www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(2), April 2025, 60-70. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2025.v16i2.1074

Enhancing Explainability, Privacy, and Fairness in Recidivism Prediction through Local LLMs and Synthetic data

Ma. Angelina Alarcón Romero[1], José Antonio Orizaga Trejo[1], Daniel Hernández Mota[2], Luis Fernando Baltazar Villalpando[1], Ma. Hidalia Cruz Herrera[1]

[1] Universidad de Guadalajara, Guadalajara, México.
[2] Instituto Tecnológico y de Estudios Superiores de Occidente, Tlaquepaque, México

E-mails: angelina.alarcon10061@alumnos.udg.mx, jose.orizaga@academicos.udg.mx, daniel.hernandezm@iteso.mx, luis.baltazar7491@alumnos.udg.mx, hidalia.cruz@academicos.udg.mx

Abstract. Predictive policing is considered a high-stake context, where the main challenges in employing an AI solution are to ensure the privacy and fairness of the system while preserving high performance. This usually implies specific demands on the technologies used and their explainability. To alleviate the emerging impediments to adopt a recidivism model, this study exploresan approach employing synthetic data in combination with state-of-the-art NLP techniques, such as transformers-based models running locally. This approach enhances the representation of crimes while preserving data privacy. In particular, we focus on comparing several language models for multilabel classification in Spanish language and techniques such as data reduction, data augmentation and in-distribution validation. The resulting methodology shows the benefits and drawbacks of selecting each language model and highlights the ability of identify and alleviate populations where the model performs significantly worse than the average.

Keywords: Recidivism prediction, Explainable Artificial Intelligence, Interpretable machine learning, Trustworthy AI, Large language models, Data Quality, Data Imbalance.

Article Info
Received Mar 2, 2025
Accepted Mar 20, 2025

1 Introduction

Modeling criminal behavior to predict detainee actions has become essential for government decision-making, as structured risk assessment approaches have been demonstrated to enhance safety outcomes [1]. Machine learning techniques are particularly valuable among these structured methods due to their vast capabilities. However, their adoption in the government sector lags primarily do to a lack of citizen trust in meeting demands for fairness, transparency and privacy [2]. This study emphasizes the importance of recidivism prediction and advocates for increased interpretability and robustness of machine learning models, especially in areas such as social security.

Criminal activity is a complex and multifaceted problem that arises from intricate interplays between several factors encompassing individual, social, economic and environmental dimensions [3]. Therefore, in compliance with the sparsity of infrastructure in México for adequately monitoring and modeling criminal behavior, we propose a system for predicting recidivism based on demographic data from INEGI census [4], in conjunction with police records from Automated Fingerprint Identification System Jalisco (AFIS).

In previous research, we explored the capacities of using a model-centric approach in predicting recidivism for robbery crimes by applying intrinsic eXplainable Artificial Intelligence (XAI) algorithms [5]. We found that this approach made the model's decision process more transparent, facilitating explanation and accountability. Nevertheless, we acknowledged two key limitations: the need for a more comprehensive representation of criminal activities and the inherent challenge that a model's performance is constrained by the quality of its training data, which in terms of fairness and accountability, it implies that the model often learns and replicates the bias present in the data. In contrast, in this work, we expand the system’s capabilities, aiming to enhance robustness through better representation and selection of the dataset.

The available data used to train the model comes from AFIS Jalisco system, which consists of two main components: demographic data of the detainee and a natural language description. The latter effectively requires the extraction of valuable categories, essentially framing the task as a multilabel classification problem. This task is further complicated by domain-specific vocabulary and several aphorisms, for which advanced natural language processing (NLP) techniques are indispensable. However, challenges arise due to two substantial factors: the limited adoption of modern NLP techniques for languages other than English, and the technical constraints of deploying models privately [6,7]. Given these limitations, this research explores the capacities of models that can operate locally on servers with modest resources, aiming to achieve an optimal solution within these constraints.

Data-centric AI is an emerging discipline that can be described as the systematic framework to develop, iterate, and maintain high-quality data in each step of the lifecycle of a machine learning system [8]. In other words, data-centric AI provides techniques to assess problems in the data used for training, validation, and maintenance. We advocate that the use of AI in government and social contexts, due to fairness requirements, demands constant revisions. Therefore, confronting these concerns with the use of data-centric techniques may result in an adequate approach for the system development. We decided to integrate the crime classification task with the prediction model to address the problem holistically and identify potential errors at any stage of the development pipeline. This approach is expected to ensure the overall trustworthiness of the model.

We adopted classification categories inspired by the penal code for both the country and the state, making changes only based on the over-representation of certain crimes during manual labeling. The resulting dataset is used for fine-tunning and evaluating the foundational models. After the classification of crimes, the records are mapped to the sociodemographic data and used for training the recidivism prediction model, using iteratively data-centric techniques such as data augmentation, data reduction and in-distribution evaluation. By applying these techniques, we addressed different issues: alleviated imbalance, enhanced performance on specific sub-populations, and avoided biases; overall, obtaining a more robust system.

This article begins by providing the context of the problem and introducing the proposed solution in Section 2. In Section 3, we define the techniques used in our experiments, with particular emphasis on the constraints for deploying language models locally and an in-depth description of data-centric techniques. We present our results in Section 4, followed by a brief discussion in Section 5. Finally, section 6 offers concluding remarks on our findings and their implications.

2 Case Study Application

This section describes the context in which the problem and the proposed solution are developed emphasizing the social, ethical and legal aspects. Technical content will follow in Section 3.

2.1 Context of the solution

Within social security, there is a demand for advanced systems that efficiently record and track the detainees' criminal behavior. For instance, in Jalisco, México, the AFIS system has been used since 2010 and is deemed suitable for documenting detention procedures. Yet, a void remains in detainee monitoring, largely contingent on the effectiveness of the public policies instituted and assessed.

In the context of criminal conduct in Jalisco, the AFIS system registers the detainee' personal information and a description of the committed crime. However, after mapping the address of residency, it becomes clear that the number of criminal records is not uniformly distributed, as can be seen in Fig. 1. This observation leads to the hypothesis that certain groups of population under specific circumstances are more prone to reoffend.

