CENTRE DE RECHERCHE
SUR LES STRATÉGIES ÉCONOMIQUES

Critical AI Challenges in Legal Practice: An application to French Administrative Decisions

Khaoula Naili

November 2023

Working paper No. 2023 – 06

CRESE
30, avenue de l'Observatoire
25009 Besançon
France
http://crese.univ-fcomte.fr/

UFR SJEPG
Sciences juridiques économiques politiques et de gestion
UNIVERSITÉ DE FRANCHE-COMTÉ

Critical AI Challenges in Legal Practice : An application to French Administrative Decisions

Naili Khaoula*

PhD student, University of Franche-Comté, CRESE, F-25000 Besançon, France.

24 novembre 2023

Résumé

We use AI methods to evaluate the accuracy of several standard machine learning models for predicting judicial decision outcomes. We highlight the key steps and challenges in predicting judicial outcomes by applying these models to a database of administrative court decisions.These findings significantly contribute to our understanding of the potential advantages of AI in the context of predictive justice. We utilize AI methods to analyze administrative court decisions sourced from the database provided by the French Council of State. This analysis has been made possible due to the Council of State's decision to make its decisions publicly accessible since March 2022. Our innovative approach pioneers the use of prediction models on the open data from the French Council of State, addressing the complexities associated with data analysis. Our primary objective is to assess the accuracy of these models in predicting outcomes in French administrative tribunals and identify the most effective model for forecasting administrative tribunal court decisions. The selected models are trained and evaluated on multi-class datasets, where decisions are traditionally categorized into various classes.

Keywords : artificial intelligence, machine learning, natural language processing, Predictive justice, Legal text.

JEL :K4

*khaoula.naili@univ-fcomte.fr

1 Introduction

Artificial intelligence (AI) is a scientific field that has a history of more than half a century, originating with John McCarthy's initial concept (Bini, 2018). This concept revolved around the idea that computers could eventually learn to perform tasks by recognizing patterns, often with minimal human intervention. While this theory may seem outdated in today's rapidly evolving technological landscape, the practical applications of AI are just now gaining prominence in the public domain as mature technology becomes increasingly accessible, thanks in part to the decreasing costs of computational power. This accessibility has paved the way for faster and more cost-effective computational solutions in various sectors, including the legal field.

Due to the distinctive ability of intelligent machines to learn and automatically adapt to new tasks based on prior experiences or provided data, there is a growing interest in harnessing AI for various purposes. Machine learning (ML), a subset of artificial intelligence (AI), relies on computational algorithms to acquire and enhance its capabilities through experiential learning. In its most basic form, this involves leveraging real-world datasets to anticipate or estimate outcomes. These datasets are essentially referred to as "training sets," which the machine meticulously studies. The algorithm then employs pattern recognition to make autonomous decisions. These inferences are subsequently assessed against a "testing set" consisting of actual outcomes to quantify the algorithm's accuracy. As the volume of data in the training sets expands and the number of testing iterations increases, much like a process of "experiential learning," the machine's algorithm continually improves its predictive accuracy.

From the 2000's, there has been an increasing fascination with harnessing the potential of artificial intelligence (AI) for the analysis of extensive datasets, including legal data. This application, commonly referred to as predictive justice, has received considerable attention. To be more specific, predictive justice involves utilizing data to make predictions about judicial decisions. It focuses on the application of data-driven methods to anticipate and forecast the outcomes of legal cases.

Many machine learning algorithms can be used for prediction. However, there are specific steps and issues that need to be addressed in order to implement these algorithms in an efficient way. The objective of our paper is to present these steps and the main challenges (e.g. the imbalanced data problem) of implementing machine learning algorithms to predict the outcomes of legal cases.

To illustrate this approach, we apply it to database of french administrative court decisions, made available by a law known as the Digital Republic Act (G'sell, 2020). This law mandated the free access to all data generated by public administrations, including decisions from all courts. [1] This reform is implemented, and enable the creation of comprehensive judicial databases that can be utilized by machine learning algorithms.

Additionally, within the realm of the French Council of State decisions, there has been a noticeable shift towards enhanced accessibility facilitated by new platforms. One such platform is the Open Data

1. Subsequently, the Justice Programming Law,Law No. 2019-222 enacted on March 23, 2019, made slight modifications to Article L. 111-13 of the Code of Judicial Organization. As a result, it now requires that French court decisions be made available to the public free of charge in electronic format.

of administrative justice,[2] which enables the availability of administrative court decisions in an open format. Another platform is the Ariana web jurisprudence base,[3] which is available online and contains over 270,000 court decisions of the Council of State and administrative courts of appeal selected for their jurisprudential interest. The base provides search functionalities, and some decisions of major jurisprudential interest are accompanied by an analysis of the case and conclusions of the public report. The decisions are presented in an XML file format, which contains various pieces of information such as the identification, date, jurisdiction, and the body of the decision rendered.

The open data of judicial decisions in France is a positive development, but it has some limitations and challenges (Cluzel-Métayer, 2016). These include the limited scope of publicly available decisions, data format issues, data quality concerns, language barriers, limited context, and accessibility issues. Despite these limitations, open data of judicial decisions in France can still promote transparency and accountability in the justice system and provide useful insights for researchers.

The open data of judicial decisions in France has been used in various ways to gain insights into the legal system and inform lawyers and policy-making. Examples include using the data to study patterns of discrimination in the justice system(Li, 2017), investigate cases of alleged police misconduct(DALE, 2019), monitor and advocate for legal reform (Shapiro, 2017), study the application of contract law (Kolt, 2022), and stay up to date on legal developments in a specific field. The open data of judicial decisions has provided valuable information to better understand the workings of the French legal system and promote transparency and accountability.

The emergence of predictive justice through AI-driven analysis of extensive datasets, including legal data, has gained considerable attention in recent times. In France, the adoption of the Digital Republic Act and subsequent legal reforms have led to the extensive dissemination of judicial decisions, creating comprehensive judicial databases accessible through open data platforms. However, while these platforms promote transparency and accountability in the justice system, they also present challenges such as limited scope, data format issues, language barriers, and data quality concerns.

Numerous research studies have delved into the ethical and societal implications of AI within the legal system, categorizing their applications into distinct areas. First, AI's role in criminal decision-making, as exemplified by Christin et al. (2015), explores the use of artificial intelligence, specifically predictive algorithms, in criminal justice processes such as sentencing, juvenile justice, and bail decisions. In Russia, Shulayeva et al. (2017) demonstrated the automatic identification of legal facts and principles in sentences through supervised learning, while Metsker et al. (2019) employed machine learning to analyze judgments from Russia's administrative process. Furthermore, Metsker et al. (2021) discussed the creation of a decision support platform for machine learning applications in areas like e-governance and internal policy modeling, drawing from court decisions and administrative data in the Russian Federation.

