Comparing Machine Learning and an Expert System for Legal Document Classification

José Jorge de Queiroz Santos Filho[a*], *Filipe* Araújo Dantas[b], *Melquezedeque* da Silva Lima[c], *Shirley* Barbosa dos Santos[d], *Galileu* Genesis[e], *Maria Gabriely* Lima da Silva[f], *Álvaro* Farias Pinheiro[g], *Eraylson* Galdino da Silva[h]

[a]University of Pernambuco - UPE, Recife, Pernambuco, Brazil, jjqsf@ecomp.poli.br, 0009-0004-6613-7752.

[b]University of Pernambuco - UPE, Recife, Pernambuco, Brazil, fad@ecomp.poli.br, 0000-0003-2452-2076.

[c]UFPE, Recife, Pernambuco, Brazil, msl5@cin.ufpe.br, 0009-0005-6308-8681.

[d]The Office of the State Attorney General of Pernambuco, Recife, Pernambuco, Brazil, 0009-0006-0286-3020.

[e]The Office of the State Attorney General of Pernambuco, Recife, Pernambuco, Brazil, galileugenesis@gmail.com, 0000-0003-2452-2076.

[f]University of Pernambuco - UPE, Recife, Pernambuco, Brazil, mgls@ecomp.poli.br, 0000-0002-3056-3985.

[g]The Office of the State Attorney General of Pernambuco, Recife, Pernambuco, Brazil, alvarofpinheiro@gmail.com, 0000-0002-6254-7293.

[h]University of Pernambuco - UPE, Recife, Pernambuco, Brazil, egs@ecomp.poli.br, 0000-0003-4287-9749 .

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

Abstract. This study assesses the performance of machine learning models and a rule-based expert system in classifying legal documents, specifically in distinguishing relevant from irrelevant cases. The evaluated models include Random Forest, Naive Bayes, XGBoost, SVM, and Decision Tree, alongside an expert system developed by a State Attorney from PGE-PE. The datasets, representing *Alvará*, *Arrolamento*, and *Inventário* legal processes, contain labeled instances of legal cases. The models were assessed based on accuracy, precision, recall, and F1-score. The results suggest that while machine learning models—particularly Random Forest—achieve higher accuracy and precision, the expert system outperforms in recall and F1-score, ensuring that no relevant cases are overlooked. The choice between machine learning models and expert systems depends on the legal context, requiring a balance between efficiency (reducing false positives) and reliability (capturing all relevant cases).

Keywords. machine learning, legal document classification, expert systems, overfitting, natural language processing
Research paper, DOI: https://doi.org/10.59490/dgo.2025.947

1. Introduction

The Office of the State Attorney General of Pernambuco (in Portuguese, *Procuradoria Geral do Estado de Pernambuco* - PGE-PE) is the government agency responsible for representing the State of Pernambuco in both judicial and extrajudicial matters. Currently, it operates with a limited staff to manage over 600,000 cases, where the efficient handling of this workload aims to protect the rights and interests of the State. This large volume of cases, combined with a workforce that partially lacks legal training, makes the proper distribution of these cases a challenge. Additionally, not all cases that go through the judiciary and involve the Attorney General's Office hold significant financial or social relevance, which affects the extent of public administration efforts. Proper classification of these cases facilitates the appropriate transfer of assets within the Attorney General's Office and ensures the protection of public funds.

This context underscores the critical need for developing an automated system capable of managing the intricate and specialized nature of legal texts, thereby enhancing operational efficiency and directing limited resources toward high-priority cases. One solution lies in implementing a system specifically designed for the classification of legal documents. The literature identifies two primary approaches to building such systems. The first approach involves rule-based expert systems (RBS), which rely on a knowledge base consisting of a set of rules written by experts(Masri et al., 2019). The second approach utilizes machine learning (ML) models, allowing the system to identify patterns that define each document category (Magalhães et al., 2022).