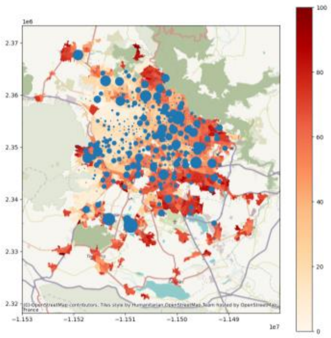

Fig. 1. The problem of recidivism prediction involves estimating how prone a detainee is to reoffend by considering the overlap of their demographic circumstances and criminal profile. The figure shows, as an example, the relationship between crimes against property and the scaled amount of population with incomplete basic education. (Image elaborated with own data using the Overpass API [10]).

This research delves into the potential to forecast a detainee's recidivism for such offenses, aiming to establish a rule set for the prediction process that elucidates the socio-demographic circumstances that endorses the problem. As suggested by existing literature, economic conditions and indicators of underdevelopment are anticipated to be particularly relevant for this endeavor [9]. This strategy helps in formulating insights that facilitate understanding, monitoring, and ultimately, intervening to address factors that adversely impact criminal behavior.

The initial record of criminal conduct is made by using the AFIS system at the time of detention. This implies that the reports are usually disorganized and untidy as they typically comprise testimonials from police officers or detainees. Despite this, these registers represent a valuable source of information; when properly processed, they offer more comprehensive and readily available data, compared to subsequent records. Therefore, finding proper methods to extract valuable information from these registers and convert it into predictions represents an essential step in adequately monitoring criminal conduct, potentially leading to results that satisfactorily model the problem of recidivism. Each register contains details about the crime, personal information and biometrics identifiers.

The proposed system is summarized on Fig. 2, which illustrates an interative process applying techniques for data selection, explainability and accountability to achieve a consensus on the mechanisms to assist and, where necessary, intervene to adjust factors shaping criminal behavior. In specific, the methodology is designed to achieve improvements by enhancing data representation using sampling techniques and synthetic data.Nonetheless, despite the tools used, any technical solution in social contexts must avoid the illusion of objectivity [11] and adhere to the ethical standards by having at its core a human-centric design.

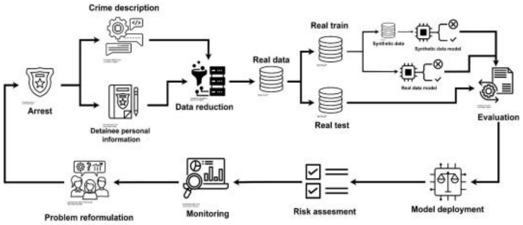

Fig. 2. Framework for developing the specific system for recidivism prediction with an emphasis on fairness and privacy, using sampling techniques and synthetic data. (Elaborated with resources from public domain and CC BY 3.0 via Noun Project [12])

2.2 Data Description

First, it is pertinent to mention that the AFIS system operates across the entire territory of Jalisco. However, due to the sparsity of the data, the study is conducted only in the municipalities of the main city, henceforth referred to as ZMG.

The nature of recidivism is inherently less common compared to single incidents, so it is expected to have unbalanced classes. However, this is amplified by data issues derived from situations such as intermittences in the service, poor planning of upgrades and several migrations of the database system. From a total of 618,332 records comprising administrative misconduct and crimes across the entire state territory between June 2014 and September 2022, only a reduced number of them were selected. These selected records are from the ZMG area and are cases where it was possible to ensure they were actual recidivism events for crimes punishable by imprisonment.

Lacking pre-established classes for crime records, these are made in natural language, which, while capturing more context, obstructs easy development for cluster and analyze similar events. To classify the crimes descriptions, we adopt a classification inspired by the penal code for both the country [13] and the state [14], making distinctions only based on the over representation of certain crimes during manual labeling, the resulting dataset is used for fine-tunning and evaluating the foundational models. After the classification of crimes, the records are mapped to sociodemographic data and used for training the recidivism prediction model. We iteratively apply data-centric techniques such as data augmentation, data reduction and in-distribution evaluation, alleviating imbalance issues and enhancing performance on specific sub-populations.

2.3 Fairness and Bias in machine learning

Fairness in machine learning is a complex topic that encompasses multi-faceted sociocultural concepts such as equality, diversity and ethical principles [15]. This multidisciplinary nature makes it a concept difficult to define and, more importantly, to apply on technical solutions. However, it has a great importance as it is recognized as a key component to increase the trustworthiness of ML [16]. Fairness in the context of machine learning can be defined in several ways; in section 3.5, we present the statistical definition adopted and the metrics to estimate it. However, intuitively it can be explained as the property of a model to work similarly well for all types of records, which in the context of this problem implies producing adequate outcomes for all types of people.

The concept of fairness is closely related to the term bias, which as well has several interpretations. For this study, we consider bias as systematic errors in prediction derived from data. It is pertinent to highlight that a model trained on a biased dataset can learn those biases and reproduce them in its predictions [17].

The dataset of crimes records presents an imbalance among certain subpopulations, as some crimes and groups are much common than others. Therefore, it is expected that the model may skew towards prioritizing such attributes over the rest of the categories, creating an unfavorable situation for the underrepresented class.

The underlying hypothesis is that subpopulations with few data points are expected to perform poorly and, consequently, are more prone to present bias as well. This is because the probability of misclassification is greater for these subpopulations, as is the probability of giving distinct results to similar instances.

2.4 Privacy in machine learning

The development of ML solution on socio-technical systems can present not only fairness concerns but also privacy issues that need to be assessed as well. The need for privacy-preserving techniques, especially in socio-technical systems, arises from the fact that these are usually built in contexts where the required dataset includes sensitive data that is essential for modeling the problem. These solutions are intended to be openly deployed without publicly disclosing sensitive information.

The problem of safeguarding sensitive data is common across many technological solutions. Nevertheless, the problems inherent to machine learning encompass specific kinds of attacks and adversarial goals. To clarify these concepts, we have encapsulated the main kind of attacks in three categories [18]: inference about members of the population, inference about members of the training dataset, and inferring model parameters.

The inference of members of the population refers to the goal of extracting statistical information about the problem just by inferencing the model. This could include inferring sensitive attributes to make representative classes of the population. Regarding model parameters attacks, these target systems whose model's secrecy is critical to its utility, such as spam or fraud detection, where the goal is to extract an equivalent model to the original or create forgery copies.

However, the more concerning problem in our study is the inference about members of the training dataset, where the focus is on protecting the identity of individuals whose personal information was used to train the model. This type of attack can be divided in two subgoals: Membership inference, which refers to determining if a certain point was used to train the model, and property inference, where the goal is to infer if a certain property is true for a subset of the training set.

