# ARGOS: A Human-in-the-Loop ML System for Legal Aid Classification in Espírito Santo

Eduardo Pacheco[a]

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 19 May 2025

**Abstract.** We present the modeling of an AI system named ARGOS to classify free legal aid requests in the Court of Justice of Espírito Santo (TJES). The model replicates the decision-making pattern of a judge from the 6th Civil Court of Vila Velha, developed within Challenge 16 of PitchGovES by Atlas.IA, promoted by the State Management Laboratory of Espírito Santo (Lab.ges). The challenge proposed an automated solution to assess applicants' eligibility by cross-referencing multiple data sources to support judicial decisions. The approach consists of three key components: (i) a machine learning model for classifying requests, (ii) an information retrieval engine to support the model, and (iii) an explainability mechanism to justify decisions. A global model for TJES was initially considered but discarded due to legal and technical constraints. Instead, a judge-specific model was developed to align with the principle of Judicial Independence in Decision-Making. Model development began with analysing past judicial decisions, classifying requests as granted or denied. A significant challenge was data imbalance, as requests were not evenly distributed across classes. To address this, balancing techniques and feature selection methods were applied. Various machine learning algorithms, including decision trees and deep learning models, were tested. The final model balanced accuracy and interpretability, ensuring magistrates could understand the factors influencing decisions. The second component involved creating an information retrieval engine to supplement applicant data. Given that free legal aid assessment relies on financial capacity, the tool gathered external data. However, access to government databases was restricted, so alternative sources were used, including public income records, social assistance histories, and judicial metadata. The solution comprises four modules: applicant data capture, a dashboard, a decision interface with model training, and a recommendation system. The initial implementation achieved 81.3% balanced accuracy, surpassing the 70% target, demonstrating potential for broader judicial adoption.

**Keywords.** Machine Learning, Legal Decisions
**Poster, DOI:** https://doi.org/10.59490/dgo.2025.930

## 1. Introduction

The right to free legal aid is fundamental for ensuring equal access to justice, particularly for economically disadvantaged individuals. In Brazil, courts face a growing demand for legal aid requests, requiring judges to evaluate the financial conditions of applicants while maintaining efficiency and fairness. TJES is no exception, and optimizing the analysis of these requests has become a pressing issue.

To address this challenge, the PitchGovES initiative, led by the Lab.ges, launched Challenge 16, which called for innovative solutions to improve the evaluation of free legal aid requests. The core question posed was: How can access to justice be facilitated through a model that assesses whether a request should be granted, cross-referencing multiple data sources to justify its decision? The winning proposal, developed by ATLAS.IA, aimed to build an AI-powered system to support judicial decision-making in this context.

Our project ARGOS, selected as the winner of the challenge, proposed the development of an AI-powered system with Human-in-the-Loop approach capable of classifying legal aid petitions as “granted” or “denied” to facilitate decision-making in current cases based on historical judicial behavior. The system integrates a

[a] Data Scientist at Atlas.IA (www.atlasia.tech) , São Paulo, Brazil. edu@atlasia.tech. ORCID: https://orcid.org/0009-0007-9887-9189

machine learning model trained on judicial decisions, an information retrieval engine to augment petitioner data, and a layer of explainability to ensure transparency and accountability. Rather than proposing a global model for the entire judiciary, we chose to replicate the decision-making logic of an individual judge from the 6th Civil Court of Vila Velha based on the rationale that each judge requires a personalized model. This design respects judicial independence, keep the judge at the center of the decision-making process aligning with legal and institutional principles while delivering practical benefits. The proposed system adheres to the guidelines for Trustworthy AI developed by the European Commission (2019), which emphasize human agency, transparency, and accountability. In line with these principles, our approach embeds a Human-in-the-Loop model that ensures judicial autonomy is maintained, while enhancing the quality and consistency of decision-making. As Bekkerman (2021) argues, HITL approaches are essential when deploying ML systems in high-stakes domains such as healthcare, finance, or law, enabling oversight and fostering trust in automated tools.