AI's impact on predicting outcomes has also been a key research focus. Studies regarding the European Court of Human Rights (ECHR) include Aletras et al. (2016), which predicted case outcomes based on textual content from ECHR cases, and Chalkidis et al. (2019), which compared prediction models across

2. https ://opendata.justice-administrative.fr/
3. https ://www.conseil-etat.fr/decisions-de-justice/jurisprudence/rechercher-une-decision-arianeweb

different justice domains. Meanwhile, Medvedeva et al. (2020) developed a system for predicting decision categories associated with ECHR legal judgments.

In the context of the U.S. Supreme Court, Sharma et al. (2015) explored various machine learning techniques to predict case outcomes, achieving high accuracy with Deep Neural Networks. Similarly, Lockard et al. (2023) emphasized the potential of natural language processing and machine learning for predicting U.S. Supreme Court case outcomes, though they noted the necessity for further research before implementing such models in the decision-making process.

Finally, within the French Supreme Court, Sulea et al. (2017) proposed a predictive model capable of determining law categories, court rulings, and the timing of decisions. Their approach, utilizing Bag-of-Words and Support Vector Machines, yielded high F-measures across various prediction tasks. These diverse applications illustrate the multifaceted exploration of AI's impact on the legal system, ranging from decision-making support to outcome prediction across different jurisdictions.

Given the potential advantages and limitations of utilizing open data of judicial decisions, there is a pressing need to explore how AI can play a pivotal role in this context. We aim to identify meaningful machine learning algorithms to exploit these new data sources. More specifically, we illustrate how to use machine learning on a dataset of administrative court decisions. Nevertheless, we recognize that employing AI in this domain requires addressing potential challenges related to data quality, accuracy, and biases. Therefore, our paper is a first step in answering the following question : how can artificial intelligence (AI) be effectively utilized to analyze transcripts of court decisions in order to identify patterns, trends, and pertinent information for legal research and decision-making processes, while addressing challenges related to data quality, accuracy, and potential biases ?

Our ultimate goal is to provide a comprehensive exploration of machine learning's role in analyzing textual legal decisions, with a particular emphasis on its capacity to enhance our understanding of the legal system, thereby promoting transparency and accountability.

In Section 1, we explore Open Data and Decision Data Collection. Section 2 is dedicated to Data Preprocessing and Feature Engineering. In Section 3, we cover Resampling Techniques, Machine Learning Models, and Results

2 Machine Learning-driven Aggregation and Data Collection of Administrative Tribunal Decisions

The legal system is inherently tied to language, making it unsurprising that natural language processing software has long played a role in certain aspects of the legal profession. However, in recent years, there has been a growing fascination with applying modern techniques to a broader spectrum of challenges. In this context, I explore the current applications of natural language processing through the application of machine learning models in the legal sector. Machine Learning (ML) is an algorithm that can learn from experiences to make predictions by training on large datasets. A plethora of ML algorithms have emerged in recent years (Das and Behera, 2017). However, not all of them have gained widespread

recognition. Some failed to address or resolve specific issues, leading to the introduction of alternative algorithms. Machine learning algorithms cover various methods. [4] In this paper, we will only utilize four of them to choose the best when applied to a dataset of french administrative court decisions.

The administrative law of France encompasses the legal framework that governs the organization, operation, and oversight of public administration. A distinctive feature of this legal domain is the presence of several tiers of jurisdiction, providing citizens with the means to challenge administrative decisions. These levels of jurisdiction consist of the administrative tribunal, the administrative court of Appeal, and the Council of State, each playing a specific role in the adjudication of administrative matters.

In the context of court decisions within the realm of administrative law, these documents predominantly manifest as textual records, characterized by a comprehensive and structured analysis. They typically include four essential elements : a statement of facts, which offers a succinct summary of the pertinent case details, encompassing what transpired and the involved parties ; an examination of the legal issues at hand, comprising relevant statutes, regulations, and applicable case law ; a detailed discussion of the legal principles and reasoning employed by the court to arrive at its final decision ; and, finally, the court's ultimate order or judgment. This meticulous analysis of court decisions ensures a comprehensive understanding of the legal processes and outcomes within the realm of administrative law.

In the initial phase of our study, we initiated data collection by acquiring the administrative tribunal decisions from the French Council of State's open data platform. These decisions were obtained in XML format, containing essential information such as the date, involved parties, and legal issues addressed. Each XML file presentsthe blocks described in table 1 (although not all blocks may be present for certain files).

Field	**Description**
Identification	Name of the .xml file
Date-Mise-Jour	Internal date for processing the decision
Code-Juridiction	Name of the jurisdiction. For the Council of State, this value will be "CE"
Nom-Juridiction	Wording of the jurisdiction
Numero-Dossier	Number of the decision or order of the presidents
Type-Decision	Will take the value "decision" or "order"
Type-Recours	Example : "Excess of authority"
Texte-Integral	This tag will contain the body of the decision rendered. Each line is separated by the <p> tag to signal a line break in the original document. The integral text is written in the French language.

TABLE 1 – Blocks description

4. Regression algorithms (e.g., Linear Regression) predict variable correlations. Instance-based algorithms (e.g., K-Nearest Neighbour) use stored data for predictions. Regularization algorithms (e.g., Ridge Regression) counteract overfitting. Decision tree algorithms (e.g., CART) construct trees for decisions. Bayesian algorithms (e.g., Naive Bayes) use Bayes' Theorem. Support Vector Machines (SVM) define decision boundaries. Clustering algorithms (e.g., K-Means) classify data based on patterns. Association rule learning algorithms (e.g., Apriori) uncover correlations. Artificial Neural Networks (ANN) mimic neural networks. Deep learning algorithms leverage abundant data. Dimensionality reduction algorithms (e.g., PCA) reduce datasets. Ensemble algorithms (e.g., GBM) integrate weaker estimators for robustness, showcasing the field's versatility.

