but are not captured by the variables in our dataset. These unobserved elements, such as procedural technicalities or external contextual factors, may introduce complexities that the model cannot fully account for. Nevertheless, the model's specificity for Dismissed cases stood out, reaching 96.41%, suggesting that while case dismissals are influenced by factors beyond the dataset, the model remains highly effective in identifying and categorizing other case outcomes. Additionally, high Positive Predictive Values (PPV) and Negative Predictive Values (NPV) across all classes further reinforce the precision and reliability of this model. The analysis of variable importance (Figure 14) revealed critical insights into the factors influencing judicial decisions. The most significant predictors were

the number of appeals and the number of applications, followed closely by decision duration and attorney presence. These variables showed a substantial mean decrease in Gini impurity, highlighting their pivotal role in shaping case outcomes. While the predictive power of the *region_ grouped* variable is comparatively lower, its level of importance suggests it may capture underlying patterns consistent with the regional disparities discussed in Section 2.1.

Expanded Random Forest Model

When the Random Forest model was expanded to include the Bias Level variable, the accuracy saw only a marginal improvement to 85.29%.

	Prediction outcome			
	Dismissed	**Grant**	**Removal**	**Total**
Dismissed	**7757**	1641	1264	10662
Grant	2079	**29078**	5208	36365
Removal	1298	1567	**38885**	41750
Total	11134	31286	44357	

Table 4: Confusion Matrix of the extended model

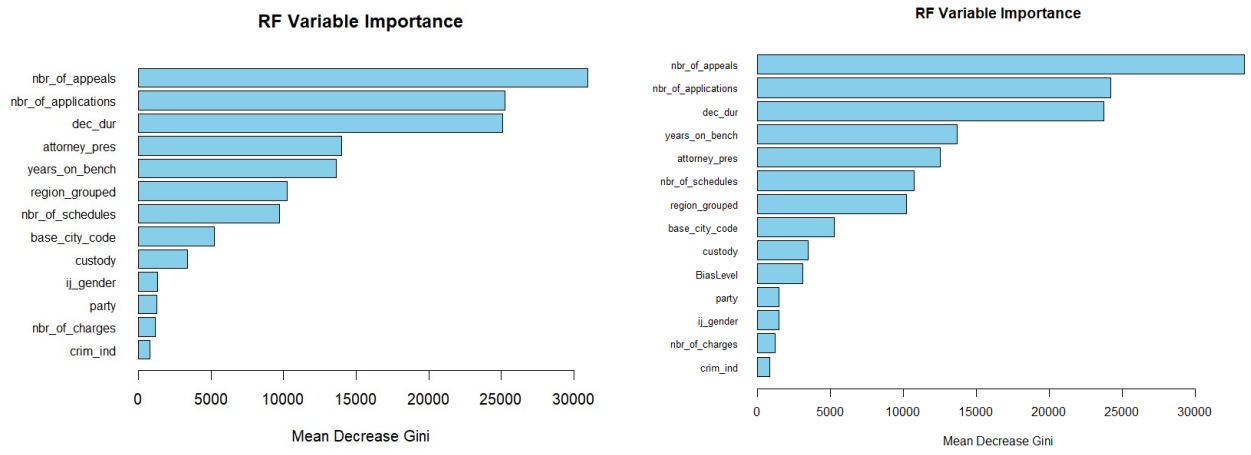

Figure 14: Variable Importance compared after the inclusion of the Bias Level variable

Its inclusion in the analysis did not significantly affect the model's predictive accuracy, and its variable importance remained relatively low compared to the primary predictors such as case-specific, judge-specific, and immigrant-specific features. This outcome aligns with our earlier expectation that supervised models, such as Random Forest, are inherently capable of capturing complex interactions and latent patterns directly from the feature set, including those related to judicial bias. Consequently, the explicit inclusion of the Bias Level as a feature did not meaningfully improve the model's performance metrics or enhance its predictive capabilities.

The importance metrics from the Random Forest model further highlight that while Bias Level may signal certain patterns of judicial behavior, its overall impact on outcome prediction is limited. This suggests that observed judicial bias, as captured by the clustering-derived Bias Level, does not fundamentally alter the predictive power of the model or significantly influence decision-making processes. Instead, the supervised model effectively identifies and utilizes other more influential variables, such as the nature of the case, the judge's characteristics, and contextual factors, to achieve high predictive accuracy.