Ensuring the privacy of the subjects for the proposed model implies that even when part of the data used is publicly available (such as census data), in combination with data obtained for inference, it is not enough to reconstruct the information of specific individuals. Recent works use synthetic data [19] as an anonymization technique, where it is relevant to mention that a common challenge in privacy-preserving machine learning is the tradeoff between privacy and utility. Synthetic data is expected to present this same challenge. Therefore, the ultimate goal consists in designing systems using synthetic data that archive an adequate level of privacy without sacrificing performance.

3 Methodology

A The system as established, requires two main models: first a NLP model for multilabel classification of crimes, and second, a model for recidivism prediction. Taking this into account, for the methodology we start by enumerating the families of languages models considered for the experiments, follow by the techniques intended to enhance the prediction model. Regarding crime classification, we mainly want to compare if LLMs, which are considered to have better capacities than BERT models, can indeed manage more complex crimes descriptions, given the restrictions of having to use a local model and operating in the Spanish language. For the recidivism prediction model, the methods are chosen with the purpose of identifying, evaluating and mitigating concerns related to fairness and privacy.

The data-centric approach is significantly less used compared to model-centric approach, resulting in the maturity of workflows and common practices that are not as mature. Therefore, we propose a combination of techniques to ensemble a workflow that has explainability, fairness and privacy at its core and is advisable, for better comprehension, to identify each method on the corresponding flow on Fig 2. In Sections 3.4, 3.5 and 3.6 the evaluation metrics are introduced followed by the methodology employed in the experiments in Section 3.7.

3.1 Transformer based language models

Natural Language Models (NLMs) have evolved through several stages [20]. They began as rule-based models, progressed to statistical models and then adopted neural networks approaches. However, the field experienced significant advances with the introduction of architecture-based models, especially those employing the Transformer architecture. This led to the rise of pre-trained language models (PLMs), that, unlike previous generations of models, are task-agnostic. This means that they are pre-trained on a large volume of data for general tasks and then, on a process called fine-tuning, they are tuned over considerably smaller amounts of data to improve their performance on specific tasks.

The Transformers architecture introduce in [21] and shown in Fig. 3, marked a new era of NLMs by introducing the concept of attention. Before this, the most adopted approaches consisted on recurrent neural networks (RNN) to represent the language as a succession of words. However, RNN present the problem of vanishing and exploiting gradients, failing to adequately represent the context of a phrase. Instead, attention represents each word in context by defining a mapping between query (Q) and a set of keys (K) to a value (V), where the three variables are vectors. The attention is calculated by computing the dot product between the queries and the keys, scaled and applying a softmax function to convert the result into a distribution of probabilities, which correspond to the weights of the values.

Although all Transformers-based models have the concept of attention as a cornerstone, each approach uses it differently. For a rough classification, the approaches for Transformers-based Language Models can be categorized into three main categories: encoder-only, decoder-only, and encoder-decoder models. The models employed in this study belong to the former two categories and are described in the subsequent sections.

Additionally, it should be noted that while larger models generally have more capabilities and perform better than smaller ones, this comes at a significant computational cost, and for certain tasks, fine-tuned smaller models can still outperform their larger counterparts [22], for smaller models fine-tuned for the specific task. Therefore, the choice of model can significantly impact the resources required.

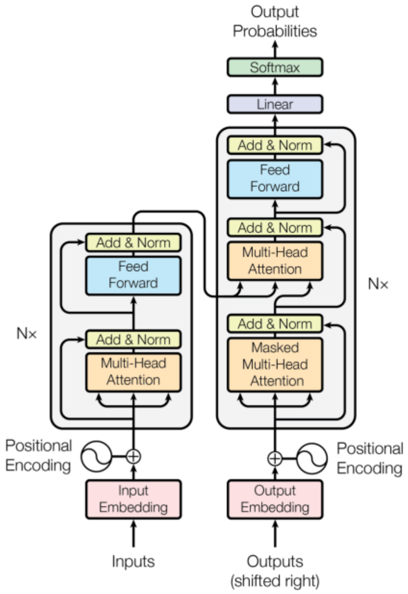

Fig. 3. Transformers architecture, from which BERT and GPT models are based on, BERT architecture uses an encoder only architecture corresponding to the part of the architecture on the left while GPT models employ decoder only architectures like the part on the right. Courtesy of [21].

3.1.1. BERT models

BERT, which stands for Bidirectional Encoder Representations from Transformers [23], is an encoder-only language model. It is distinguished by using a bidirectional approach for its attention mechanism, considering both the preceding and following context for each word in a sentence. An illustration of a bidirectional attention mechanism is shown in Figure 4. This method allows BERT to capture richer and more nuanced representations of words by considering their full context.

Since the launch of the original paper, several variants of the BERT architecture and fine-tuning dataset for diverse tasks have been released. Specifically, in the open-source community, the platform Hugging Face hosts models for each of the main NLP tasks. For this study two models based on BERT architecture are considered: the first [24] fine-tuned for Spanish language and the second [25], trained from scratch using the variant architecture RoBERTa.

Both models are open-source and by using the Transformers library, developed by the Hugging face community, models with architectures based on BERT can be fine-tuned for specific tasks. In our case this involves multi-label classification using a corpus of crimes manually labeled.

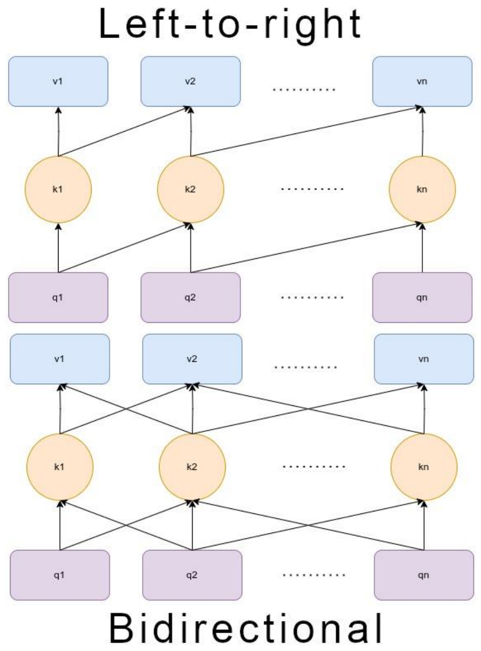

Fig. 3. Representation of the attention mechanism in the approach of each architecture, BERT models use a bidirectional architecture while GPT models uses a Left-to-right architecture. The nodes represent the components of the attention, Query (q), Key (k) and Value (v) embeddings of each token.

