Bias Evaluation in Contextual Machine Learning

Anna Dalla Vecchia[1][0000−0001−7026−5205] and Kostas Stefanidis[2][0000−0003−1317−8062]

[1] Department of Computer Science, University of Verona, Italy
anna.dallavecchia@univr.it
[2] Faculty of Information Technology and Communication Sciences, Tampere University, Finland konstantinos.stefanidis@tuni.fi

Abstract. The integration of contextual information, like time, weather, or location, into machine learning (ML) models has been shown to improve the performance and personalization of the model. However, these additional features may unintentionally introduce biases, leading to disparities across target classes or subgroups. This paper presents a novel methodology to evaluate the bias introduced by contextual features in ML models. We introduce a general metric, called Contextual Bias, which quantifies the disparity in class-level performance when each contextual feature is removed. To efficiently assess feature influence without retraining the model from scratch, we leverage DaRE (Data Removal-Enabled) forests, a machine unlearning framework that allows post-hoc removal of contextual features. This approach is applied to a real-world dataset of tourist visits to Points of Interest (PoIs) in Verona, Italy, where contextual factors play a crucial role in shaping user behavior. This work provides a foundation for developing unbiased context-aware systems, highlighting the need to consider not only accuracy but also bias metrics when integrating contextual information into ML models.

Keywords: Bias · Context · Machine Learning.

1 Introduction

In recent years, machine learning (ML) models have become essential tools in many real-world applications, playing a crucial role in supporting the analysis and understanding of heterogeneous data sources. Although traditional ML approaches rely on features derived directly from sensor data and user activity, there is a growing interest in the integration of contextual information, such as weather conditions, location, and time, in order to enrich the feature and improve the model's performance [3, 6].

The inclusion of contextual features has been shown to enhance the accuracy of the model and enable a more tailored personalization. However, their integration may also influence the learning dynamics of the model without a clear understanding of those modifications. In particular, contextual features may indirectly introduce bias into the model, affecting its behavior across different subgroups or classes that are not explicitly declared as protected. This

poses important concerns about unbalanced or skewed outcomes of the decision model [17, 24].

Understanding how contextual features influence model behavior, and under which conditions they may lead to biased outcomes, remains a challenging task. Traditional feature importance techniques, such as feature importance scores in tree-based models (e.g., Random Forests or XGBoost), permutation feature importance, and model-agnostic approaches like SHAP [16] and LIME [21] have become widely adopted to gain insight into model behavior. Although these methods are useful for general interpretability, they are not explicitly designed to detect or quantify the bias introduced by contextual features. For this reason, they often fail to reveal the impact that context can have on fairness-related aspects of a model [10]. This work addresses this gap by proposing a methodology to evaluate the potential bias introduced by contextual features in ML models. The goal is to ensure that context-aware systems are not only accurate and personalized but also fair and reliable when deployed in real-world scenarios.

To this end, we proposed a novel formalization to quantify the bias associated with contextual features in a ML model. This general metric is used to guide an evaluation algorithm that leverages machine unlearning techniques [9] to obtain a model that discards the influence of the contextual feature under analysis. Specifically, we employ the data removal-enabled (DaRE) forest [9] to simulate the removal of individual contextual features after training, without requiring full retraining of the model. By comparing the predictions of the original model with the new one, we are able to investigate the role of the contextual features and how the biases change.

Lastly, we present a real-world dataset of Points of Interest (PoIs) in the city of Verona, where contextual information plays a key role in shaping tourist behaviors [6]. This dataset, which reflects temporal, spatial, and environmental factors, will serve as the basis for our case study and will support future research on context-aware and bias-aware machine learning in the tourism domain.

The remainder of the paper is organized as follows. In Section 2, the related work is discussed. Section 3 introduces the formalization of the problem and presents the new evaluation metric, while Section 4 outlines the proposed algorithm. Section 5 discusses a preliminary analysis of the dataset. Finally, Section 6 summarizes the findings and outlines the future directions.