Overall, the Random Forest model provides a robust framework for understanding and predicting judicial decision-making. It achieves high accuracy while offering meaningful insights into the factors influencing outcomes. The additional analysis incorporating Bias Level confirms the model's reliability and highlights that observed patterns in judicial behavior merit further exploration. However, these patterns do not dominate the predictive dynamics nor significantly alter the model's performance. This result reflects the model's ability to utilize its feature set comprehensively and suggests that while Bias Level offers interpretative value in understanding systemic behavior, its role in enhancing outcome prediction is limited.

Findings from Neural Network

Feature importance analysis, depicted in Figure 15, highlights that case-specific variables, such as the number of appeals, number of applications, and decision duration, were the most critical predictors. These variables consistently exhibited high importance, reflecting their dominant role in determining outcomes. Conversely, judge-specific attributes, such as gender and political affiliation, showed lower importance, indicating a limited influence on case decisions. These results confirm that case-specific details are the primary drivers of judicial decision-making.

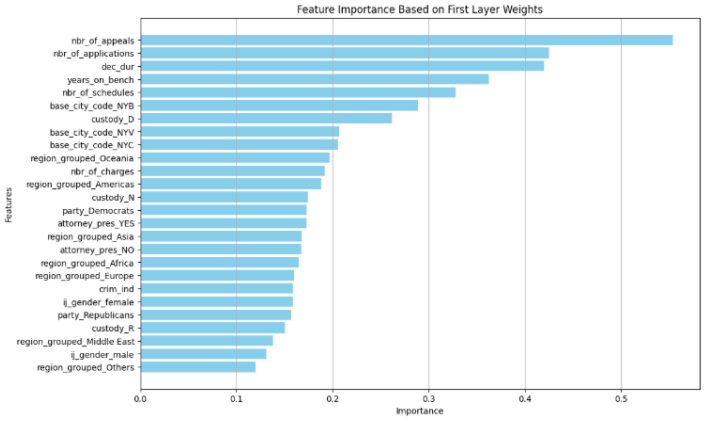

Figure 15: Feature importance based on first-layer weights.

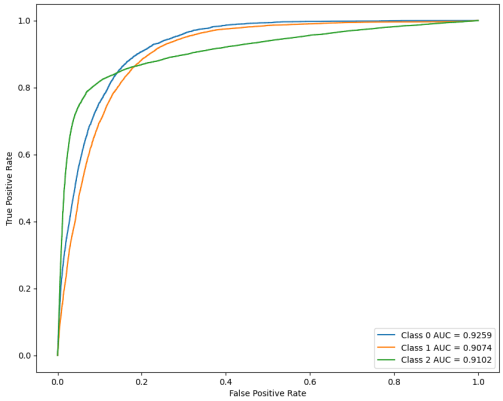

Figure 16: ROC-AUC curves for the three decision classes.

The network's classification performance was further evaluated using ROC-AUC metrics, as shown in Figure 16. The model achieved AUC scores of 0.926, 0.907, and 0.910 for the three decision classes, underscoring its strong discriminative capability across all outcomes. These scores indicate a high degree of accuracy in distinguishing between decision types, further validating the robustness of the model. In summary, the neural network achieved a training accuracy of 0.8376 and a validation accuracy of 0.8388, demonstrating its reliability and ability to generalize effectively. The exclusion of batch normalization and dropout, combined with the use of tanh activation, contributed to this robust performance. The high importance of case-specific variables and the strong ROC-AUC results further underscore the model's capacity to accurately capture patterns in judicial outcomes and provide reliable predictions.

5 Conclusions and Policy Implications

This study contributes to understanding judicial decision-making by analyzing systemic factors and individual discretion through machine learning and clustering methodologies. It reveals patterns in the adjudication of immigration cases that not only highlight variability among judges but also suggest avenues for reform. Our findings resonate with the insights of Kleinberg et al. [2018], who explored the integration of machine learning into judicial decisions to improve fairness and efficiency. Both studies underscore the potential of predictive tools to diagnose and address disparities in judicial outcomes.

A salient finding from this study is the clustering of Immigration Judges into lenient, stringent, and balanced archetypes, even though cases are ostensibly randomized upstream by the Department of Homeland Security (DHS). The emergence of these clusters with distinct outcome distributions implies that judges exercise significant discretion in their rulings. This discretionary role aligns with findings in the bail decision context, where random assignment of cases still yielded variability in judicial leniency, as documented by Kleinberg et al. [2018]. Such patterns highlight a shared systemic challenge: while randomization minimizes bias at the case-assignment level, variability in individual decision-making perpetuates disparities.