### *1.1 Project Scope and Objectives*

We propose automating judicial decisions by replicating the individual decision patterns of each judge. This can increase efficiency in judicial services and allow judges to dedicate more time to analyzing the merits of each case. This can be addressed by a machine learning model that learns the judge's decision-making patterns and optimize public resource allocation, generating benefits for society as a whole. It is possible to avoid wasting scarce resources—such as judicial time—on tasks that could be automated. It is also possible to enhance the data available by data enrichment from external sources that either corroborate or contradict the petitioner's claims, enabling more informed decision-making. The proposed system was structured around two main components: 1. **machine learning model** trained to classify legal aid requests as granted or denied, based on historical judicial decisions. 2. An **information retrieval engine** that collects external data to supplement the analysis. This system must also enhance **explainability** and enforce transparency to help users to justify the decisions.

A key challenge in designing the system was determining whether a global model for the entire court could be developed. However, legal and technical constraints led to the decision to focus on modeling the decision-making patterns of a single judge. This approach aligns with the principle of Judicial Independence in Decision-Making, ensuring that each magistrate retains autonomy in their rulings. Judges typically rely on the information provided in the initial petition to evaluate a legal aid request. However, applicants may omit or misrepresent financial details, leading to potential inaccuracies in judicial decisions. The project aimed to expand the data sources available to judges, supplementing the petition with external financial and social data. An information retrieval engine was developed to collect relevant data, including: 1. Public salary records of federal and state executive branch employees 2. Social assistance histories, such as Bolsa Familia and Emergency Aid 3. Metadata from previous legal proceedings, including records from the Public Defender's Office. Despite efforts to access government databases, legal and technical barriers restricted direct integration with certain financial records, such as tax returns and property registries. As a result, alternative web-crawling techniques and APIs were utilized to gather publicly available data. The developed solution, ARGOS, integrates multiple components into a comprehensive decision-support system. The platform consists of four key modules: 1. Data Capture – Extracts structured information from the initial petition and external sources. 2. Dashboard – Consolidates relevant data into a user-friendly interface for judicial review. 3. Decision Interface – Enables judges to classify requests and fine-tune the AI model. 4. Recommendation System – Provides AI-generated predictions to streamline decision-making.

## 2 Data

At the final stage, the data from all collected sources was combined and processed to represent a legal aid request, stored in the "request" table with the following fields:

| Field | Description |
|---|---|
| Total income | Sum of the applicant's declared income and external sources |
| Has income | Boolean indicating whether the applicant has a registered income |
| Has public defender | Indicates whether the applicant has been assisted by a public defender |
| Has social vulnerability | Self explanatory |
| Has debt | Self explanatory |
| Total Debt | Total amount of debt recorded for the applicant |
| Has expenses | Boolean flag indicating if the applicant has reported expenses |
| Has assets | Boolean flag showing if the applicant has registered assets |
| Total assets | Sum of the applicant's registered assets and external sources |

| | |
|---|---|
| Occupation | The applicant's employment status or job category |
| Decision | Target variable, indicating whether the request was granted or denied. |

**Tab. 1 –** Features and target label of the machine learning model

### *2.1 Exploratory Data Analysis*

Income distribution showed a predominance of low-income petitioners, consistent with the target demographic of public legal aid policies. However, cross-variable analysis raised inconsistencies: while most applicants declared minimal income, some reported high levels of expenses or debts that seemed disproportionate. Histograms of total debt and total expense distributions highlighted this gap. While most petitioners reported low or zero debt, many also reported unusually high expenses. This discrepancy suggested either incomplete or inaccurate self-reporting, reinforcing the need for external data verification.