2 Related Work

In recent years, the discussion about ethical concerns in the development and deployment of intelligent systems is gaining increasing attention [20, 8, 22, 7, 5]. Despite this growing interest, there is still no clear consensus on universally accepted ethical guidelines [23]. A key challenge is that ethical principles are often context-dependent, meaning that their interpretation and application can vary depending on the specific domain or use case. The authors in [13] discuss bias from a different perspective: it is not always necessary to completely remove it. The elimination of one form of bias may unintentionally give rise to another.

Instead of enforcing strict, unbiased situations, a possible approach may involve designing systems that allow users to transparently explore and adjust existing biases.

2.1 Bias in Machine Learning

In the machine learning domain, bias is typically understood as a deviation in model behavior that leads to unfair outcomes, especially for underrepresented groups in the dataset [17]. In classification tasks, for example, a biased model may achieve high overall accuracy while underperforming on specific classes or demographic groups. [10, 24] The state of the art on ethical issues related to machine learning is explored in [11], where various approaches aimed at enhancing fairness in machine learning are outlined. The authors conclude by presenting five open dilemmas for the research. This highlights that despite years of exploration of existing methodologies designed to mitigate potential ethical biases and inequities, it remains challenging to define the concept of fairness in a consistent manner [15, 19]. Recent studies have also investigated the relationship between human cognitive biases and machine learning models, highlighting how biases present in training data, originating from human behaviors and societal patterns, can be learned and amplified by machine learning systems. For instance, the study in [4] introduces a method to detect bias by swapping sensitive attributes, measuring the impact on predictions. Their results reveal biases, including lower wage predictions for females and Asians, underscoring the importance of bias detection for developing fairer machine learning models.

2.2 Context in Machine Learning

The integration of contextual information, such as time, location, weather, or user situation, has become increasingly common in machine learning, particularly in domains like recommendation systems, smart mobility, and personalized services. Contextual features enrich the input space and allow models to adapt their predictions to external conditions, often beneficial for increasing prediction accuracy, improving personalization, and enabling better interpretability [3]. For instance, in context-aware recommender systems, contextual factors allow the model to learn the user preferences that depend on the situation they are acting in [2]. Similarly, in spatiotemporal forecasting or mobile health, context is essential to model the variability of human behavior or system responses across time and space[1, 12]. In [26], the authors provide a systematic literature review describing how context is incorporated at various stages of recommender system development, identifying the most commonly used contextual features across different application domains, and discussing evaluation mechanisms in terms of datasets, metrics, and validation protocols. More recent work has investigated the synergies between context and temporal aspects, such as data aging, to enhance recommendations in dynamic environments like wearable devices and smart TVs [12]. These approaches emphasize the need for models to adapt not only to static contextual variables but also to evolving user behavior

and environmental conditions. In a similar direction, the authors in [18] study how contextual models can be leveraged to forecast Point of Interest occupation and users' preferences in the tourism domain, proposing an architecture that leverages both historical and contextual data.

To the best of our knowledge, the work most closely related to ours is presented in [25], where the authors present a system that explains fairness violations in machine learning models. This system identifies training data subsets that most contribute to fairness violations by estimating how model fairness changes when those subsets are removed. However, it primarily focuses on a single protected feature, whereas our approach considers the effects of multiple contextual factors on model behavior. Building upon these foundations, our work explores how contextual factors may influence the behavior of machine learning models, with a particular focus on their potential impact on model bias. This broader perspective allows us to better understand the complex role of context in model behavior and its potential implications in real-world applications.

3 Bias in Contextual Machine Learning

Concerns about fairness and bias have emerged as critical aspects since machine learning systems increasingly incorporate contextual information to enhance performance and personalization [3, 27]. Typically, the bias concept in machine learning refers to systematic deviations in model behavior that disadvantage certain groups or data subsets. This section first presents the traditional contextual machine learning models, and then the potential biases emerging from those models when contextual data is introduced.

