Proceedings of CONF-CDS 2026 Symposium: Machine Learning and Neural Network Applications in Engineering
DOI: 10.54254/2755-2721/2026.AST35981

Predicting Hate Crime Bias Using Machine Learning Based on Multidimensional Socio-demographic and Incident Features

Sophia Zhou

Crean Lutheran, Irvine, USA
jyzhou944@gmail.com

Abstract. This study applies a Random Forest classifier to the FBI's national Uniform Crime Reporting (UCR) dataset containing 218,069 single-bias hate crimes from 1991to 2020. Within the dataset, crimes were predicted based on bias-driven motives: race, sexual orientation, gender, religion, or disability. Thirteen incident features were analyzed to form the prediction. The primary model chosen for this study oversampled the two smallest categories and tuned hyperparameters. Ultimately, an accuracy of 68.7% with F1 = 0.80 and recall = 0.93 was achieved for the main prediction target: racial bias. Five trials were conducted with different configurations of the train/test model, compared across two metrics. It can be seen that there is a significant trade-off between total balance within all 5 categories versus overall accuracy. It was determined the incident year was the dominant predictive feature, likely due to FBI changes in definition and classification across the 1991-2020 period. Results demonstrate the predictability of indicators within official records and are limited to UCR metrics of data collection.

Keywords: Hate Crime Classification, Random Forest, Racial bias, Class imbalance, FBI Uniform Crime Reporting

1. Introduction

Hate crimes are defined as illegal acts against a person or group because of a certain bias, specifically reasons such as race, religion, sexual orientation, gender, or disability. Instead of being aimed at a specific person, hate crimes target entire groups and characteristics that members of that group carry. Therefore, the impact goes far beyond single-standing cases, as a hate crime is committed against an entire community, and generates fear and psychological harm by association [1]. The Hate Crime Statistics Act (HCSA) in 1990 was the first systematic federally backed collection of data on hate crime in the United States [2]. Later, this opened doors for the publication of annual data and further documentation, such as the FBI administering the Uniform Crime Reporting (UCR) Program in 1992, from which this study derives its database.

This study evaluates whether a machine learning model can accurately predict the bias motivation of hate crimes, given details of the incident provided within the UCR database. This research serves as a reference for accuracy costs and imbalance handling that future studies can expand upon.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AST35981

2. Related works

Previous research has examined the impact of social, regional, and demographic factors on hate crimes. The relationship between population density, ethnicity, and socio-economic conditions within each county and state, and hate crime incidents had been analyzed in past studies. In a 2022 study analyzing hate crimes from 2012 to 2016, it was found that anti-Black, Hispanic, and White incidents fluctuated significantly across rural-urban divides, differing based on the demographics of communities [3]. Machine learning has been used in crime detection by applying logistic regression, decision trees, and neural networks. Ana Ortiz Salazar trained classifiers on Seattle Police incident reports, determining if crimes were bias-motivated, using machine learning for greater efficiency and accuracy [4] Another study focused on identifying hate crimes absent from government databases by extracting information from news outlets instead. It was confirmed through this study that hate crimes are under-reported, and any classification of crimes presents the danger of human bias [5]. This study differs from the above by applying a Random Forest classifier, using national data, and generating multiclass bias predictions, with a focus on race-motivated incidents. Findings may differ based on UCR variation across states, and model performance will reflect differences in data density alongside the desired target of structural patterns.

3. Data and methodology

3.1. Source of data

This study was conducted using the FBI Dataset on Hate Crime Statistics, released under the Uniform Crime Reporting (UCR) Program, and distributed on Kaggle. The data entries are displayed as a set of national and state-level crime entries, covering the years 1991-2020. The program collects data on the date, location, information regarding the offender and the victim, and the bias motivator of the incident. This paper divides motives into 5 categories: bias driven by race/ethnicity, gender, sexual orientation, religion, and disability. Additionally, crimes are classified as single or multiple bias, depending on whether there is a record of more than one motivation applied. There have been changes throughout the years for data classification in the file used. Since 2013, agencies have been given the choice to record up to five bias motivators per offense type. In 2015, the FBI revised its race/ethnicity bias categories and added seven religion-oriented motivators, as well as an anti-Arab subcategory [6]. These changes are intentionally designed and align with the findings described in Section 3.3. The dataset used includes a national aggregate file and a separate file for each state, so in order to avoid row duplication, this study uses only the national file (N = 219,073).