RBS present both advantages and limitations when applied to the classification of legal documents. On the positive side, these systems offer simplicity and transparency, as their deterministic ”if-then” rules make the decision-making process straightforward to understand and validate (Sasikumar, 2007). This level of transparency is particularly advantageous in legal contexts, where explainability is essential to fostering user trust and ensuring compliance (Zhou et al., 2023). However, this approach also comes with significant challenges. As the rule base expands, maintaining the system becomes increasingly complex, leading to difficulties in managing rule interdependencies and preserving operational efficiency. Additionally, constructing the knowledge base is a resource-intensive process, requiring extensive input from legal experts to accurately codify rules. This dependency on expert knowledge can be a significant limitation during system development(Westermann et al., 2019).

Machine learning (ML) models offer a dynamic solution to the limitations of rule-based systems in legal document classification. Unlike static rule-based approaches, ML models learn patterns directly from data, enabling adaptation to evolving legal standards and new case types without extensive manual updates. They generalize across diverse document formats, identify nuanced patterns, and efficiently handle large datasets with minimal human intervention after training. This adaptability and scalability position ML models as an alternative or complement to traditional rule-based systems(Prentzas & Hatzilygeroudis, 2007; Westermann et al., 2019).

In this context, this research was conducted with the objective of answering two key questions: (1) Are the evaluated models capable of learning the patterns embedded in legal documents and performing accurate classifications? and (2) Can machine learning models outperform rule-based expert systems developed by domain specialists when evaluated on key classification performance metrics such as accuracy, precision, recall, and F1-score? By addressing these questions, the study conducts a comparative analysis of machine learning models commonly used in document classification, such as XGBoost (Chari et al., 2021), Support Vector Machines (SVM) (Al Hasan et al., 2022), Random Forest (Bento & Teive, 2023), Naive Bayes (Dias & Cavalcante, 2023), and Decision Tree (Gêda et al., 2021), in comparison to a rule-based system built using expert-defined rules.

The comparison evaluates the performance of these approaches in classifying the relevance of document cases related to the Tax on Inheritance and Donation of Goods or Rights, which is referred to in Brazil as ICD (in Portuguese, *Imposto sobre Transmissão Causa Mortis e Doação de quaisquer Bens ou Direitos* - ICD). This domain presents unique challenges due to the legal and procedural complexity of the documents, as well as their substantial volume. These challenges include varying page layouts within the same case, processes of different lengths, and the specialized vocabulary of legal documents. The objective is to identify the approach that provides the most reliable and efficient solution for optimizing resource allocation and supporting decision-making within legal institutions. The main contributions of this study can be summarized as follows:

- This study provides a comprehensive comparative analysis of various classification models and a rule-based expert system, evaluating their performance in classifying the relevance of processes within the context of ICD documents.
- It was verified that the use of machine learning models enhances accuracy and precision but struggles with recall, highlighting a trade-off that must be considered in legal document classification.
- To the best of our knowledge, this is the first work to compare the use of machine learning models with a rule-based expert system in the context of ICD.
- This work sheds light on the challenges machine learning models face in surpassing expert systems, highlighting their limitations in recall and generalization despite achieving higher accuracy and precision.

The remainder of this article is structured as follows: Section 2 reviews related work, providing context and background for our study. Section 3 introduces the proposed methodology and details the experimental protocol applied. Section 4 presents and discusses the results, offering insights into models performances. Finally, Section 5 concludes the paper by addressing research limitations and outlining potential directions for future work.

2. Related Works

Several studies have addressed the challenge of applying machine learning to document classification in legal contexts, highlighting its potential for improving efficiency and accuracy. The classification of legal documents has greatly advanced with the development of machine learning and natural language processing (NLP) techniques. These technologies facilitate the efficient analysis and categorization of vast amounts of legal texts, streamlining processes and significantly reducing the manual workload for law firms and judicial institutions (Noguti et al., 2020).

For instance, (Magalhães et al., 2022) conducted an experimental analysis using machine learning techniques to group and classify judicial decisions. The study examined a dataset of 430,000 decisions from the Regional Labor Court, employing text representation models based on BERT alongside clustering algorithms such as K-Means. The findings demonstrated that these methods effectively capture the semantic nuances of legal texts, offering significant support for decision-making processes within Brazil's judicial system. This research highlights the potential of advanced ML techniques to enhance the organization and analysis of complex legal data.