3.1.2. Large Language Models

The term Large Language Models (LLMs) refers to models with tens to hundreds of billions (B) of parameters, and most of them consist of decoder-only models, which gained popularity after the launching of the model ChatGPT in 2022.

LLMs, compared to the other PLMs, are much larger models that not only are task-agnostic but also exhibit stronger language capabilities and emergent abilities [26]. The interest in using this kind of models, even though they are not usually used for traditional NLP tasks, arises from the nature of the crime corpus considered in the dataset. Given that it consists of very messy data with several aphorisms, it is believed that these models could better understand this type of text compared to encoder-only models, particularly in the Spanish language.

For this study, we have special interest in models that can be hosted locally. Therefore, the models must be open source, have a permissive license and be small enough. Consequently, even if some models from OpenAI, Google, or Anthropic, such as ChatGPT, Gemini, or Claude, may be strong candidates, they were no considered because they can only be accessed via API calls. Instead, a selection of small, multilingual models, trained in high quality data were chosen.

Llama [27] is a family of models developed by Meta. As of now, Llama 3 is the latest version and is available in two sizes, 8B and 70B parameters.I It is optimized for dialogue use cases.

Mistral [28] is a model with 7B parameters designed to achieve superior performance and efficiency. It employs a sliding window attention (SWA) mechanism that allows it to handle sequences of arbitrary length, and a grouped-query attention (GQA) that grants faster inference.

Phi 3 [29] is a family of models developed by Microsoft, available in three sizes ranging from 3.8B to 14B parameters, along with a multimodal version. These models were developed focusing on being lightweight, having a high performance and dense reasoning, using synthetic data and filtered public datasets.

3.1.3. Fine-tuning and Quantization

Fine-tuning is the process of training a PLM on new data with the purpose to adapt it to specific tasks. In fact, in the early stages of PLM models, such as BERT, fine-tuning was always required for them to be useful, as they were not able to perform specific tasks out of the box. In contrast, LLMs can perform specific tasks using zero-shot learning. However, fine-tuning can still be beneficial to improve the performance on a specific task.

To fine-tune the model for solving the problem of multilabel classification for both encoder-only and decoder-only models, the process consists mainly on adapting an additional layer specific to the task. This implementation was carried out using the environment of libraries from HuggingFace, and the problem was addressed using the Sequence Classification classes.

The finetuning process for all models was carried out using a single GPU device employing an NVIDIA GeForce RTX 3090 graphic card. However, since training a LM usually requires more memory than what is required for inference, the amount of memory needed to fine-tune a LLM exceeds the available VRAM of this card. Therefore, the QLoRA method was employed. Highlighting that for final inference the LLM models can be quantized and converted for running on a CPU instead of requiring strictly GPU hardware.

3.2 Data centric techniques

The data-centric approach in AI represents a paradigm shift from the traditional model-centric approach, which primarily focuses on developing more effective models keeping the dataset unchanged. Instead, data-centric AI emphasizes the importance of data itself, ensuring it is well-suited for the tasks and methods at each stage of the machine learning pipeline [8]. Adopting a data-centric approach offers several advantages over a model-centric one particularly in terms of fairness. For example, in the model-centric approach, data quality is sometimes underestimated, which can lead to negative effects in the model, such as a deterioration on the model performance or persistent biases. However, it is important to clarify that both approaches complement each other, and the proper use of both can significantly assist in achieving the fairness objectives.

Regarding data-centric methods, it is relevant to mention that the process of crime classification already contributes to improving data quality. This is because data cleaning and preparation usually consume most of the development time.

The techniques in data-centric AI can be categorized by the three main goals in the life cycle of a machine learning system: Training, Evaluation and Maintenance. For the purposes of this investigation, we focus on training and evaluating techniques.

3.2.1. Reduction and Augmentation

Machine learning revolves around data, making it essential to develop techniques that create datasets effectively and efficiently, encoding the reality of the problem. This process typically involves collecting the necessary data, labeling it, and preparing it for the algorithm to learn from. However, techniques such as data reduction and augmentation can be utilized to alleviate several concerns within the dataset that might otherwise propagate to the model.

Starting with reduction this technique aims to decrease the complexity of the dataset while preserving its essential information. Reducing the complexity can help to lower the computational resources required and improve interpretability. Additionally, reducing the amount of redundant data can help to avoid overfitting on noise and focus instead on the relevant information enhancing performance, efficiency and interpretability. A common technique for this purpose is by simply under sampling the majority class to alleviate class imbalance, for example, in [30] the authors employ random under sampling in several ratios for class imbalance in tweets classification.

Augmentation is the process of manipulating existing data to generate variations or synthesize new data. This process is useful both for unbalanced problems and general data insufficiency. Specifically for tackling unbalanced issues, a popular approach is SMOTE [31] and its variants, which consist on generating new data points by linearly interpolating between the minority class and its neighbors. The implementation used in this study is present in smote-variants python library.

3.2.2. Synthetic data

Synthetic data can be defined as artificially annotated information generated by algorithms. This data, as well as interpolation methods, is used to augment data in domains with scarcity. However, synyhetic data differs from techniques like SMOTE in that it can also be used to replace real data in the training process. Additionally, synthetic data can address various issues beyond unbalanced classes, such as ensuringe overall data quality and meeting requirements for data privacy and fairness.

The generation of data involves the use of a generative model. This model aims to learn the true distribution of a dataset which it can then be used to generate new pointsh the same distribution. The approach originates from the technique known as Generative Adversarial Nets (GANs) [32], which employ a generative model that learns the distribution and a discriminative model that estimates the probability of a sample being real or synthetic. The idea behind GANs can be visualized in Fig. 4.

Since then, several variants for specific applications have emerged. In this study we employ the variant CTGAN [33] to generate tabular data implemented in the data-synthetic library.