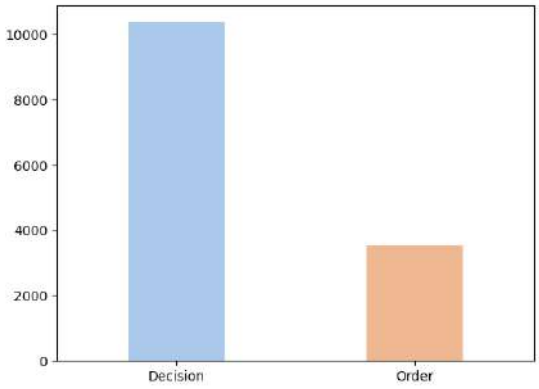

FIGURE 1 – Type of decisions

To perform a more comprehensive analysis, we manually examine legal decisions to ensure that the texts contain all the relevant parts. In our observation, we notice that the text often contains both the decision's components and is divided by terms such as "DECIDE" or "ORDONNE." This enables us to effectively identify and differentiate the various sections within the decision, including the text of the description of the case and the text of the final decision. An uncomplicated extraction algorithm is utilized to gather all files from their individual folders and merge them into a Dataset. [5]

After analyzing the data, we have obtained a distribution of the different types of decisions (figure 1). This variable, "types of decisions," provides us with information about the specific categories or classifications of the decisions present in the dataset. Both "orders" and "decisions" refer to official rulings or outcomes made by a court, tribunal, or administrative body. We have specifically chosen to focus on decisions rather than orders due to the nature of a decision itself. By working with decisions, we can delve into the substantive aspects of the legal process, including the judge's analysis and the final judgment reached.

As our database exclusively consists of decisions, our objective is to divide the text integral of each decision into two parts : the description of the case and the text of the final decision. As previously mentioned, these two sections are typically separated by the keyword "DECIDE". To accomplish this, we have developed an algorithm that extracts the description of the case ('description') from the full text. The decision text includes the text of the solution ('Resultat'), which is identified by utilizing the keyword "DECIDE" as a dividing point.

Our database now includes new variables such as 'Identification,' 'Code-Juridiction,' 'Nom-Juridiction,' 'Type-Decision,' 'Type-Recours,' 'Solution,' 'Description,' and 'Resultat.' These variables provide additional information about the cases, including unique identifiers, codes and names of the jurisdiction, types of decisions, types of recourse, the solution or judgment rendered, the full text of the decision, and the outcome of the case. For the present study, we focus solely on two variables : 'Description' and 'Solution'. We will compare different ML models to predict the outcomes of the 'Solution' variable from the 'Description' variable. The 'Description' refers to the full and unaltered version of a legal document, like a court judgment in this case. It includes all relevant details, provisions, and explanations, leaving nothing

5. GitHublink

out.[6] These outcomes are introduced and represented in the collected data. In table 2, we delve into each term in detail within the "Solution" variable.

Name	Translation	Description
Rejet	Dismissal	Claim or request rejected, no favorable decision.
Satisfaction partielle	Partial satisfaction	Partial grant of claim or request, some relief.
Satisfaction totale	Full satisfaction	Complete grant of claim or request, full relief.
Non-lieu	Discontinuance	Case dismissed due to lack of evidence or legal basis.
Rejet défaut de doute sérieux	Rejection due to lack of serious doubt	Claim rejected for insufficient evidence.
Expertise / Médiation	Expertise / Mediation	Alternative dispute resolution methods.
Désistement	Withdrawal	Voluntary withdrawal of claim, case termination.
Sursis à statuer	Stay of proceedings	Temporary suspension of legal proceedings.
Radiation du registre	Removal from the register	Case removed from court's register.
Renvoi autres juridictions	Referral to other jurisdictions	Transfer of case to another court.
Satisfaction partielle (susp. exécution)	Partial satisfaction with suspension of execution	Partial grant with temporary suspension of execution.
Supplément d'instruction	Additional investigation	Further investigation or evidence gathering
Renvoi au Tribunal des conflits	Referral to the Tribunal of Conflicts	Referral of jurisdictional conflict case to specialized Tribunal of Conflicts.

TABLE 2 – Explanation of the "Solution" Categories

Figure 2 shows that some of the solution categories have only one or two examples, which means they do not represent a substantial group of cases suitable for applying machine learning and extracting meaningful insights from them. Therefore, as an initial step, it is necessary to reduce the number of categories. This step aims to consolidate similar or closely related categories to ensure an adequate representation of cases and facilitate the application of machine learning techniques.

To reduce the number of categories, it is crucial to consult legal professionals and leverage their expertise in regrouping and consolidating similar categories. We can effectively identify and merge relevant categories that exhibit similarities or share common characteristics.

We have grouped the categories in table 3 :

Decisions resulting from expertise or mediation are removed from the database due to their indeterminate nature. Indeed, expertise is an evaluative procedure, and based on its findings, the court will determine the merits of each party's arguments. If an expertise is requested and granted, prevailing on that issue can be seen as a favorable outcome, but it does not guarantee overall success in the case. Similarly, in mediation, the absence of a settlement means the uncertainty of who will prevail in the lawsuit. When a decision is not precisely rendered, it cannot be treated or regarded as a final judgment.

We obtain a database with four categories of final solutions. The histogram 3 illustrates the resulting

6. This comprehensive presentation ensures that everyone involved, including lawyers, judges, legal practitioners, and the public, can access complete and accurate information to interpret and apply the law correctly.

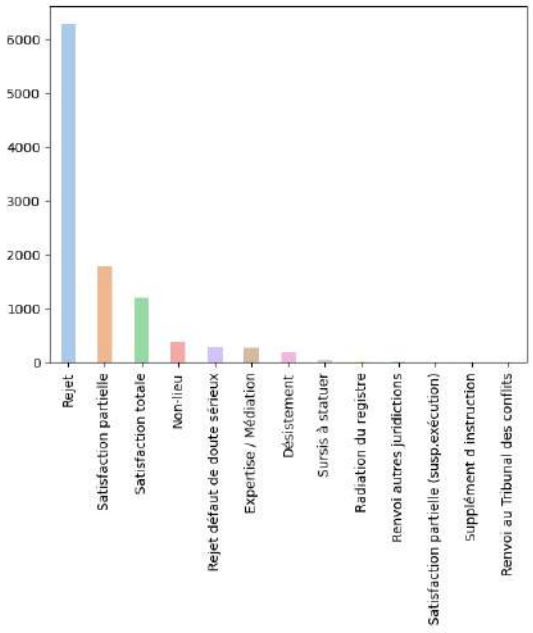

FIGURE 2 – The solution's categories

Name	Description
Rejet	This category includes sub-categories related to cases that have been rejected or dismissed : 'Rejet', 'Rejet défaut de doute sérieux', 'Radiation du registre'.
Satisfaction partielle	This category involves cases where partial satisfaction has been achieved : 'Satisfaction partielle', 'Satisfaction partielle (susp. exécution)'.
Satisfaction totale	'Satisfaction totale'.
Renvoi	This category covers cases that have been referred or transferred to other jurisdictions or authorities for further processing : Non-lieu, Désistement, Sursis à statuer, Renvoi autres juridictions, Supplément d'instruction, Renvoi au Tribunal des conflits.
Expertise / Médiation	This category encompasses cases that involve expertise or mediation as part of the resolution process : 'Expertise / Médiation'.