The work of (Polo et al., 2021) evaluated the performance of several classifiers combined with different feature extraction approaches to classify Brazilian court cases into three status categories: archived, active, and suspended The study achieved an accuracy of over 93% in the classification tasks, highlighting the effectiveness of XGBoost, MLP, and LSTM models when integrated with feature extraction methods such as BERT and TF-IDF.

Serras and Finger, (Serras & Finger, 2022) conducted a study on the use of attention-based algorithms to automate the categorization of Brazilian legal documents. The authors highlighted several challenges in the classification of legal documents, including redundancy, the use of generic terms, inconsistencies in vocabulary, and poorly defined syntactic structures within and between terms. These challenges are also present in ICD legal documents, a context in which this study aims to evaluate machine learning models.

The works found in the literature corroborate the argument that machine learning approaches can enhance the analysis and classification of legal documents. To the best of our knowledge, this is the first study to investigate the use of machine learning models for classifying the relevance of legal documents in the ICD context within a governmental organization. Furthermore, it compares these models with an expert system based on rules developed by a State Attornes.

3. Methodology and Experimental Protocol

The methodology used to build machine learning models for legal document classification consists of two main steps: preprocessing and modeling training.

3.1. Preprocessing

In the preprocessing stage, data cleaning, normalization and standardization, outlier detection and treatment, and data transformation techniques were applied. This step is crucial for improving data quality, thereby contributing to the accuracy, efficiency, and reliability of the analyzed models. The goal is to prepare raw data for analysis, ensuring it is clean, consistent, and suitable for modeling algorithms or subsequent analyses. This is essential because raw data often contains issues such as missing values, inconsistencies, noise, or inappropriate formats, which can negatively impact the results.

As previously mentioned, during the data preparation stage, a conversion function was used to transform classifications into binary values, enabling the use of vectorization with term frequency-inverse document frequency (TF-IDF). TF-IDF is a measure that combines the local importance of a word in a specific document (TF) with the global importance of that word across a collection of documents (IDF)(Cahyani & Patasik, 2021). The interpretation of TF-IDF provides insights into the relative significance of words within a specific document compared to the entire corpus. The higher the TF-IDF value, the more important the word or term.

Words with higher TF-IDF values can serve as key terms that help distinguish different documents or define the primary topics or themes of a document. These words can be used to summarize document content and construct classification models. In the TF-IDF implementation, we included only tokens that appeared in at least five documents and excluded those with a frequency greater than 90% across the corpus. Another consideration in this technique was limiting the total vocabulary size to 25,000 tokens, which corresponds to the total number present in the analyzed dataset.

3.2. Model Training

For the classification of legal texts using Machine Learning, we studied the application of five models. The first was SVM (Support Vector Machine), a supervised learning algorithm that utilizes a decision boundary, known as a hyperplane, to segregate the analyzed data and define different classes (Al Hasan et al., 2022). A notable observation about this model is that it can become inefficient when dealing with large datasets.

Another model used was Naive Bayes, a probabilistic algorithm based on Bayes' theorem that assumes feature independence. This simplification facilitates classification by analyzing only the presence or absence of specific features, which can also lead to inaccuracies in its results (Dias & Cavalcante, 2023).

Moving on to tree-based models, we also employed Decision Tree, which generates an inverted tree to make predictions. At its nodes, questions about data features are asked, branching based on the answers (Gêda et al., 2021). A critical limitation of this algorithm is its tendency toward overfitting.

The fourth model used was Random Forest. This classifier combines multiple decision trees, each trained on a random sample of the data, and considers a random subset of features to predict classifications through majority voting (Bento & Teive, 2023). However, a drawback of this algorithm is its high computational cost.

Finally, we employed XGBoost (Extreme Gradient Boosting), a tree-based algorithm where multiple weak trees are combined sequentially to create a strong model (Chari et al., 2021). This method is also prone to overfitting and requires careful tuning of its numerous hyperparameters to achieve optimal performance.

Due to the limitation of computational resources, hyperparameter optimization was not performed. For this reason, we used the values reported in the literature. The hyperparameters for these models are detailed in Table 1.