3.1 Context Definition in Machine Learning

Nowadays, we are flooded by data generated from different sources, like sensors, wearable devices, or other types of platforms. As a result, tools that help users understand data values and trends are becoming an invaluable assistant both for users and service providers. In many real-world applications, supplementary information, such as the geographical location, weather conditions and holidays, plays a crucial role in improving the performance and the interpretability of such tools. In this work, we refer to this kind of additional information as contextual features, which encompass any external or auxiliary data that can describe the circumstances under which user interactions or events occur.

To formally capture the role of contextual features in a machine learning setting, we define the model input as a combination of both intrinsic and contextual data. Let $\mathcal{M}$ be a general machine learning model, $X \in \mathbb{R}^n$ the original input feature vector, and $Y \in \mathbb{R}$ the target variable. Each element $x \in X$ represents an intrinsic attribute of the instance. Let $C \in \mathbb{R}^m$ be the set of contextual features, where each $c \in C$ encodes external information related to the instance, such as

temporal or environmental context, i.e. contextual features. The complete input to the model is then defined as the augmented feature vector:

$$\hat{X} = [X \parallel C] \in \mathbb{R}^{n+m}$$

where $\parallel$ denotes the concatenation of the original features and the contextual features. The model is trained to learn a function:

$$\mathcal{M} : \mathbb{R}^{n+m} \to \mathbb{R}$$

mapping the enriched input $\hat{X}$ to the predicted output Y.

3.2 Bias Definition in Contextual Machine Learning

In recent years, the concept of bias and fairness has emerged as a topic to pay attention to in different areas of computer science. Also, machine learning is becoming a domain in which the implications of these concepts have to be explored [17]. In general, bias in machine learning can be seen as a deviation during the learning process in which some aspects of the datasets are considered more important, resulting in unfair, skewed, or wrong outcomes of the model [14, 17]. These problems, in most cases, originate from an unbalanced dataset. For instance, if a dataset is biased towards a particular class, the resulting model will excel for that class, while it may underperform for the less represented group. However, analyzing a well-processed dataset can reveal how each feature influences the outcome and whether it is still biased. Specifically, when contextual features are introduced, they can reveal an unexpected bias due to the provided additional information.

In this work, we investigate the possible presence of bias by studying how contextual features influence the learning process and whether they introduce disparities in model performance across different target classes. Rather than treating context as a neutral input, we propose a novel perspective in which context is analyzed as a potential source of bias that may affect some prediction outcomes. More in detail, we investigate the precision of the model in a classification task, analyzing when there is a significant difference between the model trained with and without each contextual feature. We define contextual bias in a classification model as the disparity in predictive performance across classes of different models. This bias can be quantified by measuring the variation in precision

Let $\mathcal{M}$ be a classification model trained to predict class labels $Y \in \{1, \ldots, K\}$ from input instances $\hat{X} = [X \parallel C]$, where $X \in \mathbb{R}^n$ represents the original feature and $C \in \mathbb{R}^m$ the contextual features. Let $\mathcal{M}_{-c}$ be the model trained without the feature $c \in C$. We define the *Contextual Bias* introduced by the feature c, denoted as $CB(c)$, as the disparity in class-level predictive performance between $\mathcal{M}$ and $\mathcal{M}_{-c}$. Using the precision evaluation metric, the contextual bias is measured as the standard deviation of the change in per-class precision:

$$CB(c) = \sqrt{\frac{1}{K} \sum_{t=1}^{K} \left(\Delta Prec_t(c) - \overline{\Delta Prec(c)} \right)^2}$$

where:

- $\Delta Prec_t(c) = Prec_t(\mathcal{M}) - Prec_t(\mathcal{M}_{-c})$
- $\overline{\Delta Prec(c)} = \frac{1}{K} \sum_{t=1}^{K} \Delta Prec_t(c)$
- $Prec_t(\cdot)$ denotes the precision on class t under the given model.