The supervised models further illuminated key predictors of case outcomes, focusing on case-specific features. Among these, the presence of an attorney was particularly impactful, strongly correlating with favorable outcomes for defendants. This finding parallels extensive literature emphasizing the critical role of legal representation in achieving equitable judicial processes. For example, Abrams and Rohlfs [2011] documented the influence of counsel in improving case outcomes across criminal justice settings. Expanding access to affordable or publicly funded legal assistance emerges as a clear policy priority to mitigate structural inequities.

Among other significant variables, the number of appeals was the most influential, reflecting their role in shaping case trajectories. Cases with multiple appeals often involve contentious issues, highlighting the importance of appellate oversight as a corrective mechanism in the judicial system. Similarly, the number of applications for relief and decision duration emerged as substantial predictors. Longer decision durations may indicate increased case complexity, suggesting that certain procedural factors influence both outcomes and perceptions of fairness. These variables point to the need for streamlined administrative processes that ensure efficiency without compromising deliberative depth.

Lastly, the interpretive value of clustering underscores its utility beyond predictive accuracy. The behavioral archetypes identified among judges provide insights for training and oversight. Tailored interventions could address specific tendencies, such as encouraging lenient judges to apply greater procedural rigor and stringent judges to emphasize proportionality in their rulings.

In conclusion, the shared insights between this study and related works underscore the transformative potential of integrating machine learning into judicial contexts. These tools not only enhance predictive capabilities but also reveal latent disparities and areas for systemic improvement. By addressing the identified gaps—through expanded access to representation, revised detention practices, and tailored judicial training—policymakers can foster a more equitable and transparent system. This study affirms the broader applicability of machine learning to diagnose and reform complex institutional systems.

References

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

Theodore Dunn and David Sagun. Human versus machine: A comparison of human forecasting and machine learning models. *Artificial Intelligence and Law*, 26(3):323–344, 2018.

Daniel L Chen and Ingrid V Eagly. Judging in a crises: Court reform and judicial decision making in times of immigration caseloads. *Journal of Empirical Legal Studies*, 17(4):810–855, 2020.

Jaya Ramji-Nogales, Andrew I. Schoenholtz, and Philip G. Schrag. Refugee roulette: Disparities in asylum adjudication. *Stanford Law Review*, 60:295–411, 2007.

Catherine Hausman. What drives variation in immigration decisions? evidence from five million immigration court cases. *American Economic Journal: Applied Economics*, 8(4):1–34, 2016.

Emily Ryo and Ian Peacock. The impact of legal representation and judge characteristics on immigration decisions. *Law & Society Review*, 55(1):102–131, 2021.

Susan B Haire and Laura P Moyer. *Judicial Politics: Readings from Judicature*. CQ Press, 2015.

Zhi-Hua Zhou. *Semi-Supervised Learning*. MIT Press, 2009.

Charu C Aggarwal and Chandan K Reddy. *Data Clustering: Algorithms and Applications*. CRC Press, 2013.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2010.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93, 2016.

Lee Epstein, William M Landes, and Richard A Posner. *The Behavior of Federal Judges: A Theoretical and Empirical Study of Rational Choice*. Harvard University Press, 2013.

Carlos Berdejó and Noelle Y Chen. Disparities in sentencing: Evidence from california's three strikes law. *The Journal of Empirical Legal Studies*, 14(2):327–368, 2017.

David S. Abrams and Chris Rohlfs. Optimal bail and the value of freedom: Evidence from the philadelphia bail experiment. *Economic Inquiry*, 49(3):750–770, 2011. doi:10.1111/j.1465-7295.2010.00339.x.

Appendix

A. Data Description

Variable Name	Definition and Source
Dataset Overview	Observations: 443,888; Variables: 14.
dec_code	The decision of the immigration judge for a case depending on the case type.
dec_dur	Duration to take the decision, measured as the difference in time between `cosc_date` and `ccomp_date`.
nbr_of_charges	Number of charges for a case.
base_city_code	Immigration court having jurisdiction over the assigned hearing location.
nbr_of_applics	The number of applications for relief filed by aliens. Abbreviated from *number of applications*.
nbr_of_schedus	Number of hearings scheduled for each proceeding. Abbreviated from *number of scheduled hearings*.
custody	Current custody status of the alien.
crim_ind	Indicates whether a charge is a criminal charge.
nbr_of_appeals	Number of appeals to BIA for each proceeding filed by aliens.
attorney_pres	Indicates whether the asylum seeker has an attorney.
ij_gender	Gender of the immigration judge.
yr_app	Year the judge was appointed.
party	Political affiliation of the appointers.
region_grouped	Native continent of immigrant. This represents a generalized grouping based on geographic origin.