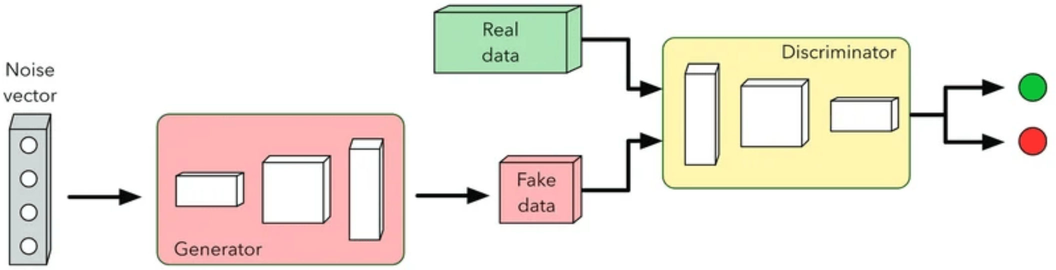

Fig. 4. Architecture of Generative Adversarial Nets (GANs), composed of a generator and a discriminator, the goal of the generator is to minimize the probability of the discriminator to distinguish real data from fake generated data. Courtesy of [34].

3.2.3. Evaluation: Data Slicing

The necessity of generating representative data evaluation originates from the fact that performance metrics overlook critical aspects of the model such as robustness, generalizability, and the presence of rational decisions [8]. Robustness, for instance, can be defined as the model's ability to make consistent predictions acrossall the present subpopulations on the dataset, this means, that a model that lacks robustness can have inconsistent performance and be more vulnerable to adversarial attacks.
To obtain a robust model, data slicing is a technique that aims to evaluate the performance on specific subpopulations with the purpose of evaluating the robustness of the system or identifying cases where the model performs significantly worse, which can subsequently uncover biases and explain controversial results.

Slices are usually made using domain knowledge to identify possible combinations of features where we might expect to find biases (such as sex or race). However, for some contexts, it can be hard to find the specific combination of values and categories. Therefore, a more systematic approach can be adopted by previously creating classes on certain ranges for numerical features and evaluating the performance on several combinations of features for each class.

A possible application of this approach, that is in fact the one use in this study, is by employing opt binning [35], a library for automatically find the best partition for each feature and algorithmically find the slices with worse performance using Slice Line [36] with a log loss as metric.

3.3 Multilabel classification problem

The multilabel classification differs from binary classification in that it involves predicting more than just two classes. However, it also distinguishes from multiclass classification where each instance is assumed to belong to only one of the categories. In multilable classification, each instance can be assigned any number of categories (or labels) [37].

Formally, the multilabel classification can be formulated as follows, for an instance x on an input space X, with Λ binary labels, the problem consist in learn a function φ that maps χ to λ, this is:

$$\varphi : \chi \rightarrow 2^{\lambda} \tag{1}$$

Multi-label classification is a much more complex problem than binary classification or multiclass classification due to the potential correlations between labels especially if the problem implies high dimensionality of the label space.

3.4 Performance metrics

The performance of the classification model, even when employing apparently complicated processes, can still be evaluated with the common metrics applicable to any other classification problem, such as accuracy and F1 score. In the case of crimes classification, it implies making the necessary adjustments to accommodate a multilabel problem [38]. For instance, in binary classification, we identify four possible outcomes: a positive instance correctly classified (True Positive or TP), a negative instance correctly classified (True Negative or TN), a negative instance incorrectly classified (False Positive or FP) and a positive instance incorrectly classified (False Negative or FN). With these terms we can define the True Positive Rate (TPR) or recall and the False Positive Rate (FPR) as:

$$TPR = \frac{TP}{TP+FN}, \tag{2}$$

$$FPR = \frac{FP}{FP+TN}, \tag{3}$$

From where we can define the metric of Area Under the Curve (AUC) as the probability that a randomly chosen positive example is higher correctly than a random chosen negative example, mathematically this is:

$$AUC = \int_0^1 \square \, TPR(FPR)d(FPR), \tag{4}$$

A common, general metric for evaluate overall performance of a classification model is the accuracy defined as:

$$Acc = \frac{TP+TN}{P+N}, \tag{5}$$

For purpose of introduce the F1 score, we define the Positive Predictive Value (PPV) or precision metrics as:

$$PPV = \frac{TP}{TP+FP}, \tag{6}$$

This way the F1 score for binary classification is:

$$F1 = 2 \cdot \frac{TPR \cdot PPV}{TPR+PPV}, \tag{7}$$

However, as previously mentioned, we are dealing with a multilabel problem. Therefore, this metric can only be applied to a specific label. To evaluate the overall performance, we mustconsider some method of averaging over each label. First, we consider the macro F1 score as the average of F1 score of each label, this is defined as follows:

$$F1_{macro} = \frac{1}{L} \sum_{i=1}^{k} \square \, F1_i, \tag{8}$$

Similarly, F1 weighted metric considers the number of instances per class called support S, as:

$$F1_{weighted} = \frac{\sum_{i=1}^{K} \square s_i \cdot F1_i}{\sum_{i=1}^{K} \square s_i}, \tag{9}$$

The f1 micro metric is defined as the F1 score calculated using the TPR and PPV per class:

$$TPR_{micro} = \frac{\sum_{i=1}^{k} \square TP_i}{\sum_{i=1}^{k} \square (TP_i + FN_i)}, \tag{10}$$

$$PPV_{micro} = \frac{\sum_{i=1}^{k} \square TP_i}{\sum_{i=1}^{k} \square (TP_i + FP_i)}, \tag{11}$$

Besides, these metrics to evaluate the model’s performance on the decision of the classification process, we introduce log loss function (also called logistic loss and cross-entropy loss). This metric evaluates the proximity between a prediction probability and the actual value, implying that the function, not only evaluates the incorrect predictions but also incorporates a penalization for biggest confidence on wrong predictions. The function L for a single prediction y with assign probability p is defined as:

$$L(y,p) = -\left(ylog(p) + (1-y)log(1-p)\right) \tag{12}$$

This function is employed mainly to evaluate the performance of the model over specific populations of the dataset to identify those that are more prone to misclassify.

3.5 Fairness metrics

For estimation of fairness in the model we mainly focus group fairness metrics, which consist on compare the classification outcomes from two groups, which commonly are selected through a sensitive feature or a class that for domain knowledge can be expected to discriminate in privileged and unprivileged groups [15].