Tab. 1 – Hyperparameters used in model configurations

Model	**Hyperparameter**	**Value**
XGBoost	max_depth	2
	learning_rate	0.1
	n_estimators	600
	random_state	42
	scale_pos_weight	11.5
SVM	C	0.01
	gamma	0.001
	kernel	'linear'
Random Forest	n_estimators	980
	max_depth	48
Naive Bayes	var_smoothing	1e-9
	priors	None
Decision Tree	max_depth	10

The rule-based system developed utilizes a dictionary created by State Attorneys from the Succession and Donation Division of PGE-PE, the department responsible for overseeing the collection of taxes on inheritances and donations from individuals and legal entities. In this dictionary, weights were assigned to words and expressions such as legal terms, assets, and values, indicating the degree of relevance of these terms, thus contributing to the final classification of the processes.

3.3. Datasets

The processes utilized in this study originate from the Succession and Donation Division (NSD) of the Pernambuco State Attorney General's Office (PGE-PE), a department responsible for monitoring the collection of taxes on successions and donations from individuals and legal entities in the state of Pernambuco.

For this study, three distinct datasets were analyzed, each containing cases labeled as RELEVANT and IRRELEVANT. Each dataset corresponds to a judicial class: *Inventário*, characterized as the formal process of asset division among heirs; *Alvará*, representing a formal authorization issued by an authority for specific activities related to the inventory; and *Arrolamento*, a streamlined inventory process for low-value cases.

In the *Inventário* dataset, which includes 11,839 records, a total of 11,566 RELEVANT cases and 273 IRRELEVANT cases were identified. In the *Arrolamento* dataset, containing 647 records, 354 RELEVANT cases and 293 IRRELEVANT cases were detected. In the *Alvará* dataset, with 392 records, 58 RELEVANT cases and 334 IRRELEVANT cases were recorded.

For the *Alvará* cases, the average word count per case was 6,849.64, with a maximum of 11,251 words and a minimum of 920 words. For the *Arrolamento* cases, the average word count per case was 7,637.70, with a maximum of 11,710 words and a minimum of 1,051 words. For the *Inventário* cases, the average word count per case was 8,341.11, with a maximum of 22,714 words and a minimum of 1 (one) word.

To conduct the experiment, the datasets were divided into training and testing sets, with **80%** allocated for training and **20%** for testing.

3.4. Performance Metrics

The system evaluation uses metrics already applied in the literature for classification tasks, which include: Accuracy – used to determine the proportion of correct predictions relative to the total predictions made; Precision – measures the proportion of true positives among all positive predictions; Recall – identifies the proportion of true positives correctly identified among all actual positive samples; F1-Score – defines the harmonic mean between Precision and Recall. However, this metric does not account for the class proportions in the dataset; and Confusion Matrix – used to graphically visualize the model's performance in classifying each class, displaying true positives, true negatives, false positives, and false negatives.

In this study, we classify processes as RELEVANT for the positive class (P) and IRRELEVANT for the negative class (N). Thus, TP (True Positive) refers to the number of samples correctly classified as positive; TN (True Negative) is the number of samples correctly classified as negative; FP (False Positive) represents the number of samples incorrectly classified as positive, and FN (False Negative) is the number of samples incorrectly classified as negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

4. Experiments and Results

To address the question, ”Are the evaluated models capable of learning the patterns embedded in legal documents and performing accurate classifications?” the performance of the classifiers was assessed using both training and test samples for each dataset. This evaluation aimed to determine their ability to generalize beyond the data on which they were trained.

Table 2 show the performance of the models on the *Alvará* dataset. Regarding accuracy, most models, including XGBoost, Random Forest, and Naive Bayes, achieve comparable results on the test set (0.88). Accuracy measures the proportion of correctly classified instances across all classes and provides a general overview of the model's performance. However, the *Alvará* dataset exhibits an imbalanced class distribution, with only 18.47% of instances belonging to the relevant class. Therefore, metrics such as Precision and Recall are more critical for evaluating the models' effectiveness in identifying relevant instances.