A high value of $CB(c)$ indicates that the contextual feature c has a disproportionate effect on the model's performance across different classes, suggesting it may be a source of contextual bias. This formulation is based on the per-class precision in a multi-class setting, as:

$$Prec_t(\cdot) = \frac{TP_t}{TP_t + FP_t}$$

where TP_t and FP_t denote the number of true positive and false positive predictions for class t. However, the contextual bias measure $CB(c)$ is orthogonal and can be generalized to any class-level evaluation function ϕ, such as recall, F1-score, or calibration error, depending on the desired fairness criteria.

4 Contextual Bias Evaluation

Evaluating the influence of individual features in a trained model typically involves retraining the model multiple times, each time excluding a single feature. Let $\hat{X} = [X \parallel C]$ be the full set of input features, composed of input features $x \in \mathbb{R}^n$ and contextual information $C \in \mathbb{R}^m$. The traditional approach trains m separate models $\mathcal{M}_{-c}$, where $c \in C$ and each model is trained on $\hat{X} \setminus \{-c\}$ While this method is straightforward and widely used, it becomes computationally expensive when applied to a model, especially when the number of contextual features is large.

To address this limitation, in this study, we adopt a more efficient alternative based on machine unlearning introduced in [9], which enables efficient removal of training data from Random Forest models through a specialized structure called DaRE (Data Removal-Enabled) forests. DaRE forests support exact and efficient unlearning by modifying only the necessary subtrees when data is removed, avoiding the need for full retraining. This is achieved by strategically introducing random nodes in the upper levels of the trees and caching sufficient statistics at decision and leaf nodes. In other words, retraining the model $\mathcal{M}$ is not required each time feature c needs to be excluded. We refer to the resulting model, in which c has been removed, as $\mathcal{M}_{-c}^{DaRE}$.

Specifically, we apply a feature-level unlearning procedure that allows us to estimate the effect of removing a specific feature from the trained model, without requiring full retraining. This enables a more scalable analysis of feature influence, particularly in scenarios involving high-dimensional or context-rich data, where the traditional approaches based on retraining would be computationally expensive.

To analyze the potential bias introduced by each feature, we adopt an iterative approach based on DaRE presented in Algorithm 1. Particularly, the

Algorithm 1

Input: Dataset $\mathcal{D}$, contextual features C
Output: DaRE Model set $\mathcal{M}_{all}$, contextual bias score set S

$\mathcal{M}_{all} \leftarrow \{\ \}$
$S \leftarrow \{\ \}$
$\mathcal{M} \leftarrow trained(\mathcal{D})$
for each $c \in C$ **do**
    $\mathcal{M}_{all} \leftarrow \mathcal{M}_{all} \cup \mathcal{M}_{-c}^{DaRE}$
    $S \leftarrow S \cup eval(\mathcal{M}, \mathcal{M}_{-c}^{DaRE}, \mathcal{D})$
end for
return $\mathcal{M}_{all}, S$

unlearning procedure is applied iteratively, removing one feature at a time until all features have been exhausted. For each resulting model, we evaluate its performance on the test set, saving the previously defined contextual bias measure (CB) computed by the *eval* function. By observing the results collected in S, it is possible to identify which contextual feature introduces the most bias by analyzing its corresponding CB value. A value close to zero indicates that the feature has a uniform impact across all classes. Conversely, a value significantly greater than zero suggests that the feature affects the model differently across classes, thus introducing contextual bias. The study focuses not only on the overall accuracy of the model, but also on the outcomes obtained for each target class, in order to capture possible shifts in predictive performance across different classes. Thus, we aim to identify whether the feature affects the prediction quality for specific subsets of the data, which may indicate the presence of feature-specific or context-induced bias.