Disparate Impact is a metric to evaluate fairness. It compares the proportion of individuals that receive a positive output for two groups: an unprivileged group and a privileged group.

$$\frac{Pr(Y=1 \vee D=unprivieged)}{Pr(Y=1 \vee D=privileged)}, \tag{13}$$

The statistical parity difference metric calculates the difference between in the ratio of favorable outcomes between monitored groups and reference groups

$$gap = \frac{P(D=unprivileged)-(D=unprivileged)}{P(D=privileged)-(D=privileged)}, \tag{14}$$

3.6 Privacy metrics

Usually the property of privacy is accomplished by modifying the original dataset inputting just the necessary amount of noise such that the main correlations between features is preserved, which makes the derived model useful while reduce the risk of re-identification [18], therefore in the literature several approaches has emerged to tackle this specific problem, for instance, one relevant work related to our problem performs an empirical evaluation of synthetic data as anonymities technique [39].

In our study, for ensuring the privacy of the system we evaluate the possibility of recreate the original training dataset from the synthetic dataset. In the context of privacy preserving machine learning, for other kinds of attacks some approaches focus on making harder to reconstruct the training dataset by inference attacks, here however we assume that even if the whole synthetic dataset is leaked the real information is still preserved, just as is assumed when using anonymization techniques, where even if we do not insert noise to the dataset, the synthetic dataset can be considering an anonymized version of the dataset. For this case the best strategy that an attacker can adopt is to try to reconstruct the original dataset with the public information in combination with the inferenced data, for evaluate the possibility of this approach, we measure the number of registers that can be linked to the original dataset assuming a perfect inference of the synthetic dataset.

Record Linkage is a method used to identify corresponding records between datasets, which helps in evaluating the risk of re-identification by comparing pairs of records from the original and synthetic datasets. In our experiment, we utilized the Record linkage library (Python Record Linkage Toolkit) to compare record pairs, which follows the methodology design in [40]. Here, we assume the attacker has access to the public information of the census and the whole synthetic training dataset, which implies that an effective attack would be able to identify which registers correspond to a specific suburb.

Euclidean distance metrics, on the other hand, measure the similarity between records by calculating the straight-line distance between data points in a multi-dimensional space. This allows us to determine the likelihood of an attacker matching a synthetic record to an original one, with smaller distances indicating a higher risk of information leakage. Together, these techniques provide a comprehensive assessment of the potential privacy risks associated with synthetic data.

We employed distance-based metrics to assess the probability of an attacker identifying an individual from synthetic data. Specifically, we calculated the Euclidean distance d between a record in the synthetic dataset and its nearest counterpart in the original dataset. A synthetic record with d=0 indicates a complete leak of actual information.

3.7 Experiment design

The dataset used in the project presents several concerns respect to imbalance, in first place with respect to the predicted class, it presents an intrinsic between-class imbalance, since as can be expected the number of registers of recidivism subjects are significantly smaller respect to single infractors. Subsequently, the dataset present both intrinsic and extrinsic within-class imbalances, which refers to inbalance on certain subpopulations that arise from the nature of the problem and from proceses of collecting, transforming and filtering the data, these can be observed in the fact that some crimes are more common than others but also that some descriptions can be harder to classify for the NL model.

In the context of the technical challenges to overcome, the experiments are design to evaluate the the effect of imbalances over the fairness and privacy at the most critical parts of the pipeline of the model, and in this way determine which combination of techniques result more appropriate.

For the evaluation of the multilabel classification problem for crime description, we compare the performance of fine-tuning three LLMs and two Bert based models, one that was previously finetuned for Spanish language and the second trained since the beginning on large corpus of multilingual data. After this a confident learning for detect hard cases and mislabels is applied.

The evaluation of the predictive model in the data-centric approach implies several comparisons, first to evaluate the utility and fairness, the approach is evaluated comparing their respective metrics over several classification models trained on the real dataset using two distributions of under sampling with 50:50 and 35:65 ratio for the target feature, the real data set augmented with SMOTE and the synthetic data for 50:50 ratio, emphasizing that in all the cases the validation is made using the test split of the real dataset and the total number of registers vary. All the experiments were performed using the library sklearn, where the tunning of hyperparameters was made using a grid of parameters searching for the best performance in the validation with the real test partition. For robustness validation and search of bad performance slices on the original dataset, the models were compared using both manual and automated slicing using log loss as the loss function, and after that using the augmentation of SMOTE and synthetic data on those slices, table 2 shows the values of the final iteration.

Finally, only consernient to the synthetic dataset, as it the only case related to the privacy of the model an empiric evaluation was made using Record linkage and Euclidean distance to nearest point, this way determines the privacy threshold obtained by the exchange between utility and privacy.

4 Results

A First, we present the results of the crime classification, as this serves as the basis for predicting recidivism. Table 1 displays the performance metrics, where the first two models are based on BERT and the remaining models are decoder-only.

Table 1. Comparative of classification metrics between models

Model	F1 micro	F1 macro	F1 weight	AUC
Beto	0.9713	0.8195	0.9778	0.9428
Bertin	0.8701	0.7024	0.8541	0.7606
Mistral	0.8192	0.7431	0.8053	0.7257
Phi	0.8109	0.7693	0.7942	0.8543
Llama3	0.8356	0.7759	0.8073	0.8761

As can be seen, the Beto model outperforms the rest in all metrics. The Bertin model, which is also an encoder-only model, takes second place in terms of F1-micro. However, its results are not as robust as Beto's, since it has the lowest F1-macro among all the models and is outperformed by the Phi and Llama 3 models in the AUC metric. Finally, the three decoder-only models showed similar performance across all metrics, with Llama 3 emerging as the best of the three, while Mistral and Phi produced mixed results.

Starting with the results of the recidivism model in Table 2, we present the final values of the GANs and SMOTE techniques for alleviating in-class imbalance compared to the approach of random undersampling. The first two columns show results for automatic slicing, while the following two correspond to the groups with a "high" degree of social backwardness and female sex, respectively. Overall, Table 2 shows that automatic slicing did not identify a significant slice, whereas for the other two groups, the loss function was significantly reduced.

Table 2. Results of loss function using slice finder approach and manual slicing for each dataset

Dataset	Logloss	% registers	Logloss GRS	Logloss SEX
Real 50:50	0.8392	5.14	0.7470	0.5414
Real 35:65	0.9611	6.28	0.7596	0.5521
Smote	-	-	0.5647	0.4335
Synthetic	-	-	0.5214	0.4303

Utility metrics, as shown in Table 3, display the values of accuracy (Acc) and AUC for real datasets as a baseline. It can be observed that the SMOTE technique barely improved performance, while the results with synthetic data were slightly worse.