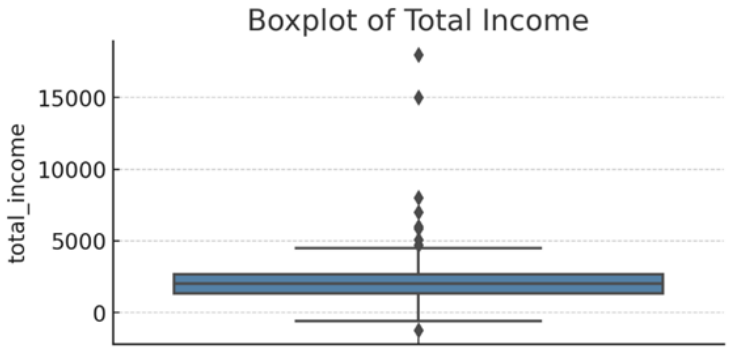


**Fig. 1 -** Total Income

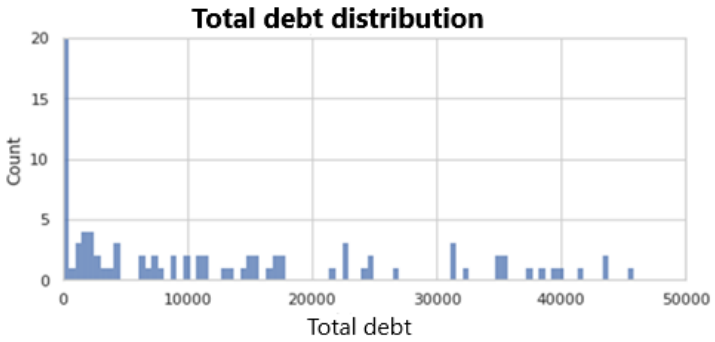


**Fig. 2 -** Total Debt

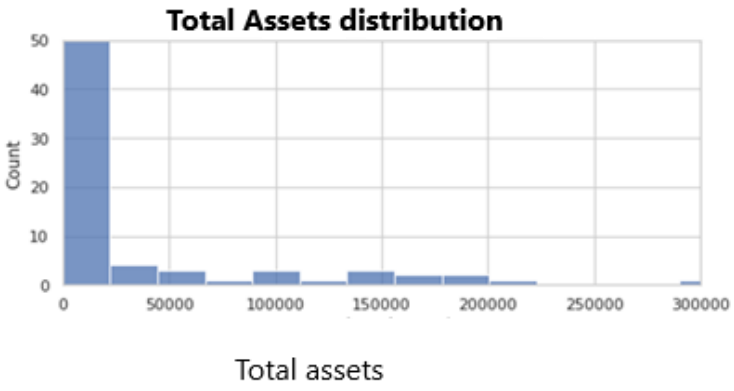


**Fig. 3 -** Total Assets

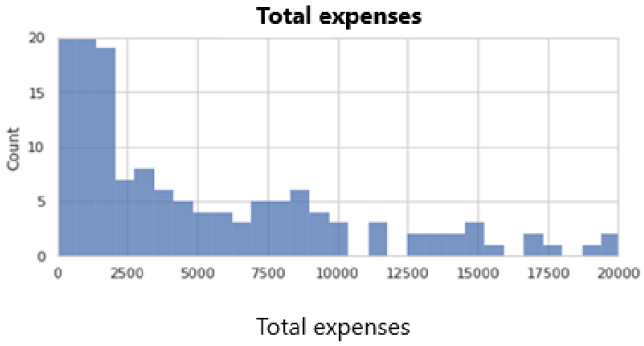


**Fig. 4 -** Total Expenses

Similarly, the distribution of total assets confirmed that most applicants owned few or no significant properties or financial holdings, yet there were outliers. These insights informed the feature selection process and pointed to the potential benefit of integrating external databases to validate self-declared information.

# 3 Modeling Approach

The modeling strategy followed a structured pipeline: 1. **Baseline Creation** – A dummy classifier always predicting the majority class ("granted") was used as a lower performance bound 2. **Preliminary Models** – Linear models such as logistic regression (Hosmer, Lemeshow, & Sturdivant, 2013), support vector machines (Cortes & Vapnik, 1995), and K-nearest neighbors (Altman, 1992) were tested to assess model fitness. 3. **Decision Tree Models** – Given their explainability, several decision tree-based (Quinlan, 1986) models were trained and fine-tuned via grid search 4. **Ensemble Models** – Random forests (Breiman, 2001) were evaluated to identify whether ensemble techniques (Breiman, 1996) improved predictive power 5. **Evaluation** – Models were evaluated using accuracy, balanced accuracy, and AUC under 10-fold cross-validation and hold-out test sets. The final training set comprised 85% of the dataset, with the remaining 15% held out for testing. Stratification ensured that class balance was preserved in both partitions.