4.1 Complexity Analysis

As demonstrated by the authors in [9], the complexity of training a DaRE model is equivalent to that of training a standard Random Forest, i.e.,

$$\mathcal{O}(T \cdot n \cdot \tilde{p} \cdot k \cdot d_{\max})$$

where T is the number of trees in the forest, n is the number of training instances, $\tilde{p}$ is the number of attributes randomly selected at each node, k is the number of candidate thresholds per attribute, and $d_{\max}$ is the maximum tree depth. The time complexity of deleting a single instance from a DaRE tree, when no structural modification is required and all attribute thresholds remain valid, is:

$$\mathcal{O}(\tilde{p} \cdot k \cdot d_{\max})$$

However, if a node's attribute thresholds become invalid, an additional cost of $\mathcal{O}(|D| \log |D|)$ is incurred to select a new valid threshold, where $|D|$ is the number of data points reaching that node. Furthermore, if the node at depth d with $|D|$ instances requires retraining, the cost of rebuilding the affected subtree is:

$$\mathcal{O}(\tilde{p} \cdot |D| \cdot (d_{\max} - d))$$

The overall time complexity of Algorithm 1 can be decomposed into three components. The first corresponds to the initial training of the DaRE model in line 3, which takes $\mathcal{O}(T \cdot n \cdot \tilde{p} \cdot k \cdot d_{\max})$. Next, for each contextual feature $c \in C$, the algorithm performs machine unlearning to obtain the model $\mathcal{M}_{-c}^{DaRE}$ at line 5, whose complexity depends on the specific operations required as described above. Finally, the evaluation step of line 6 adds an additional cost of $\mathcal{O}(n)$ per feature. Consequently, the total complexity is the sum of these costs over all contextual features.

5 Early Findings on PoIs Case Study

In this work, we apply the proposed methodology to a case study focused on Points of Interest (PoIs) in the tourism domain. Understanding and incorporating contextual information is crucial in this setting, as tourists' behaviors and visit patterns are strongly influenced by external factors such as time, weather, or events. To validate the importance of context in this domain, we build upon previous work presented in [6], where the authors investigate how contextual factors influence the occupancy of PoIs, focusing on the city of Verona, Italy.

More in detail, the study introduces the notion of a *Touristic Visit* representing the set of input feature vectors X with the corresponding *Visit Context*, which defines the tuple of contextual features C defined in this work. The objective of the study was to forecast the occupation of each PoI at different timeslots of the day, then support the development of a crowd-aware recommendation system capable of suggesting to tourists the less crowded attractions.

The problem was modeled using both traditional machine learning (Random Forest) and deep learning approaches. These models were evaluated in two settings: using raw historical data only, and using data enriched with contextual features. The results obtained shows that including context significantly improves accuracy. Furthermore, models trained with contextual data demonstrated better generalization capabilities when applied to PoIs with limited historical information. That study demonstrates how integrating contextual features improves the forecast of the PoI crowding, benefiting both tourists, who can better plan their visits, and city tourist service providers, who can manage urban crowding more efficiently.

The role of context in the tourism domain is further investigated in this work from a different perspective. While previous studies, including [3, 6], have demonstrated the benefits of incorporating contextual information to improve recommendations, the current investigation focuses on understanding when such contextual features may act as sources of bias in the model. The aim is to ensure that context-aware models are not only more accurate but also fair and reliable when applied to real-world tourism scenarios.

5.1 Dataset

To validate our approach, this work focuses on the use case presented in [6] regarding the touristic scenario. More in detail, the real-world dataset tracks

touristic activity in the city of Verona in Italy. The dataset spans nine years from 2014 to 2022 and includes approximately 4.4 million visit entries, involving about 1 million unique tourists and seventeen Points of Interest (PoIs). Each entry consists of a timestamped visit linked to an anonymized user ID and a PoI identifier. To enhance the dataset, we integrated contextual variables, including meteorological conditions and calendar-based indicators such as public holidays. Specifically, the augmented feature vector $\hat{X}$ is represented by the concatenation of the input features