The precision metric evaluates the proportion of correctly identified relevant instances among all instances retrieved by the model, with higher values indicating a lower rate of false positives. In this regard, the Random Forest and Naive Bayes models achieved the best performance, each with a precision of 94% on the test sample. On the other hand, the recall metric measures the proportion of relevant instances correctly identified by the model out of all relevant instances in the dataset, with higher values reflecting a lower rate of false negatives. For this metric, the Decision Tree model demonstrated the best performance, achieving a recall of 0.72.

When comparing the training and testing results, it is evident that most models achieved a perfect performance on the training sample. However, this performance dropped drastically on the test sample, indicating that the models were significantly affected by the overfitting problem.

To address the second question, “Can machine learning models outperform rule-based expert systems developed by domain specialists when evaluated on key classification performance metrics such as accuracy, precision, recall, and F1-score?”, the performance of the classifiers was evaluated using test samples from each dataset and compared against the performance of an expert system developed by a State Attorney from the PGE-PE.

Model	Training				Testing			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
XGBoost	1.0000	1.0000	1.0000	1.0000	**0.8861**	0.8459	0.6592	0.7034
SVM	0.8675	0.4338	0.5000	0.4645	0.8481	0.4241	0.5000	0.4589
Random Forest	1.0000	1.0000	1.0000	1.0000	**0.8861**	**0.9408**	0.6250	**0.7685**
Naive Bayes	1.0000	1.0000	1.0000	1.0000	**0.8861**	**0.9408**	0.6250	0.6685
Decision Tree	1.0000	1.0000	1.0000	1.0000	0.8228	0.6791	**0.7245**	0.6962

Tab. 2 – Performance metrics of models on the *Alvará* dataset during training and testing, with the best results for each metric highlighted in bold.

Model	Training				Testing			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
XGBoost	1.0000	1.0000	1.0000	1.0000	0.7154	0.7168	0.6957	0.6982
SVM	0.5581	0.2791	0.5000	0.3582	0.5692	0.2846	0.5000	0.3627
Random Forest	1.0000	1.0000	1.0000	1.0000	**0.7462**	**0.7438**	**0.7314**	**0.7347**
Naive Bayes	0.9948	0.9942	0.9954	0.9948	0.5923	0.5778	0.5000	0.5377
Decision Tree	0.9793	0.9821	0.9766	0.9789	0.7077	0.7050	0.6911	0.6935

Tab. 3 – Performance metrics of models on the *Arrolamento* dataset during training and testing, with the best results for each metric highlighted in bold.

For the *Arrolamento* dataset, Table 3 presents the performance metrics of the models during training and testing. In this case, the Random Forest model achieved the best results across all evaluated metrics. However, it is evident that overfitting also affected the models in this dataset. For example, the Naive Bayes model achieved a performance above 0.99 on the training sample but dropped to approximately 0.60 on the testing sample.

Model	Training				Testing			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
XGBoost	0.9958	0.9978	0.9143	0.9520	0.9764	0.6553	0.5087	0.5113
SVM	0.9754	0.4877	0.5000	0.4938	0.9768	0.4884	0.5000	0.4941
Random Forest	0.9959	0.9979	0.9171	0.9538	**0.9772**	**0.9886**	0.5091	0.5121
Naive Bayes	0.9879	0.8352	0.9938	0.8983	0.9700	0.5725	**0.5320**	**0.5430**
Decision Tree	0.9835	0.9834	0.6685	0.7469	0.9709	0.5639	0.5236	0.5326

Tab. 4 – Performance metrics of models on the *Inventário* dataset during training and testing, with the best results for each metric highlighted in bold.

Table 4 show the performance of the models on the *Inventário* dataset. The Random Forest model demonstrated a superior performance with the highest accuracy (0.9772) and precision (0.9886), highlighting its ability to correctly identify relevant cases. For the Recall and F1-Score, the best results are achieved by the Naive Bayes, demonstrating its ability to effectively balance the trade-off between precision and recall, particularly in identifying relevant instances in the dataset. When comparing the training and testing results, it is possible to observe that the accuracy metric values are relatively close in both samples. However, when analyzing the Precision and Recall results, the impact of overfitting becomes evident in most of the models. This highlights the importance of evaluating models using multiple metrics to gain a comprehensive understanding of their performance.