# 4. Model Performance

The baseline dummy classifier achieved 63% accuracy and 50% balanced accuracy—highlighting the challenge posed by class imbalance. Logistic regression improved upon this with 71.7% test accuracy and 75.2% balanced accuracy. Support vector machines and KNN performed comparably, though KNN was more computationally intensive. Decision trees emerged as the most promising model class due to their transparency. A tree optimized via grid search (with max depth = 9 and min samples per leaf = 8) achieved 73.3% test accuracy and 77.9% balanced accuracy. Feature importance analysis revealed that income, unemployment status, and educational attainment were significant predictors.

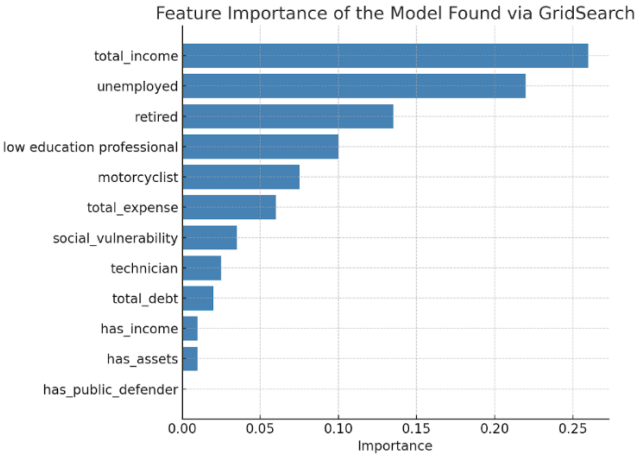


**Fig. 5 -** Feature importances of Decision Tree model

Random forests slightly outperformed decision trees in certain metrics and exhibited superior performance in low false-positive regions of the ROC space. However, decision trees were favored for high true-positive rates and greater interpretability. The model's ability to identify denial cases improved markedly over the baseline.

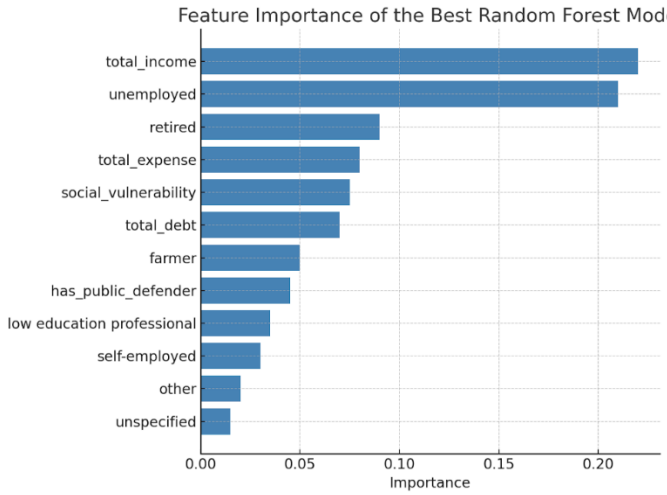


**Fig. 6 -** Feature importances of Random Forest model

After training all candidate models we then compared the top-performing models. The following picture shows the results.

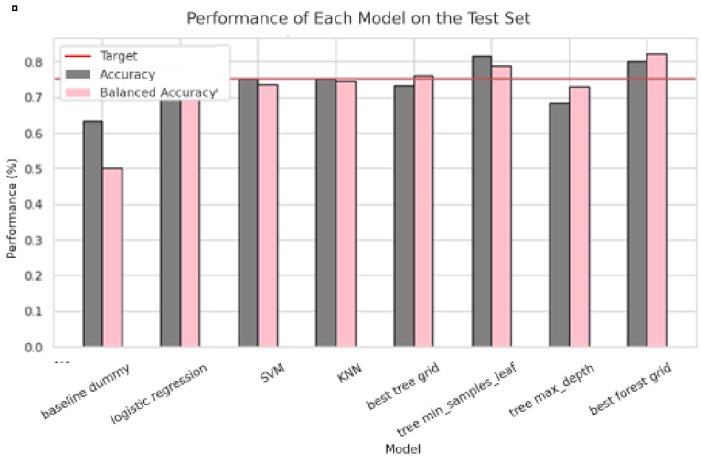


**Fig 7 -** Comparison of best models

ROC curves confirmed these findings, with decision trees and random forests significantly outperforming KNN.