$$X = \{u, p, t, lat, long\}$$

where u is the user identifier, p is the PoI identifier, t is the timestamp of the visit, and lat and $long$ are the geographical coordinates of the PoI, and of the contextual features

$$C = \{ts, day, month, year, doy, dow, week, rain, temp, festive, hol\}$$

where ts is the timeslot of the performed visit (e.g. morning, noon, afternoon), day, $month$, $year$ are the day of the month, the month, the year respectively, doy is the day of the year, dow and $week$ are the day of the week and the week number in a year, $rain$ is a string for the weather condition (e.g., rainy, sunny), $temp$ is the temperature, $festive$ and hol are two boolean indicators for the day is a weekend or a public holiday.

5.2 The Role of the Context

Firstly, it is important to highlight the relationship between the crowding levels at PoIs and the contextual factors added to the original dataset.

As expected, the weather conditions strongly influence the tourists' affluence in different PoI, and the tourist visit patterns. Fig. 1 illustrates that the number of visits to certain PoIs on the same calendar day in two different years with different weather conditions can notably vary. Outdoor locations, such as POI 49, show a clear drop in attendance on rainy days. Conversely, indoor venues such as PoIs 62 and 63 may even benefit from adverse weather, attracting more visitors who will prefer indoor activities when outdoor conditions are unfavorable.

Fig. 2 presents another illustrative case, showing the number of visits to Juliet's House in February. This PoI, recognized as the symbol of love, attracts more visitors around Valentine's Day. Looking at the plot, the red bars represent the number of visits on February 14th across different years, while the blue bars represent the average number of visits for the same weekday throughout the month. There is a notable spike on Valentine's Day compared to the baseline, highlighting how special events can significantly influence tourist behavior, leading to atypical crowding on typically less congested days.

5.3 The per-class evaluation

The Fig. 3 illustrates the class-wise evaluation of Precision, Recall, and F1-score for the Random Forest classifier applied to our dataset enriched with contextual

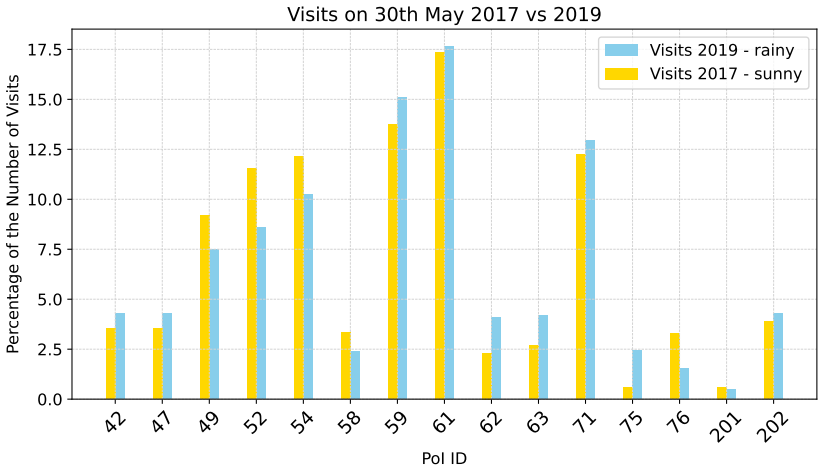

Fig. 1. Number of visits on the 30th of May of two years with different weather conditions.

factors. We trained a Random Forest classifier where occupancy is discretized into three classes (i.e., Low, Medium, High) based on percentile thresholds computed for each PoI. Each subplot represents one metric, with classes denoted on the x-axis and the respective metric on the y-axis. It is clear that the class *Medium* exhibits lower scores across all three metrics compared to the other classes. This indicates that the model systematically penalizes instances from class *Medium*, highlighting a potential bias or imbalance in model performance.

Consequently, a deeper analysis is required to discern whether the observed underperformance of *Medium* is driven by contextual features that influence the model's predictions or whether it reflects input features. Such an investigation is essential to properly interpret these results and to guide the training in order to obtain an unbiased model.