TABLE 3 – The grouped categories

data. We believe that the regrouped categories accurately represent the nuances and complexities of the legal domain while reducing the overall number of distinct categories. As a result, the number of categories for the final judgment solution has been reduced from 13 to 4.

3 Data preprocessing and Features engineering

3.1 Imbalanced data

The histogram 3 reveals an imbalanced data problem. This means that the distribution of data points across different categories in the histogram is uneven, with some categories having a significantly higher number of instances compared to others. The machine learning community faces a significant challenge when dealing with imbalanced data sets (Japkowicz and Stephen, 2002), where one class (usually the majority class) has a much larger number of examples compared to the others. This problem refers to

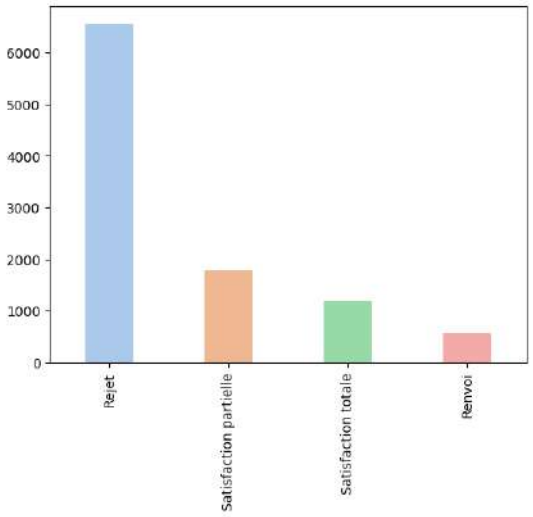

FIGURE 3 – The categories of final outcomes

a situation in a dataset where the distribution of classes or categories is significantly unequal. In other words, one class is represented by a large number of instances, while another class has significantly fewer instances.

This imbalance can occur in various classification problems, where the goal is to predict the class or category of a given sample based on its features. Imbalanced data can pose challenges in machine learning and statistical modeling. It can lead to biased and inaccurate predictions, as models tend to favor the majority class due to its higher representation. This can result in poor performance, low recall, and misleading evaluation metrics.

To address the issue of imbalanced data, it is necessary to apply resampling methods, which will be discussed in the upcoming section. Additionally, we employ a technique called Natural Language Processing (NLP) to analyse the legal data. Performing NLP and resampling methods simultaneously is a common practice to ensure that both the textual information is preserved and the class imbalance is addressed effectively, leading to better machine learning model performance when working with imbalanced text datasets.

3.2 Natural Language Processing (NLP)

NLP is a field of artificial intelligence and computational linguistics that focuses on the interaction between computers and human language. It involves the development of algorithms and models to enable computers to understand, interpret, and generate human language in a meaningful way. NLP encompasses a wide range of tasks and applications, including text classification, sentiment analysis, named entity recognition (NER), machine translation, question-answering systems, and text summarization. [7] These NLP techniques help extract meaning and insights from unstructured text data, enabling applications that involve language understanding, generation, and interaction.

7. Text classification involves categorizing text documents into predefined categories or classes based on their content (Kuriyozov et al., 2023). Sentiment analysis determines the sentiment or opinion expressed in a piece of text, whether it is positive, negative, or neutral(Stine, 2019). NER involves identifying and extracting specific entities from text, such as names of people, organizations, locations, or dates(Mohit, 2014) . Machine translation focuses on automatically translating text from one language to another (Lopez, 2008). Text summarization involves generating concise summaries of longer text documents(Allahyari et al., 2017).

In our study, we will apply text classification techniques to categorize the "description" text based on the labels of the "Solutions" categories.

3.3 Numerical representations and Data preprocessing for administrative text decisions

Once the data collection process is completed, the first step involves obtaining a data representation that can be effectively utilized by the learning algorithm. To achieve this, there are several approaches available that convert a collection of text documents into a numerical dataset. We had to preprocess it to make it usable for our analysis. This included cleaning the data to remove any irrelevant or erroneous information (missing values, outliers...), as well as transforming the data into a format that could be easily analyzed using AI techniques. This data preprocessing step is essential for addressing the imbalanced data problem and facilitating the application of machine learning models.

The initial step, tokenization, involves breaking down the text into units known as tokens, such as words or phrases, to enable efficient processing. Following this, the removal of punctuation, numbers, and unimportant words (stopwords) further simplifies the text for NLP algorithms. Subsequently, feature engineering transforms the raw data into a comprehensive set of features, enhancing machine learning model performance. In this process, Term Frequency-Inverse Document Frequency (TF-IDF) (Jing et al., 2002) is used, a statistical measure that highlights the importance of a term in a document within a large corpus. Ultimately, this transformation expands the data from a single variable to 34,116 features.

3.4 Data Description

In the case of applying NLP, we have chosen to represent the texts on which we work to better visualize them. The figure 4 presents the four-word clouds of each category of the 'solution' variable.

The significant size of the words presented in the four-word clouds, such as 'administrative justice,' 'article,' 'code,' and 'French territory,' signifies that we are analyzing texts related to French administrative decisions, taking into account the laws, codes present in the texts, as well as the procedures. We observe that legal texts generally focus more on residence permits for foreigners and asylum rights. For example, according to the 'Rejet' word cloud, we can analyze that matters related to residence in French territory lead to a rejection that imposes the obligation to leave French territory. Within the texts in the 'Renvoi' category, the reason might be a lack of documentation. The same words appears in the 'Rejet' word cloud but with a different intensity in the 'Renvoi' category.

For both 'satisfaction total' and 'satisfaction partielle' categories, the two-word clouds are similar in terms of keywords, suggesting that the prediction models do not distinguish between texts followed by complete satisfaction and texts followed by partial satisfaction.

((a)) WordCloud of the category Rejet

((b)) WordCloud of the category Renvoi

((c)) WordCloud of the category Satisfaction Partiel

((d)) WordCloud of the category Satisfaction total

FIGURE 4 – Text representing the solution category.

4 Resampling techniques and machine learning models

In this section, we will primarily concentrate on data analysis to identify the most suitable combination of resampling techniques and machine learning models. Our objective is to utilize this combination to effectively predict the final decision outcome. Alternatively, there are two approaches.The first approach is focused on predicting the four categories of the final solution using a single unified machine learning model. In contrast, the second approach involves selecting the initial machine learning model to predict three of the categories, and then employing a different model to predict the remaining two categories.