Table 3. Utility metrics comparison for each dataset over two algorithms

Model	Dataset	Acc	AUC
Random Forest	Real 50:50	0.7897	0.7130
	Real 35:65	0.7482	0.6892
	Smote	0.8187	0.7723
	Synthetic	0.7564	0.6962
Logistic Regression	Real 50:50	0.6851	0.5698
	Real 35:65	0.6711	0.5433
	Smote	0.6429	0.6121
	Synthetic	0.6908	0.5772

The fairness metrics, as shown in Table 4, indicate that the baseline values on real datasets were significantly greater than those used to alleviate bias. Both techniques produced similar results, although synthetic data performed better in terms of the degree of social backwardness.

Table 4. Fairness metrics comparission for each dataset for SEX and GRS (degree of social backwardness) features

Model	Dataset	GAP GRS	DIR GRS	GAP SEX	DIR SEX
Random Forest	Real 50:50	9.82	1.58	12.15	2.06
	Real 35:65	11.05	1.27	25.65	1.87
	Smote	8.78	1.25	5.80	1.63
	Synthetic	-0.62	0.93	5.57	1.30
Logistic Regre-ssion	Real 50:50	8.55	3.02	8.15	2.29
	Real 35:65	12.48	1.28	19.18	1.68
	Smote	7.63	4.67	4.61	1.40
	Synthetic	0.91	0.76	0.32	0.63

Finally, regarding privacy, the analysis focuses solely on the synthetic dataset, as it is the only one with a privacy emphasis. The analysis involves determining the number of possible pairs subject to a specific degree of certainty for reconstructing the original dataset. This process begins by mapping the synthetic dataset to the census values using the degree of social backwardness as an index. Initially, there were 1,569,604 candidate pairs. According to the degree of certainty, the number of pairs decreases significantly; higher certainty means fewer identifiable records. This trend is illustrated in Table 5.

Table 5. Number of linked records according to grade of certainty

Score (max=2)	Whole dataset
1.900	1,131,593
1.950	523,019
1.990	41,643
1.995	12,573
1.999	708

By taking the nearest point and mapping it to the original dataset, we find that only 873 locations (less than one tenth) are actually used in the original dataset. From matching both frames, we observe an overlap of only 62 records that have a true equivalent. This result can be attributed to noise, making it impossible to distinguish the correct pairs from the incorrect ones.

Calculating the euclidean distance, with the features scaled for similitude in the range from 0 to 1 for the 36 features, we find that the distance of the the nearest record is 0.0308± 0.0069.

5 Discussion

The results presented in Table 1 demonstrate a clear difference between the Beto model and the other models. This disparity can largely be attributed to the fact that Beto was trained on the largest corpus of Spanish data. Another noteworthy point is that the large language models (LLMs) achieved results close to those of the Bertin model. This raises the question of whether fine-tuning the LLMs with more Spanish data before task-specific fine-tuning could potentially enable them to outperform Bertin.

Given that Beto emerged as the clear winner, we decided against conducting an exhaustive comparison of error types. However, it is worth noting that empirical observations revealed differences in performance between Bertin and the LLMs. For complex crime descriptions, Bertin often failed to assign any label or overused the spurious label, whereas the LLMs typically failed to assign all relevant labels but were more precise in assigning at least one correct label. This suggests that, despite a similar number of errors, the LLMs' approach is more useful as it retrieves part of the information in most cases.

The use of sy nthetic data for privacy preservation proved to be an appropriate approach. The total number of records that could be linked to the real dataset, assuming a perfect copy of the synthetic dataset was leaked, is negligible. Furthermore, the synthetic data still yielded comparable utility metric results.

6 Conclusions

The proposed approach assembling state of the art natural language processing and data-centric techniques result in an adequate solution to target the special needs of interpretability, fairness and privacy that a socio-technical problem requires.

The natural language techniques used for classify crimes allowed to extract valuable information from raw data, however, it was found that some of these techniques still lacks development to be applied to problems in Spanish language with vocabulary of narrow domains, since as could be expected the model with the best performance overall is the only with a previous finetune in Spanish.

Respect to the model for predict recidivism employing socio-demographical information of the detainee, it is relevant to consider the development of the system as a continuous effort to identify bias and fairness concerns between the algorithmic fairness, problem formulation and the intrinsic bias of society.

Finally, the study can be expanded to compare new metrics and techniques both for privacy and fairness with the objective to literally obtain the most robust systems, although the adopted approach represents a considerable baseline for deploying and maintaining a system with trustworthiness in consideration.

Acknowledgements

The authors thank the Jalisco government and the State Public Safety Council for their support. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsor.