6 Conlcusion and Future Work

This work presents a preliminary investigation into the role of contextual features as potential sources of bias in machine learning models. While the use of contextual information to enhance model performance is widely discussed in the literature, its interactions with class labels and other input features remain insufficiently explored. To address this challenge, we present a general metric to quantify the bias introduced by contextual features and propose a novel methodology based on machine unlearning. Our approach evaluates new models without the contextual feature being analyzed, eliminating the need for full model retraining, thanks to the DaRE forest approach. Additionally, we presented a real-world dataset of tourist visits in the city of Verona, where contextual information plays a crucial role in user behavior and recommendation dynamics.

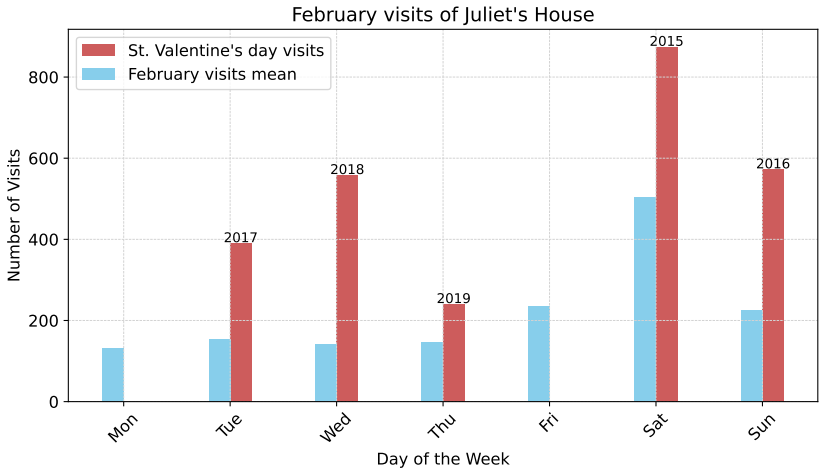

Fig. 2. February weekly visits.

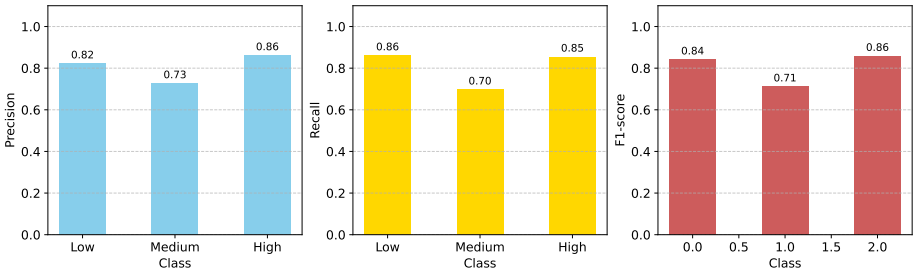

Fig. 3. Precision, Recall, and F1-score of the Random Forest Classifier.

This dataset will serve as a starting point for future studies on context-aware, unbiased, and explainable machine learning models. Beyond its methodological contribution, this study lays the foundation for an alternative explanation of feature importance, based not only on model accuracy but also on its potential to introduce biases across classes.

Future work will focus on conducting a more comprehensive empirical evaluation of the proposed methodology using the presented dataset. Additionally, we plan to investigate its applicability across other domains to further assess the generalizability and orthogonality of the contextual bias problem discussed. Another direction involves integrating our approach with traditional feature importance techniques, such as feature importance scores in tree-based models, permutation-based feature importance, and other model-agnostic approaches.