4.1 First approach :

In this section, we will present a methodology that enables us to select the optimal combination of resampling techniques and machine learning models. This selection aims to predict the four categories of the final solution using a unified machine learning model.

4.1.1 Training Data Re-sampling techniques

Resampling methods are utilized to modify the dataset in order to address the issue of class imbalance. There are two main scenarios commonly employed for this purpose : oversampling, which involves generating additional instances for the minority class, and the hybrid approach, which combines different resampling techniques to achieve a more balanced representation of the classes.

Synthetic Minority Over-sampling Technique (SMOTE) addresses the imbalanced data issue by generating synthetic samples for the minority class (Chawla et al., 2002). It works by selecting a minority class instance and finding its k nearest neighbors. Synthetic samples are then created by interpolating between the feature vectors of the selected instance and its neighbors. This process helps to increase the

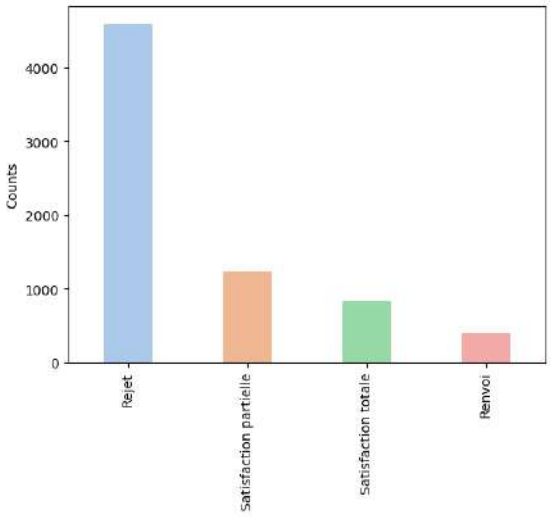

FIGURE 5 – Categories of final solutions to be predict

representation of the minority class in the dataset.

The synthetic samples generated by SMOTE are not duplicates of existing instances but rather new instances that capture the characteristics of the minority class. By introducing these synthetic samples, SMOTE helps to balance the class distribution and improve the performance of machine learning models on imbalanced datasets.

BorderlineSMOTE is proposed by Han et al. (2005) as an extended version of the SMOTE algorithm that specifically tackles the challenge of misclassification near the decision boundary in imbalanced datasets.

While the standard SMOTE algorithm generates synthetic samples by interpolating between existing minority class samples, it can inadvertently introduce noisy or irrelevant synthetic samples. This includes instances that are misclassified near the decision boundary. To address this issue, Borderline SMOTE has been developed, which focuses on generating synthetic samples only for the minority class instances that are located near the decision boundary.

SMOTETomek (A hybrid method) is suggested by Wang et al. (2019) as an effective resampling technique specifically designed to address imbalanced datasets. SMOTETomek combines the advantages of both undersampling and oversampling techniques, aiming to overcome the limitations of SMOTE and Tomek Link [8] methods. To implement SMOTETomek, we utilized the imbalanced-learn library, which provides functions for both SMOTE-based oversampling and Tomek Link-based undersampling.

The SMOTETomek method follows a two-step algorithmic flow to address imbalanced datasets. Firstly, it applies the SMOTE technique to generate synthetic minority samples, resulting in an extended dataset. This step aims to increase the representation of the minority class. Secondly, it utilizes the Tomek Link method to identify and remove Tomek Link pairs from the augmented dataset. By eliminating overlapping instances, the method enhances the separability between classes. The combination of these steps in SMOTETomek provides a resampling approach that effectively tackles class imbalance in datasets.

8. Tomek Link are pairs of data points in imbalanced datasets, one from the majority class and one from the minority class, that are nearest neighbors to each other with no other opposite-class data points in between. Identifying and removing Tomek links can help improve classification performance on imbalanced data by enhancing class separation.

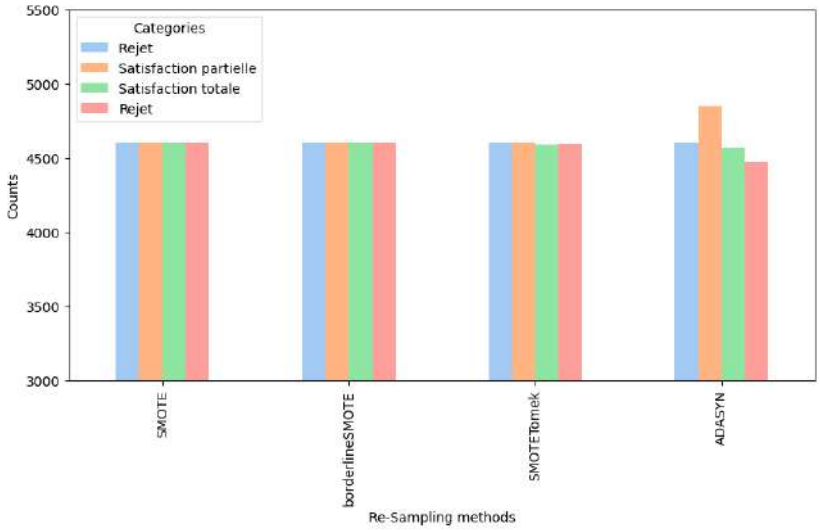

FIGURE 6 – Data augumentation

The oversampling step with SMOTE increases the representation of the minority class, while the under-sampling step with Tomek Link eliminates overlapping instances that may lead to misclassification.

Adaptive synthetic sampling (ADASYN) is an innovative technique introduced by He et al. (2008), designed specifically for addressing imbalanced datasets in machine learning. ADASYN aims to tackle the imbalance issue by generating synthetic data points in the minority class based on their level of difficulty for learning. It assigns higher weights to minority examples that are more challenging to learn, resulting in the generation of additional synthetic data points for these difficult instances compared to the relatively easier ones. This approach offers two benefits :

— it reduces bias stemming from class imbalance,

— it adaptively adjusts the classification decision boundary to better handle difficult examples.

The histogram 6 presents an augmentation of data using the resampling techniques for our four variables. Notable changes can be observed. The histogram displays a more balanced distribution across the four variables, with the bars representing each variable showing similar heights. This indicates that the resampling techniques all adress successfully the initial class imbalance, resulting in a more representative dataset.

4.1.2 Machine Learning Methods and Results

After experimenting with multiple resampling methods, we will utilize a machine learning classifier to determine the optimal combination of the best classifier and resampling technique.

In our case, we will refer to the literature to select the best classifier specifically designed for text analysis (Korde and Mahender, 2012). The selected classifier will be cited as follows :

— DecisionTree (DT) is widely used machine learning technique for both classification and regression tasks. It is a non-parametric supervised learning algorithm that builds a tree-like model to make predictions based on feature values.