Addressing the first research question that guided this study, the results suggest that while the evaluated models exhibit the capacity to learn patterns in legal documents, their practical utility is constrained by certain limitations. The high accuracy achieved across models highlights their potential for the task of legal document classification; however, the reliance on this metric alone can be misleading, particularly in imbalanced datasets where metrics such as precision and recall are more indicative of real-world performance. Furthermore, the significant drop in performance from training to testing data underscores the pervasive issue of overfitting, raising concerns about the models' ability to generalize beyond the training set.

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Alvará	XGBoost	**0.8861**	0.8459	0.6592	0.7034
	SVM	0.8481	0.4241	0.5000	0.4589
	Random Forest	**0.8861**	**0.9408**	0.6250	0.6885
	Naive Bayes	**0.8861**	**0.9408**	0.6250	0.6685
	Decision Tree	0.8228	0.6791	0.7245	0.6962
	Expert System	0.7595	0.8557	**0.7595**	**0.7883**
Arrolamento	XGBoost	0.7154	0.7168	0.6957	0.6982
	SVM	0.5692	0.2846	0.5000	0.3627
	Random Forest	**0.7462**	**0.7438**	**0.7314**	**0.7347**
	Naive Bayes	0.5923	0.5778	0.5702	0.5677
	Decision Tree	0.7077	0.7050	0.6911	0.6935
	Expert System	0.7000	0.6977	0.7000	0.6985
Inventário	XGBoost	0.9764	0.6553	0.5087	0.5113
	SVM	0.9768	0.4884	0.5000	0.4941
	Random Forest	**0.9772**	**0.9886**	0.5091	0.5121
	Naive Bayes	0.9700	0.5725	0.5320	0.5430
	Decision Tree	0.9709	0.5639	0.5236	0.5326
	Expert System	0.9624	0.9658	**0.9624**	**0.9658**

Tab. 5 – Comparison of model performance with the Expert System on the test set for the datasets *Alvará*, *Arrolamento* and *Inventário*, with the best results for each metric highlighted in bold.

To address the second question, "Can machine learning models outperform rule-based expert systems developed by domain specialists when evaluated on key classification performance metrics such as accuracy, precision, recall, and F1-score?", the performance of the classifiers was evaluated using test samples from each dataset and compared against the performance of an expert system developed by a State Attorney from the PGE-PE.

For the *Alvará* dataset, Random Forest, Naive Bayes, and XGBoost achieve the highest accuracy (88.61%), outperforming the Expert System (75.95%). This suggests that the ML models are generally better at correctly classifying cases across all classes. However, the Expert System achieves the highest recall (0.7595) and F1-score (0.7883), indicating that it is more effective in identifying relevant legal cases (i.e., those that should be classified as positive). This could be due to the rule-based nature of the Expert System, which is likely designed with legal expertise to minimize false negatives—ensuring that important cases are not overlooked. In contrast, while ML models have higher precision (Random Forest and Naive Bayes: 0.9408), their lower recall (0.6250) means they may fail to capture some relevant cases. This trade-off is critical, higher precision reduces false positives but may lead to missed relevant cases, whereas higher recall ensures that fewer important cases are ignored.

For the *Arrolamento* dataset, Random Forest outperforms the Expert System across all metrics, achieving the highest accuracy (0.7462), precision (0.7438), recall (0.7314), and F1-score (0.7347). This suggests that, for this dataset, the ML model is not only more consistent in correctly classifying cases overall, but it also maintains a better balance between precision and recall. The Expert System achieves slightly lower but comparable performance across all metrics (accuracy: 0.7000, precision: 0.6977, recall: 0.7000, and F1-score: 0.6985). This indicates that for *Arrolamento* cases, the ML model is more effective than the rule-based system, likely because legal patterns in this dataset are more learnable from the data than from fixed rules. The fact that ML models can optimize both precision and recall in this case suggests they may be better suited for handling the complexities of classification in this specific legal context.