References

1. B **Singh, J. P., Kroner, D. G., Wormith, J. S., Desmarais, S. L., & Hamilton, Z. (Eds.). (2018).** *Handbook of Recidivism Risk/Needs Assessment Tools* (1st ed.). Wiley. https://doi.org/10.1002/9781119184256
2. **Straub, V. J., Morgan, D., Bright, J., & Margetts, H. (2023).** Artificial intelligence in government: Concepts, standards, and a unified framework (arXiv:2210.17218).arXiv. http://arxiv.org/abs/2210.17218
3. **Jacobs, L. A., & Skeem, J. L. (2021).** Neighborhood Risk Factors for Recidivism: For Whom Do They Matter? American Journal of Community Psychology, 67(1–2), 103–115. https://doi.org/10.1002/ajcp.12463
4. ***INEGI. (2021).*** *Censo de Población y Vivienda 2020. Marco conceptual*. https://www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825197520.pdf
5. **Alarcon R. M. A., Orizaga T. J. A., Oliva D., Baltazar F., Cruz H. M. H. (2024).** Prediction of recidivism on robbery crimes through XAI models and sociodemographic factors for mass surveillance https://intranet.matematicas.uady.mx/journal/descargar.php?id=324
6. **Marchisio, K., Dash, S., Chen, H., Aumiller, D., Üstün, A., Hooker, S., & Ruder, S. (2024).** *How Does Quantization Affect Multilingual LLMs?* (arXiv:2407.03211). arXiv. http://arxiv.org/abs/2407.03211
7. **Bumgardner, V. K. C., Mullen, A., Armstrong, S., Hickey, C., & Talbert, J. (2023)**. Local Large Language Models for Complex Structured Medical Tasks (arXiv:2308.01727). arXiv. http://arxiv.org/abs/2308.01727
8. **Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2023).** Data-centric Artificial Intelligence: A Survey (arXiv:2303.10158). arXiv. http://arxiv.org/abs/2303.10158
9. **Murray, J., & Farrington, D. P. (2010).** Risk Factors for Conduct Disorder and Delinquency: Key Findings from Longitudinal Studies. The Canadian Journal of Psychiatry, 55(10), 633–642. https://doi.org/10.1177/070674371005501003
10. **OpenStreetMap Contributors**. OSM Server Side Scripting: Overpass API. [Computer software]. Available: https://overpass-api.de/api/interpreter. [Accessed: 30/June/23].
11. **Murgai, L. (2023).** Mitigating Bias in Machine Learning. mitigatingbias.ml.
12. “Book”, “Police” by kholifah, “database” by Sri Utami, “file processing” by memet, “filter data” by ProSymbols, “Meeting” by Hermine Blanquart, “Natural Language Processing” by ND’studios, “Monitoring” by kliwirt art, wit license CC BY 3.0, via Noun Project Inc, https://creativecommons.org/licenses/by/3.0/
13. **Diario Oficial Federal (2024).** CÓDIGO PENAL FEDERAL. Accesed: 07/June/24. Available: https://www.diputados.gob.mx/LeyesBiblio/pdf/CPF.pdf
14. **CNDH (2021),** CÓDIGO PENAL PARA EL ESTADO LIBRE Y SOBERANO DE JALISCO Accesed: 07/June/24, Available: https://normas.cndh.org.mx/Documentos/Jalisco/C%C3%B3digo_PE_Jal.pdf
15. **Caton, S., & Haas, C. (2024).** Fairness in Machine Learning: A Survey. ACM Computing Surveys, 56(7), 1–38. https://doi.org/10.1145/3616865
16. **European Commission. Directorate High Level Expert Group on Artificial Intelligence. (2019).** *Ethics guidelines for trustworthy AI.* Publications Office. https://data.europa.eu/doi/10.2759/177365
17. **Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022)**. A Survey on Bias and Fairness in Machine Learning (arXiv:1908.09635). arXiv. http://arxiv.org/abs/1908.09635
18. **De Cristofaro, E., (2020).** An Overview of Privacy in Machine Learning. (arXiv:2005.08679) https://arxiv.org/abs/2005.08679
19. **Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., & Wei, W. (2024).** Machine Learning for Synthetic Data Generation: A Review (arXiv:2302.04062). arXiv. http://arxiv.org/abs/2302.04062
20. **Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021).** Transformers: “The End of History” for NLP? (arXiv:2105.00813). arXiv. http://arxiv.org/abs/2105.00813
21. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023).** Attention Is All You Need (arXiv:1706.03762; Issue arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762
22. **Singh, S., & Mahmood, A. (2021).** The NLP Cookbook: Modern Recipes for Transformer based Deep Learning Architectures. IEEE Access, 9, 68675–68702. https://doi.org/10.1109/ACCESS.2021.3077350
23. **Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019).** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805
24. **Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2023).** Spanish Pre-trained BERT Model and Evaluation Data (arXiv:2308.02976). arXiv. http://arxiv.org/abs/2308.02976

25. **de la Rosa, J., Ponferrada, E. G., Villegas, P., Salas, P. G. de P., Romero, M., & Grandury, M. (2022).** BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling (arXiv:2207.06814). arXiv. http://arxiv.org/abs/2207.06814
26. **Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024).** Large Language Models: A Survey (arXiv:2402.06196). arXiv. http://arxiv.org/abs/2402.06196
27. **Llama Team. (2024).** The Llama 3 Herd of Models. arXiv. https://llama.meta.com/
28. **Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023)**. Mistral 7B (arXiv:2310.06825). arXiv. http://arxiv.org/abs/2310.06825
29. **Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, Q., Cai, M., Mendes, C. C. T., Chen, W., … Zhou, X. (2024).** Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone (arXiv:2404.14219). arXiv. http://arxiv.org/abs/2404.14219
30. **Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., & Napolitano, A. (2015).** Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. 2015 IEEE International Conference on Information Reuse and Integration, 197–202. https://doi.org/10.1109/IRI.2015.39
31. **Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002).** SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
32. **Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014).** Generative Adversarial Networks (arXiv:1406.2661). arXiv. http://arxiv.org/abs/1406.2661
33. **Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019).** Modeling Tabular data using Conditional GAN (arXiv:1907.00503). arXiv. http://arxiv.org/abs/1907.00503
34. **de Rosa, G. H., & Papa, J. P. (2021)**. A survey on text generation using generative adversarial networks. Pattern Recognition, 119, 108098. https://doi.org/10.1016/j.patcog.2021.108098
35. **Navas-Palencia, G. (2022).** Optimal binning: Mathematical programming formulation (arXiv:2001.08025). arXiv. http://arxiv.org/abs/2001.08025
36. **Sagadeeva, S., & Boehm, M. (2021).** SliceLine: Fast, Linear-Algebra-based Slice Finding for ML Model Debugging. Proceedings of the 2021 International Conference on Management of Data, 2290–2299. https://doi.org/10.1145/3448016.3457323
37. **Herrera, F., Charte, F., Rivera, A. J., & Del Jesus, M. J. (2016).** Multilabel Classification. In F. Herrera, F. Charte, A. J. Rivera, & M. J. Del Jesus, Multilabel Classification (pp. 17–31). Springer International Publishing. https://doi.org/10.1007/978-3-319-41111-8_2
38. **Herrera, F., Charte, F., Rivera, A. J., & Del Jesus, M. J. (2016).** Case Studies and Metrics. In F. Herrera, F. Charte, A. J. Rivera, & M. J. Del Jesus, Multilabel Classification (pp. 33–63). Springer International Publishing. https://doi.org/10.1007/978-3-319-41111-8_3
39. **Lu, P.-H., Wang, P.-C., & Yu, C.-M. (2019).** Empirical Evaluation on Synthetic Data Generation with Generative Adversarial Network. Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics, 1–6. https://doi.org/10.1145/3326467.3326474
40. **Christen, P. (2012).** Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31164-2