3.2. Label construction

The focus of this study is the BIAS_DESC column, which records bias motivation and is semicolon-delimited. In order to eliminate ambiguity, this study centers its analysis around single-bias incidents, which account for 218,059 out of 219,073 cases (99.5%). The incidents that contain multiple values are eliminated (MULTIPLE_BIAS == 'S'), which numbers 1,014 (0.5%) of the entire dataset.

In total, there are 35 distinct bias motivators, which were sorted deterministically into 5 umbrella classifications. This study centers its focus on Race-motivated incidents, as that is the primary represented motivator in the data, and thus reached the highest accuracy level in testing. Secondary

DOI: 10.54254/2755-2721/2026.AST35981

categories include Sexual Orientation, Gender, Religion, and Disability, which were analyzed for context. The Race category contains the FBI bias descriptors of Anti-African American/Anti-Black, Anti-American Indian/Alaska Native, Anti-Arab, Anti-Asian, Anti-Hispanic, Anti-Multi Race, Anti-Other Race/Ethnicity/Ancestry, and Anti-White. The Sexual Orientation category contains real FBI bias descriptors of Anti-Bisexual, Anti-Gay(Male), Anti-Heterosexual, Anti-LGBT(Mixed), and Anti-Lesbian. The Gender category includes Anti-Female, Anti-Gender Non-Conforming, Anti-Male, and Anti-Transgender. Religion-related biases include: Anti-Atheism/Agnosticism, Anti-Buddhist, Anti-Catholic, Anti-Eastern Orthodox, Anti-Hindu, Anti-Islamic, Anti-Jehovah's Witness, Anti-Jewish, Anti-Mormon, Anti-Multiple Religions Group, Anti-Other Christian, Anti-Other Religion, Anti-Protestant, Anti-Sikh. Finally, the Disability category consists of Anti-Mental Disability and Anti-Physical Disability.

One limitation of the original FBI dataset is the "Anti-LGBT (Mixed Group)" category combines biases aimed towards sexual orientation and gender identity. This analysis assigns the group under Sexual Orientation, as ¾ of the terms can be classified under it.

3.3. Encoding

The raw FBI dataset used collected 28 metrics of data. Table 1 details what information each column provides, and how it was encoded.

Table 1. Data disposition

Column	Description
DATA_YEAR	Incident year
OFFENSE_NAME	Crime type; frequency-encoded
LOCATION_NAME	Location type; frequency-encoded
STATE_NAME	State name; frequency-encoded
OFFENDER_RACE	Offender race; one-hot encoded
VICTIM_TYPES	Victim type; frequency-encoded
POPULATION_GROUP_DESC	Population size tier; ordinal 1–8 scale
INCIDENT_DATE	Months cyclically encoded as sine/cosine
MULTIPLE_OFFENSE	Multiple offenses were flagged and removed; one-hot encoded
AGENCY_TYPE_NAME, DIVISION_NAME, REGION_NAME	Geographic context; one-hot encoded
VICTIM_COUNT	Total victims; log-transformed
BIAS_DESC	Prediction target; 5 categories

Out of the 28 columns, 13 were used as predictors for the model, 1 was used as the prediction target, 1 was used as a filter prior to analysis, and 13 were excluded from the study. Exclusions were made due to the data lacking descriptive content regarding the final predictor, as in the case of INCIDENT_ID and ORI; PUB_AGENCY_NAME was excluded due to risk of inflating accuracy; POPULATION_GROUP_CODE, STATE_ABBR, and TOTAL_INDIVIDUAL_VICTIMS were redundant with columns used as predictors; and PUB_AGENCY_UNIT, OFFENDER_ETHNICITY,

DOI: 10.54254/2755-2721/2026.AST35981

ADULT_VICTIM_COUNT, JUVENILE_VICTIM_COUNT, ADULT_OFFENDER_COUNT, and JUVENILE_OFFENDER_COUNT were excluded due to substantial missing information. Finally, TOTAL_OFFENDER_COUNT was not included despite having lower missingness than specific counts of adult or juvenile counts, but it was not included in this study and can be a point of future work.