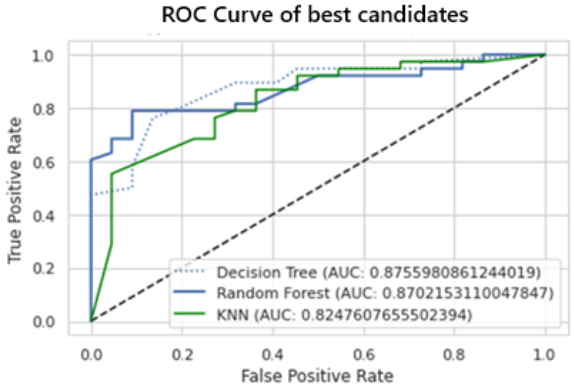


**Fig 8 –** ROC and AUC of best candidates

Confusion matrices indicated that the decision tree was more conservative (favoring grants), while random forests were more effective in denying non-eligible requests.

# 5. Discussion

Across all evaluation metrics, decision tree and random forest models emerged as the best-performing options. Decision tree model achieved the highest balanced accuracy and was better suited for legal contexts where minimizing false negatives (wrongly denying aid) is paramount. Model performance on the test set was summarized as follows: **Decision Tree**: 81.66% acc, 78.82% balanced acc; **Random Forest**: 80% acc, 82,29% balanced acc. Slightly higher performance in low-FPR zones, better at correctly denying aid. These results confirmed the validity of the proposed approach and its potential applicability to other judges or courts.

# 6. Next Steps and Policy Considerations

From a technical standpoint, the system could be improved through the use of ensemble methods such as Boosting or Stacking. However, these enhancements must be weighed against the loss of interpretability. We argue that explainability can be preserved through global feature importance analysis and counterfactual explanation tools. SHAP-IQ (Fumagalli et al., 2023) methods must be explored to better-grounded explainability. Another major frontier for improvement lies in feature quality. The inclusion of more granular asset data—such as vehicle ownership and real estate—would provide stronger predictive signals. As these features were underrepresented in the current dataset, they had limited influence on model performance. Access to such data is legally possible and technically feasible via integration with existing government platforms. On the systems side, user interfaces could be optimized based on feature relevance. We also propose the development of a *decision draft generator*—a system that automatically prepares preliminary judicial decisions for review—thus saving time while maintaining final control with human judges.

# 7. Conclusion

This project demonstrates the feasibility and utility of modeling judicial behavior through machine learning, with direct application to legal aid classification in the Brazilian judiciary. By combining historical decision analysis, external data retrieval, and interpretability techniques, the ARGOS system provides a powerful tool for increasing efficiency and transparency in legal decision-making. More broadly, the experience highlights how AI can respect legal autonomy while delivering practical benefits to public administration. The approach taken here—replicating an individual judge's logic, rather than imposing a top-down standard—offers a scalable, ethically sound model for digital governance. As courts worldwide face growing caseloads and limited resources, systems like this one represent a compelling frontier for responsible and human-centered AI in the public sector.

## Acknowledgement


- **Funding or Grant :** This work was supported by Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES), under grant TERMO DE OUTORGA DE SUBVENÇÃO ECONÔMICA No 311/2021 within EDITAL FAPES/SEGER/SEG No 05/2020.
- **Use of AI**: During the preparation of this work, the author used ChatGPT in order to translate to english. After using this tool/service, the author(s) reviewed, edited, made the content their own and validated the outcome as needed, and take full responsibility for the content of the publication.
- **Conflict Of Interest (COI)**: There is no conflict of interest.


## References


Bekkerman, R., Bilenko, M., & Culp, M. (2021). *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning Publications.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI*. European Commission. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Altman, N. S. (1992). *An introduction to kernel and nearest-neighbor nonparametric regression.* The American Statistician, 46(3), 175-185.

Breiman, L. (1996). *Bagging predictors.* Machine Learning, 24(2), 123-140.

Breiman, L. (2001). *Random forests.* Machine Learning, 45(1), 5-32.

Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273-297.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. John Wiley & Sons.

Quinlan, J. R. (1986). *Induction of decision trees.* Machine Learning, 1(1), 81-106.

Fumagalli, F., Muschalik, M., Kolpaczki, P., Hüllermeier, E., & Hammer, B. (2023). *SHAP-IQ: Unified Approximation of any-order Shapley Interactions.* In Advances in Neural Information Processing Systems (NeurIPS 2023).