For the *Inventário* dataset, Random Forest achieves the highest accuracy (0.9772) and precision (0.9886), meaning it is highly effective at reducing false positives, which is crucial when the goal is to ensure that only highly relevant cases are classified as such. However, the Expert System achieves the highest recall (0.9624) and F1-score (0.9658), meaning it is more effective at capturing all relevant cases—even at the cost of slightly lower precision. However, the Expert System achieves the highest recall (0.9624) and F1-score (0.9658), meaning it is more effective at capturing all relevant cases—even at the cost of slightly lower precision. This suggests that, for the *Inventário* dataset, the Expert System ensuring that relevant cases are not ignoreted, which is particularly important in legal contexts where missing a critical instance could have significant consequences.

Addressing the second research question, the results indicate that ML models generally achieve higher accuracy and precision, making them valuable for reducing false positives—an essential factor if the legal system aims to streamline case classification and avoid unnecessary workload on legal professionals. However, the Expert System often performs better in recall and F1-score, suggesting it is more cautious and prioritizes the inclusion of all potentially relevant cases. This is particularly important in legal contexts where missing a relevant case (false negatives) could have serious consequences, such as delays in legal proceedings or overlooked claims.

Ultimately,it can be inferred that the choice between ML models and the Expert System depends on the legal priorities for classification. If efficiency and reducing false positives are the primary goals, ML models—particularly Random Forest—are superior. However, if ensuring that no relevant case is missed is more critical, the Expert System remains highly competitive, particularly in recall-focused applications.

Furthermore, in this study, the models were evaluated using a baseline pipeline, applying TF-IDF for text embedding and implementing classical classifiers with hyperparameters derived from the literature. However, the results indicate that, despite achieving higher performance in metrics such as accuracy and precision across most datasets, the models were significantly affected by overfitting and failed to attain high performance in recall and F1-score, which are crucial for identifying relevant cases.

Thus, to enhance the performance of these models in legal document classification, future research should prioritize mitigating overfitting through advanced regularization techniques (Moradi et al., 2020), the use of contextual document embeddings (Morris & Rush, 2024), under-sampling or over-sampling strategies (Taha et al., 2021), use of hybrid approach Villena Román et al., 2011, and hyperparameter fine-tuning (Bischl et al., 2023). Effectively addressing these challenges would improve the models' generalization to unseen data, ultimately making them more practical, reliable, and well-suited for real-world legal applications.

5. Conclusion

This study evaluated the performance of machine learning models and a rule-based expert system in classifying legal documents, specifically in categorizing relevant and irrelevant cases. The evaluated models included Random Forest, Naive Bayes, XGBoost, SVM, and Decision Tree, alongside an expert system developed by a State Attorney from PGE-PE. The datasets, representing *Alvará*, *Arrolamento* and *Inventário* legal processes, contained labeled instances of legal cases. The models were assessed based on accuracy, precision, recall, and F1-score, providing a comprehensive assessment of their effectiveness.

The findings of this study reinforced the potential of machine learning models for legal document classification, particularly in terms of precision and accuracy. The Random Forest, Naive Bayes, and XGBoost models achieved promising performance, with accuracy rates ranging from 74.62% to 97.72% across different datasets. However, as observed in previous research on machine learning in the legal domain, these models exhibited susceptibility to overfitting, particularly in recall and F1-score, emphasizing the need for techniques that could address this issue.

In contrast, the rule-based expert system demonstrated superior recall and F1-score, ensuring the identification of most relevant cases, a crucial aspect in legal contexts where false negatives could be highly problematic. However, consistent with findings in the literature, this approach had limitations regarding flexibility and precision, particularly when applied to large-scale unstructured data.

Despite the advancements presented in this study, some limitations needed to be acknowledged. First, the models were trained and tested on a specific dataset, and their generalizability to other legal domains might have required additional fine-tuning. Additionally, the absence of advanced class-balancing techniques may have influenced the recall and F1-score of the machine learning models. Future research should focus on mitigating these challenges by exploring enhanced regularization, contextual embeddings, sampling methods, hybrid approaches, and hyperparameter tuning to improve performance and robustness.