— GradientBoostingClassifier (GBC) builds an ensemble of weak prediction models, typically decision trees, and combines them to create a strong predictive model (Friedman, 2001).

— Support Vector Machine (SVM) is particularly effective in cases where the data is linearly separable or can be transformed into a higher-dimensional space where separation is possible. The main objective of SVM is to find the optimal hyperplane that separates data points belonging to different classes with the maximum margin. This hyperplane allows to define a decision boundary that can be used to classify new, unseen data points (Drucker et al., 1996).

— KNeighborsClassifier (KNC) is a simple yet effective supervised learning algorithm. During the training phase, KNC stores the feature vectors and corresponding class labels of the training data. To make predictions, KNC calculates the distances to all training data points and selects the K nearest neighbors. The algorithm then performs a majority vote among these neighbors to determine the predicted class or regression value for the query point. The number of neighbors K , is a crucial hyperparameter that impacts the algorithm's performance.

Evaluation parameters In our study, the categorization of future results is viewed as a multi-classification task. Therefore, evaluation of the accuracy of prediction results involves utilizing performance metrics such as Accuracy, Precision, Recall, and F1-Score. These metrics are calculated using the following formulas :

$$Precision = \frac{TP}{TP+FP}$$
$$recall = \frac{TP}{TP+FN}$$
$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$
$$F1 - score = \frac{2*precision*recall}{precision+recall}$$

Table 4 provides the definitions for TP (true positives), FP (false positives), TN (true negatives), and FN (false negatives). Our results will be evaluated mainly with respect to accuracy and F1-score. Accuracy is the ratio of the number of correct predictions to the total number of predictions. F1-Score, which is a trade-off between Precision and Recall, serves as a reliable measure for assessing classifier performance.

		Actual values	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

TABLE 4 – The confusion table

Discussion of results The table 5 displays the results of the different resampling methods (SMOTE, borderSMOTE, ADASYN, SMOTETomek) evaluated in terms of accuracy and F1-score for the various machine learning models (GBC, DecisionTree, SVM, KNC).

The values shown demonstrate that the performance of the resampling methods varies depending on the selected machine learning models. As an illustration, when considering the GBC model, we notice comparable accuracy results across different techniques like SMOTE, borderSMOTE, and SMOTETomek. All these methods achieve an average accuracy of 81%. However, the GBC model achieves a slightly higher F1-score with SMOTETomek (71.29%).

In our case, we will select the optimal combination based on the best value of the F1 score. This choice is because the F1 score takes into account both precision and recall, making it more suitable for evaluating model performance, especially when dealing with imbalanced datasets whereas accuracy only looks at overall correctness and may be biased towards the majority class. By prioritizing the F1 score, we aim to achieve a balanced and reliable assessment of our model's ability to correctly predict positive instances while minimizing false positives and false negatives.

Consequently, the best combination that yields the highest performance is the Gradient Boosting Classifier with the SMOTE-Tomek technique. This pairing has demonstrated superior results in our evaluation, achieving a balance between precision and recall while effectively handling imbalanced datasets. It is important to note that these results are based on the data provided in the table, and actual performance may vary depending on the specific dataset and other factors related to machine learning.

	SMOTE		borderSMOTE		ADASYN		SMOTETomek	
Metrics	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
GBC	81.19	70.73	**81.33**	71.07	81.09	70.9	81.19	**71.29**
DecisionTree	72.76	61.5	71.22	59.41	72.07	60.06	71.28	59.62
SVM	79.19	65.75	78.66	64.54	79.12	65.53	78.89	65.14
KNC	47.31	42.93	47.64	42.93	44.94	41.04	46.85	42.58

TABLE 5 – The accuracy and F1-score of the resampling methods vary depending on the ML models chosen

Table 6 presents the micro metrics values for each category. [9]

	Precision	Recall	F1-score	Support
Rejet	0.89	0.93	0.91	1963
Satisfaction partielle	**0.69**	**0.51**	**0.59**	545
Satisfaction totale	**0.53**	**0.58**	**0.55**	345
Renvoi	0.74	0.83	0.78	184
accuracy			0.81	3037
macro-avg	0.71	0.77	0.71	3037
weight-avg	0.81	0.81	0.81	3037

TABLE 6 – micro-metrics of GBC Results

We observe that the metric values for the two classes, "Satisfaction totale" and "Satisfaction partielle," are lower compared to the other metric values. This suggests that the GBC model performs better in classifying the "Rejet" and "Renvoi" categories compared to the other categories. To address this issue, we will utilize hyperparameter optimization methods to improve the metric values.

Optimization of hyparameters and Results Hyperparameter optimization is a critical procedure in machine learning that aims to identify the optimal values for the hyperparameters of a model. Hyperparameters are predetermined configuration settings that shape the behavior of the model during the training process and cannot be learned from the data itself. These settings exert a profound influence on the model's performance.

9. Support is the number of cases presented for each category in the test sample.

The primary objective of hyperparameter optimization is to discover the specific combination of hyperparameter values that yield the best model performance, typically measured by metrics like accuracy or error rate. This optimization process entails exploring a predefined space of possible hyperparameter values and evaluating the model's performance for each combination. This evaluation is typically performed using a separate validation set or through cross-validation techniques.

One of the common techniques for hyperparameter optimization is Grid Search. Grid search (Ahmad et al., 2022) involves specifying a grid of possible values for each hyperparameter and evaluating the model's performance for every combination in the grid. It performs an exhaustive search over the entire parameter space.

In our study, we utilized grid search with cross-validation to determine the optimal parameters for the gradient-boosting classifier. Subsequently, we applied these optimized parameters, and the table 7 below illustrates the updated results.

	Precision	Recall	F1-score	support
Rejet	0.88	0.97	0.92	1963
Satisfaction partielle	0.69	0.61	0.65	545
Satisfaction totale	0.67	0.45	0.54	345
Renvoi	0.84	0.73	0.78	184
accuracy			0.83	3037
macro-avg	0.77	0.69	0.72	3037
weight- avg	0.82	0.83	0.82	3037

TABLE 7 – micro-metrics of GBC Results using Gradient searchCV

The table 7 shows that the recall value for "Satisfaction partielle" increased from 0.51 to 0.61, resulting in an improved F1 score of 0.59 to 0.65. Additionally, the precision value for the "Satisfaction totale" increased from 0.53 to 0.67. However, there was a decrease in the recall value from 0.58 to 0.45, which caused a slight decline in the F1 score from 0.55 to 0.54.