References

1. Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P.: Towards a better understanding of context and context-awareness. In: Proceedings of Handheld and Ubiquitous Computing HUC'99. LNCS, vol. 1707, pp. 304–307. Springer (1999)
2. Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating contextual information in recommender systems using a multidimensional approach **23**(1) (2005)
3. Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Recommender Systems Handbook, pp. 191–226 (2015)
4. Alelyani, S.: Detection and evaluation of machine learning bias. Applied Sciences **11**(14), 6271 (2021)
5. Araújo, T.B., Efthymiou, V., Christophides, V., Pitoura, E., Stefanidis, K.: TREATS: fairness-aware entity resolution over streaming data. Inf. Syst. **129**, 102506 (2025)
6. Belussi, A., Cinelli, A., Dalla Vecchia, A., Migliorini, S., Quaresmini, M., Quintarelli, E.: Forecasting POI Occupation with Contextual Machine Learning. In: Advances in Databases and Information Systems. pp. 361–376. Springer International Publishing (2022)
7. Borges, R., Sahlgren, O., Koivunen, S., Stefanidis, K., Olsson, T., Laitinen, A.: Multi-objective fairness in team assembly. In: New Trends in Database and Information Systems - ADBIS 2023 Short Papers, Doctoral Consortium and Workshops: AIDMA, DOING, K-Gals, MADEISD, PeRS, Barcelona, Spain, September 4-7, 2023, Proceedings. Communications in Computer and Information Science, vol. 1850, pp. 106–116 (2023)
8. Borges, R., Stefanidis, K.: Feature-blind fairness in collaborative filtering recommender systems. Knowl. Inf. Syst. **64**(4), 943–962 (2022)
9. Brophy, J., Lowd, D.: Machine unlearning for random forests. In: International Conference on Machine Learning. pp. 1092–1104. PMLR (2021)
10. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR (2018)
11. Caton, S., Haas, C.: Fairness in machine learning: A survey. ACM Computer Surveys **56**(7) (2024)
12. Dalla Vecchia, A., Marastoni, N., Oliboni, B., Quintarelli, E.: The synergies of context and data aging in recommendations. In: Big Data Analytics and Knowledge Discovery. pp. 80–87. Springer Nature Switzerland (2023)
13. Demartini, G., Roitero, K., Mizzaro, S.: Data bias management. Commun. ACM **67**(1), 28–32 (2023)
14. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Trans. Inf. Syst. **14**(3), 330–347 (1996)
15. Hutchinson, B., Mitchell, M.: 50 years of test (un)fairness: Lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. p. 49–58. FAT* '19, Association for Computing Machinery (2019)
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777. NIPS'17, Curran Associates Inc. (2017)

17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6) (2021)
18. Migliorini, S., Dalla Vecchia, A., Belussi, A., Quintarelli, E.: ARTEMIS: a context-aware recommendation system with crowding forecaster for the touristic domain. Information Systems Frontiers pp. 1–27 (2024)
19. Narayanan, A.: Translation tutorial: 21 fairness definitions and their politics. In: Proc. conf. fairness accountability transp. vol. 1170, p. 3 (2018)
20. Pitoura, E., Stefanidis, K., Koutrika, G.: Fairness in rankings and recommendations: an overview. VLDB J. **31**(3), 431–458 (2022)
21. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144. KDD '16, Association for Computing Machinery (2016)
22. Sacharidis, D., Giannopoulos, G., Papastefanatos, G., Stefanidis, K.: Auditing for spatial fairness. In: Proceedings 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28-31, 2023. pp. 485–491 (2023)
23. Steen, M.: Ethics as a participatory and iterative process. Communications of the ACM **66**(5), 27–29 (2023)
24. Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. In: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. EAAMO '21, Association for Computing Machinery (2021)
25. Surve, T., Pradhan, R.: Explaining fairness violations using machine unlearning (2025)
26. Villegas, N.M., Sánchez, C., Díaz-Cely, J., Tamura, G.: Characterizing context-aware recommender systems: A systematic literature review. Knowledge-Based Systems **140**, 173–200 (2018)
27. Zhao, Y., Wang, Y., Liu, Y., Cheng, X., Aggarwal, C.C., Derr, T.: Fairness and diversity in recommender systems: A survey. ACM Trans. Intell. Syst. Technol. **16**(1) (2025)