The specific encoding method used for each metric is dependent on the number of unique values that the predictor carries; for high cardinality values with greater than 15 unique values, frequency encoding was used to compare relative frequencies without assuming ordinal relationships (OFFENSE_NAME, LOCATION_NAME, VICTIM_TYPES, STATE_NAME). One-hot encoding was used for AGENCY_TYPE_NAME, DIVISION_NAME, REGION_NAME, OFFENDER_RACE, MULTIPLE_OFFENSE , which contained fewer than 15 unique values, and binary indicator columns were employed for each value. Ordinal encoding was used for POPULATION_GROUP_DESC, which converted values according to a 1–8 integer scale in order to order population sizes from smallest to largest. INCIDENT_DATE was cyclically encoded in a sine/cosine form, so that December and January are adjacent rather than maximally distant. Finally, a log transformation prior to training addressed the right skew in data for VICTIM_COUNT.

3.4. Model and evaluation

The model selected in this project is a Random Forest classifier, which does not look for linear separability in data, accounts for categorical and numeric columns without scaling, and generalizes the dataset to take the best average [7]. This choice is made based on comparisons with several models that performed worse under the same conditions, as shown in Table 2.

Table 2. Model comparison

Algorithm	Accuracy	Macro-F1
Decision Tree (max depth = 20)	0.468	0.304
Decision Tree (unpruned)	0.463	0.296
MLP Neural Network	0.633	0.251
HistGradientBoosting	0.456	0.325
Random Forest — Configuration E	0.687	0.368

Decision Trees utilize a single tree rather than the Random Forest's averaging, which gives it an advantage. The MLP treats frequency encoded integers as continuous, which disregards the encoding that is designed to indicate an ordinal relationship. HistGradientBoosing is technically a more advanced algorithm, as trees are built sequentially and designed to rectify predecessors. However, it ended up underperforming Random Forest when tested.

In the dataset used, there is a severe imbalance of data from each bias category (Race 64.1%, Religion 18.6%, Sexual Orientation 15.7%, Disability 0.8%, and Gender 0.7%). This study was designed to primarily focus on racial bias; thus, models were selected based on performance in overall accuracy. Macro-F1, which considers all 5 categories equally, was also tested for comparison, and the cost of doing so drops the accuracy significantly in each category (see below). All configurations were tested based on splitting the dataset 80/20 between train and test.

DOI: 10.54254/2755-2721/2026.AST35981

Table 3. Model comparison

Model	Accuracy	Macro-F1
A: Random Forest, unweighted	0.665	0.352
B: Random Forest, class-weighted	0.660	0.354
C: Random Forest, tuned + class-weighted	0.617	0.405
D: HistGradientBoosting, class-weighted	0.456	0.325
E: Random Forest, tuned + oversampled	0.687	0.368

Table 3 reveals the significant difference between overall accuracy. Macro-F1 is attributed to the different priorities of each metric. The significant imbalance in raw data saw the solution of oversampling, achieving the highest degree of balance for both measurements. Configuration E, which oversampled the two smallest classes (Disability, Gender), reported the highest overall accuracy (0.687) and also was the most accurate in predicting the Racial Bias category specifically (Section 3.2). The oversampling tactic had its drawbacks, primarily shifting the model's focus away from the Sexual Orientation category, leaving its recall at 0.06, which was significantly worse compared to other tests. For a study where all categories are designed to be considered with equal weight, Configuration C would yield the most balanced choice. However, the significant gap in data available easily induces bias in that case, as trends are difficult to judge based on a few standalone cases, when compared to the thousands of cases that report racial bias. Another model, HistGradientBoosting, is introduced in Configuration D. Nevertheless, the metrics fared worse than those of the Random Forest, hence the latter was nonetheless employed in this investigation.

Table 4. Metrics under configuration E

Bias	Precision	Recall	F1	Support
Disability	0.26	0.18	0.21	367
Gender	0.17	0.30	0.22	306
Race	0.70	0.93	0.80	27,975
Religious Group	0.68	0.40	0.51	8,129
Sexual Orientation	0.55	0.06	0.10	6,835

Table 4 demonstrates that Configuration E is capable of recalling 93% of Racial bias incidents, and predict 70% correctly, yielding the strongest performance. It ourperforms the null Configuration A (F1= 0.78) and the balanced Configuration C (F1 = 0.72).