In conclusion, the utilization of the hyperparameter improvement technique proves to be ineffective for both the "partial satisfaction" and "total satisfaction" categories. Therefore, our next course of action involves modifying the prediction methodology for the final decision. The following section presents an alternative solution to achieve optimal metrics for predicting the final solution.

4.2 Second approach

To address the issue of poor metric values for the two categories, namely total satisfaction and partial satisfaction, we have decided to adopt a different approach. Initially, we will merge these two categories into a single category called "satisfaction." The entire process that was previously implemented will be applied again to select the best machine learning model for predicting these three categories. In the next step, we will proceed to train another dataset consisting only of the "satisfaction total" and "satisfaction partielle" categories. This additional dataset will help us identify the most suitable model for accurately predicting these two specific categories. In order to generalize the process, once our initial model predicts satisfaction, we employ additional machine learning techniques to ascertain whether the satisfaction

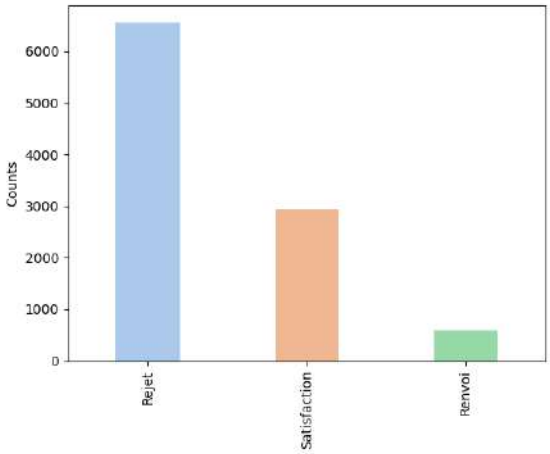

FIGURE 7 – New solutions categories

is partial or total. This modification was put in place to improve the training efficiency of the machine learning model. The resulting histogramme 7 reflects the modifications made to the category distribution and indicates that the problem of imbalanced data persists.

In order to implement this method, we applied the same preprocessing techniques. Additionally, we utilized identical resampling methods and a consistent machine learning model to address the issue of imbalanced data. The table 8 shows that the Gradient Boosting Classifier achieved the best metric values with different resampling methods. The highest accuracy score of 88.08% was obtained with the SMOTETomek resampling technique, while the best F1 score of 84.03% was achieved with the ADASYN resampling method.

In our scenario, where we are primarily interested in assessing the performance of the minority class, our emphasis lies on the F1 score. The F1 score, which is the harmonic mean of precision and recall, offers a well-rounded evaluation of the classifier's effectiveness, especially in the context of imbalanced datasets.

To determine the optimal combination of classifier methods and resampling techniques for our research, we will give precedence to achieving the highest F1 score. Notably, the ADASYN resampling method exhibited the highest F1 score, establishing it as the preferred choice for our analysis.

	SMOTE		borderlineSMOTE		ADASYN		SMOTETomek	
Metrics	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
GBC	87.84	83.87	87.81	83.77	87.94	**84.03**	**88.08**	83.79
DecisionTree	80.14	74.64	80.24	72.58	79.22	70.97	80.24	74.14
SVM	84.62	76.84	83.96	76.26	84.35	76.61	84.03	76.14
KNC	56.10	49.60	55.31	49.56	53.86	49.18	55.61	50.42

TABLE 8 – Variability in Accuracy and F1-Score of Resampling Methods Based on Chosen ML Models

After selecting the most suitable model for predicting the three different categories, we will assess their micro metrics to evaluate their predictive ability. The table 9 displays the micro metrics, which exhibit very high values even when default hyperparameters are utilized.

To summarize, the GradientBoosting classifier stands out as the superior machine learning model for successfully predicting the three categories of the final solution. This conclusion is drawn from the

	Precision	Recall	F1-score	support
Rejet	0.91	0.92	0.92	1963
Satisfaction	0.84	0.79	0.81	890
Renvoi	0.75	0.84	0.79	184
accuracy			0.88	3037
macro-avg	0.83	0.85	0.84	3037
weight- avg	0.88	0.88	0.88	3037

TABLE 9 – micro- metrics of GBC Results using the ADASYN resampling methods

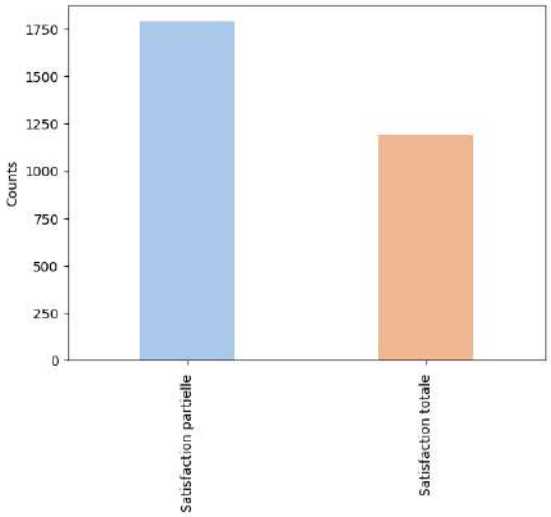

FIGURE 8 – Satisfaction categories

utilization of the ADASYN resampling method on the dataset, which further enhances the classifier's performance.

In the subsequent step, we will select the most suitable model for training, which includes only two categories : partial satisfaction and total satisfaction, as depicted in histogram 8.

The histogram 8 indicates that the training data is still affected by the issue of imbalanced data. Therefore, we proceed similarly to identify the optimal combination of resampling techniques and machine learning models. The outcomes are presented in the table 10. Subsequently, we will choose the most suitable model for fitting the data, which comprises solely two categories : partial satisfaction and total satisfaction, as illustrated in the table 10.

	SMOTE		borderlineSMOTE		ADASYN		SMOTETomek	
Metrics	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
GBC	68.42	67.41	67.30	60.44	68.42	67.61	68.42	67.54
DecisionTree	64.38	62.79	64.16	62.86	63.38	62.07	64.27	62.83
SVM	71.66	69.61	71.55	69.55	**71.89**	**70.21**	71.55	69.73
KNC	61.14	60.79	60.24	59.99	59.57	59.27	60.35	60.12

TABLE 10 – The Accuracy and F1-Score Variability of Resampling Methods Across Different ML Models

The table 10 demonstrates that ADASYN appears to be particularly effective for the SVM model, while the other resampling methods yield similar results for the GBC and DecisionTree models. The KNC model generally has lower performance regardless of the resampling method used.