By addressing these challenges, machine learning models could become even more efficient and reliable for practical applications in the legal sector, contributing to the automation and optimization of legal document classification.

References

Al Hasan, S., Hussain, M. G., Protim, J., Rahman, M. M., Fahim, N., Chowdhury, M. Z., & Pritom, A. I. (2022). Classification of multi-labeled text articles with reuters dataset using svm. *2022 International Conference on Science and Technology (ICOSTECH)*, 01–05.

Bento, F. M., & Teive, R. C. G. (2023). Classificação de documentos jurídicos utilizando a arquitetura transformer: Uma análise comparativa com algoritmos tradicionais de machine learning e chatgpt. *Brazilian Journal of Development*, *9*(6), 20208–20224.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., et al. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(2), e1484.

Cahyani, D. E., & Patasik, I. (2021). Performance comparison of tf-idf and word2vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, *10*(5), 2780–2788.

Chari, H., Aswale, S., Pawar, V. N., Shetgaonkar, P., & Kumar, K. C. (2021). Advertisement click fraud detection using machine learning techniques. *2021 International Conference on Technological Advancements and Innovations (ICTAI)*, 109–114.

Dias, L. C. M., & Cavalcante, L. G. M. (2023). Aplicação do classificador naive bayes para detecção de fraudes. *Ciência Da Computação: Avanços E Tendências Em Pesquisa*, *1*, 9–26.

Gêda, B. M., et al. (2021). Classificação de textos de decisões judiciais.

Magalhães, D., Pozo, A., & Machado, S. (2022). Técnicas de aprendizado de máquinas aplicadas à classificação de decisões judiciais. *Revista de Estudos Empíricos em Direito*, *9*.

Masri, N., Sultan, Y. A., Akkila, A. N., Almasri, A., Ahmed, A., Mahmoud, A. Y., Zaqout, I., & Abu-Naser, S. S. (2019). Survey of rule-based systems. *International Journal of Academic Information Systems Research (IJAISR)*, *3*(7), 1–23.

Moradi, R., Berangi, R., & Minaei, B. (2020). A survey of regularization strategies for deep models. *Artificial Intelligence Review*, *53*(6), 3947–3986.

Morris, J. X., & Rush, A. M. (2024). Contextual document embeddings. *arXiv preprint arXiv:2410.02525*.

Noguti, M. Y., Vellasques, E., & Oliveira, L. S. (2020). Legal document classification: An application to law area prediction of petitions to public prosecution service. *2020 International joint conference on neural networks (IJCNN)*, 1–8.

Polo, F. M., Ciochetti, I., & Bertolo, E. (2021). Predicting legal proceedings status: Approaches based on sequential text data. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 264–265.

Prentzas, J., & Hatzilygeroudis, I. (2007). Categorizing approaches combining rule-based and case-based reasoning. *Expert Systems*, *24*(2), 97–122.

Sasikumar, M. (2007). A practical introduction to rule based expert systems.

Serras, F. R., & Finger, M. (2022). Verbert: Automating brazilian case law document multi-label categorization using bert. *arXiv preprint arXiv:2203.06224*.

Taha, A. Y., Tiun, S., Abd Rahman, A. H., & Sabah, A. (2021). Multilabel over-sampling and under-sampling with class alignment for imbalanced multilabel text classification. *Journal of Information and Communication Technology*, *20*(3), 423–456.

Villena Román, J., Collada Pérez, S., Lana Serrano, S., & González Cristóbal, J. C. (2011). Hybrid approach combining machine learning and a rule-based expert system for text categorization.

Westermann, H., Šavelka, J., Walker, V. R., Ashley, K. D., & Benyekhlef, K. (2019). Computer-assisted creation of boolean search rules for text classification in the legal domain. In *Legal knowledge and information systems* (pp. 123–132). IOS Press.

Zhou, X., Du, H., Sun, Y., Ren, H., Cui, P., & Ma, Z. (2023). A new framework integrating reinforcement learning, a rule-based expert system, and decision tree analysis to improve building energy flexibility. *Journal of Building Engineering*, *71*, 106536.