4. Discussion

Tuning in focus on racial bias, the primary model achieved F1= 0.80 and recall = 0.93. Racially motivated incidents constitute 64.1% of the data set; therefore, a model predicting the classes uniformly would be expected to yield 64% accuracy with zero recall for the other categories. The model built in this study exceeds that baseline substantially, demonstrating that the achieved results are not simply due to majority-class prediction.

DOI: 10.54254/2755-2721/2026.AST35981

Out of the columns, DATA_YEAR was the largest contributor to model prediction accuracy (~23%), followed by offense type and location. All of the above were frequency encoded and held the same significance in the balanced configuration as well. The columns OFFENDER_RACE and VICTIM_TYPES were most susceptible to being dependent on the predicted value; however, the categories combined only accounted for a small share of importance. However, one area of concern is the prevalence of DATA_YEAR, which could be subject to changes in the FBI's classification of terms. Most notably, the 2015 edit to race and ethnicity categories, and the addition of seven sub-categories in religion [6]. Therefore, the model could be inferring predictions based on which values were actually available during the year, rather than identifying relationships. Another explanation would be genuine variation across different years, adhering to political movements and rallies that would have induced variation to bias motivation. The two explanations cannot be separated from this data, as the very limitation is built into the dataset itself, representing a limitation on generalizability.

A data ceiling for the categories of Disability and Gender (1,836 and 1,531 reported incidents, respectively) resulted in poor performance across all model configurations. This is due to a lack of information that no model can account for. A possible explanation is the structural underreporting documented by Pezzella et al., where groups face police barriers or are unwilling to speak on discriminatory violence [1]. The UCR data cannot distinguish between the former explanation and whether these 2 bias motivators are genuinely uncommon compared to other categories.

5. Conclusion

This study analyzes 218,059 hate crimes across all 50 US states in the timeframe 1991-2020, using a Random Forest classifier to predict bias motivations from 13 metrics regarding each incident. The findings are as followed: Bias motivation is predictable from structural details above a majority-class baseline, with Race-category incidents predicted primarily from offense type, location, geography, and date; The dominance of incident year as a predictor could be driven by FBI taxonomy changes rather than inherent patterns, placing a limit on generalizability and reflecting the difficulty of a comprehensive and holistic analysis, as all datasets would have human limits on categorization during reporting or documentation of the crime. Finally, testing various combinations to offset the imbalance of entries belonging to each category quantified the loss of accuracy if prioritizing strictly equal analysis for categories that vary significantly in data points.

The results of this study are bound by structural limitations within the UCR reporting system, reflecting only incidents reported and classified as bias-motivated by local agencies. There exists a significant gap between official records and the true occurrence of hate crimes within daily activity. However, the under-reporting of hate crimes is a societal issue that cannot be completely addressed in the short term. This model maximizes available information on the subject and finds predictability within incidents stored in official records.

References

[1] Pezzella, F. S., Fetzer, M. D., & Keller, T. (2019). The dark figure of hate crime underreporting. *American Behavioral Scientist*. Advance online publication.

[2] Hate Crime Statistics Act of 1990, Pub. L. No. 101-275, 28 U.S.C. § 534. https://www.congress.gov/bill/101st-congress/house-bill/1048

[3] Holder, E., Edwards, B., & Osborne, D. (2022). An exploratory study on the structural and demographic predictors of hate crime across the rural-urban divide. *Victims & Offenders, 19*(3).

DOI: 10.54254/2755-2721/2026.AST35981

[4] Salazar, A. O. (2024). Detecting hate crimes through machine learning and natural language processing. *Police Practice and Research, 25*(6), 746–768.
[5] Davani, A. M., Yeh, L., Atari, M., Kennedy, B., Portillo-Wightman, G., Gonzalez, E., Delong, N., Bhatia, R., Mirinjian, A., Ren, X., & Dehghani, M. (2019). Reporting the unreported: Event extraction for analyzing the local representation of hate crimes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 5753–5757).
[6] U.S. Department of Justice, Federal Bureau of Investigation. (2021). Hate crime data collection: Methodology.
[7] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.