The table 11 provides a comprehensive evaluation of the SVM model's performance. These metrics, especially the F1-score for 'Satisfaction totale,' may be considered relatively low values for evaluation,

even though we have noted improvements in the chosen model's evaluation metrics. It is worth mentioning that the best models for each step in the second approach are different than the best one in the first approach. Consequently, as a next step, we may consider applying another machine learning models or even explore deep learning models in an attempt to achieve better, or potentially worse, results.

	Precision	Recall	F1-score	support
Satisfaction partielle	0.77	0.69	0.73	544
Satisfaction totale	0.65	0.62	0.63	349
accuracy			0.72	893
macro-avg	0.70	0.70	0.70	893
weight- avg	0.72	0.72	0.72	893

TABLE 11 – micro- metrics of SVM Results using the ADASYN resampling methods

5 Conclusion

In conclusion, this research has highlighted the growing interest in Machine Learning and the potential benefits it offers for legal research and decision-making. Throughout this effort, we are attentive to minimizing biases and ensuring the accuracy and quality of the data.

Increasing availability of large database of court decisions, like France's commitment to making judicial decisions accessible as open data has paved the way for extensive exploration of this field. However, predicting court decisions is not without its challenges, including data limitations, format issues, language barriers, and data quality concerns. Addressing these challenges is crucial for realizing the full potential of AI in this domain.

One notable limitation that researchers may face is the lack of insight into the decision-making process of the models at a certain point in our analysis. For example, in our application developed on administrative court decisions, we saw that our models are proficient at making decisions, we encounter a challenge in comprehending the underlying reasons for these decisions. This limitation brings to light the 'black box' nature of some machine learning algorithms, where the internal mechanisms remain elusive, and the decision outputs may seem like a 'mystery.' Without a clear understanding of why certain decisions are reached, it becomes challenging to provide comprehensive explanations or insights based solely on the model's output. As a result, we must acknowledge this limitation in our study, as it underscores the need for more transparent and interpretable models, which would not only enhance our understanding but also improve the credibility and utility of our findings.

Bibliographie

G. N. Ahmad, H. Fatima, S. Ullah, A. S. Saidi, et al. Efficient medical diagnosis of human heart diseases using machine learning techniques with and without gridsearchcv. *IEEE Access*, 10 :80151–80173, 2022.

N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos. Predicting judicial decisions of the

european court of human rights : A natural language processing perspective. *PeerJ Computer Science*, 2 :e93, 2016.

M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut. Text summarization techniques : a brief survey. *arXiv preprint arXiv :1707.02268*, 2017.

S. A. Bini. Artificial intelligence, machine learning, deep learning, and cognitive computing : What do these terms mean and how will they impact health care ? *The Journal of Arthroplasty*, 33(8) : 2358–2361, 2018. ISSN 0883-5403. doi : https://doi.org/10.1016/j.arth.2018.02.067. URL https://www.sciencedirect.com/science/article/pii/S0883540318302158.

I. Chalkidis, I. Androutsopoulos, and N. Aletras. Neural legal judgment prediction in english. *arXiv preprint arXiv :1906.02059*, 2019.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.

A. Christin, A. Rosenblat, and D. Boyd. Courts and predictive algorithms. *Data & CivilRight*, 2015.

L. Cluzel-Métayer. Les limites de l'open data. *L'Actualité juridique. Droit administratif*, (2), 2016.

R. DALE. Law and word order : Nlp in legal tech. *Natural Language Engineering*, 25(1) :211–217, 2019. doi : 10.1017/S1351324918000475.

K. Das and R. N. Behera. A survey on machine learning : concept, algorithms and applications. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(2) :1301–1309, 2017.

H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9, 1996.

J. H. Friedman. Greedy function approximation : a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

F. G'sell. Les progrès à petits pas de la «justice prédictive» en france. In *ERA Forum*, volume 21, pages 299–310. Springer, 2020.

H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote : a new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing : International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1*, pages 878–887. Springer, 2005.

H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn : Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

N. Japkowicz and S. Stephen. The class imbalance problem : A systematic study. *Intelligent data analysis*, 6(5) :429–449, 2002.

L.-P. Jing, H.-K. Huang, and H.-B. Shi. Improved feature selection approach tfidf in text mining. In *Proceedings. International Conference on Machine Learning and Cybernetics*, volume 2, pages 944–946. IEEE, 2002.

N. Kolt. Predicting consumer contracts. *Berkeley Tech. LJ*, 37 :71, 2022.

V. Korde and C. N. Mahender. Text classification and classifiers : A survey. *International Journal of Artificial Intelligence & Applications*, 3(2) :85, 2012.

E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov. Text classification dataset and analysis for uzbek language. *arXiv preprint arXiv :2302.14494*, 2023.

Y. Li. Algorithmic discrimination in the us justice system : A quantitative assessment of racial and gender bias encoded in the data analytics model of the correctional offender management profiling for alternative sanctions (compas). 2017.

K. Lockard, R. Slater, and B. Sucrese. Using nlp to model us supreme court cases. *SMU Data Science Review*, 7(1) :4, 2023.

A. Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3) :1–49, 2008.

M. Medvedeva, M. Vols, and M. Wieling. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28 :237–266, 2020.

O. Metsker, E. Trofimov, S. Sikorsky, and S. Kovalchuk. Text and data mining techniques in judgment open data analysis for administrative practice control. In *Electronic Governance and Open Society : Challenges in Eurasia : 5th International Conference, EGOSE 2018, St. Petersburg, Russia, November 14-16, 2018, Revised Selected Papers 5*, pages 169–180. Springer, 2019.

O. Metsker, E. Trofimov, and G. Kopanitsa. Application of machine learning for e-justice. In *Journal of Physics : Conference Series*, volume 1828, page 012006. IOP Publishing, 2021.

B. Mohit. Named entity recognition. *Natural language processing of semitic languages*, pages 221–245, 2014.

A. Shapiro. Reform predictive policing. *Nature*, 541(7638) :458–460, 2017.

R. D. Sharma, S. Mittal, S. Tripathi, and S. Acharya. Using modern neural networks to predict the decisions of supreme court of the united states with state-of-the-art accuracy. In *Neural Information Processing : 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings, Part II 22*, pages 475–483. Springer, 2015.

O. Shulayeva, A. Siddharthan, and A. Wyner. Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law*, 25(1) :107–126, 2017.

R. A. Stine. Sentiment analysis. *Annual review of statistics and its application*, 6 :287–308, 2019.

O.-M. Sulea, M. Zampieri, M. Vela, and J. Van Genabith. Predicting the law area and decisions of french supreme court cases. *arXiv preprint arXiv :1708.01681*, 2017.

Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang. Smotetomek-based resampling for personality recognition. *Ieee Access*, 7 :129678–129689, 2019.