REDUCING THE ETHNIC AND GENDER BIAS WHEN PREDICTING RECIDIVISM WITH LOGISTIC REGRESSION AND RANDOM FOREST

LONG MA

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

1236464

COMMITTEE

dr. Grzegorz Chrupała
dr. Michal Klincewicz

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Tilburg, The Netherlands

DATE

June 24, 2022
Word Count: 8393

ACKNOWLEDGMENTS

A short thank you to all the family and friends that supported me throughout the process of writing this thesis: your love and support is appreciated!

REDUCING THE ETHNIC AND GENDER BIAS WHEN PREDICTING RECIDIVISM WITH LOGISTIC REGRESSION AND RANDOM FOREST

LONG MA

Abstract

In this thesis the ethnic and gender bias is measured in 2 Machine Learning models designed to predict whether a convict will recommit a crime. This was tested on 3 different datasets containing offenders from 3 different states from the United States and whether they recommitted a crime within 3 years. Previous papers have focused on either achieving a high accuracy predicting recidivism or by testing bias reduction methods. This thesis compares a high accuracy model (Random Forest) with a traditional model (Logistic Regression) and tests 5 bias reduction methods in order to measure what methods can reduce the bias while maintaining a high accuracy. Three public datasets have been used in this thesis with features such as gender, race, historic arrest record, education level or income to predict whether an offender will reoffend or not. The results show that LR can outperform RF with certain datasets and/or mitigation methods, although RF generally performs better on accuracy and bias in this thesis. Both models have their own mitigation methods that are more effective. Logistic Regression had the best results, with the Balanced Gender Model and the Gender Eliminated Model. Random Forest had the best results with the Proxy-Eliminated Gender Model. The bias mitigation methods seem to have more effect on the Random Forest Model, by outscoring the base model more often than the Logistic Regression Model after using them. The characteristics of a dataset furthermore also influenced how effective the bias mitigation methods were. Lastly the bias mitigation methods were much more successful on gender based models than on race based ones.

Data Source: The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. The author of this thesis acknowledges that they do not have any legal claim to this data or code.

1 INTRODUCTION

Machine learning is becoming more relevant in the field of predicting recidivism: even simple ML-models can now outperform specific software made to predict recidivism as was recently demonstrated in a paper (Dressel & Farid, 2018) by getting a higher accuracy with two Logistic Regression models compared to the widely used Correctional

Offender Management Profiling for Alternative Sanctions (COMPAS) software (p. 1). Ethnic bias in ML-models has made the news a couple of times recently. The Dutch IRS for example had to pay a €2.75 million fine in 2021 for using ML-models that were discriminatory. This thesis will investigate how biased a more accurate model is in the field of predicting recidivism. The assumption is that models with a higher accuracy are more useful in practice when not accounting for bias.

In the past five years several methods have been used to lower the bias from ML-models for predicting recidivism with varying degrees of success. Several papers have also been written on how accurate ML-models can predict recidivism. By combining the findings of these papers published in the past five years, this thesis will try to add new findings to the existing scientific discussion on predicting recidivism and the reduction of bias. A comparison will be made between two Machine learning models (ML-models), namely a Logistic Regression model (LR) and a Random Forest model (RF). These two models will be tested on three publicly available datasets that contain information on (re)offenders and several characteristics such as age, gender, ethnicity, crime committed and crime type. In the literature on predicting recidivism with ML-models, LR is usually used as a baseline model and RF is used in papers that aim for a high accuracy in their models (see subsection 2.2). By measuring the bias in both, this thesis aims to compare the results of a RF model to a LR model on multiple datasets. Another important aspect this thesis will research is how the five bias reduction methods discussed in this thesis will work on the two different models.

The three datasets used in this thesis have different amount of features ranging from 11 (dataset 2) to 50 (dataset 3) after data processing and some differences in the distribution of ethnicities and minor differences in the distribution of genders and some differences in the percentage of recidivists (see subsection 3.1) By testing the reduction methods on different datasets, the importance of distribution and the number of features can be taken into account as well.

The main research question that this thesis is going to answer is 'To what extent can the bias be lowered while maintaining the accuracy when predicting recidivism with a Random Forest and a Logistic Regression model?'. The main question is not easily answered with a single paragraph, so it will be divided into three sub-questions:

1. How accurate can Logistic Regression and Random Forest predict recidivism in the three datasets?

2. How do the five bias reduction methods individually influence the prediction accuracy and bias of the Logistic Regression and Random Forest models?

3. What bias reduction methods had the best results on which model? And on which dataset?

The Random Forest model outperformed the Logistic Regression model in general as expected, but the LR model scored better on the Iowa dataset (dataset 2) almost every time. After using the bias mitigation methods it became clear that some methods only had impact on 1 model and barely any effect on the other. This was true for the most successful mitigation method for both LR and RF. Surprisingly some mitigation methods were so successful that they improved 2 out of three metrics or even all of the metrics compared to the baseline results. For more detail on this see section 4 Results.

The thesis will be structured as follows: in section 2.1 the literature written on predicting recidivism and bias will be discussed. In 2.2 the literature on predicting recidivism and Machine Learning will be discussed. In 2.3 some recent bias mitigation methods will be explained. Lastly in 2.4 the focus will be on the research gaps left in the literature discussed in section 2.

In section 3 the experimental setup will be discussed which consists of two subsections: subsection 3.1 will be about the three recidivism datasets and subsection 3.2 will explain more about the two Machine Learning models and the evaluation metrics.

In section 4.1 the results of the baseline models will be shown and discussed, after which the results of the bias mitigation methods will be shown in a subsection in 4.2 per individual dataset and ML-model. These results will be then combined and examined per dataset in section 4.3. The results will then be examined when purely looking at the results per ML-model in section 4.4.

In section 5 there will be room for some discussion on the results of this thesis and how it relates to the literature from section 2. The societal relevance of the findings in this thesis will explained. Furthermore the limitations of this thesis will be discussed.

Finally in section 6, a conclusion will be drawn based on the previous sections. Finally some recommendations will be made for future research.

2 RELATED WORK

The notion that a machine can assist with the assessment of whether a criminal will reoffend or not started with software specifically designed for this purpose. In 2000 The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) was first released and used to determine the likelihood of a criminal reoffending (Dressel & Farid, 2018). In the past two decades this role has been taken over by Machine Learning based tools in countries such as the US, the UK, Canada & Spain because of the access to much more powerful computers and the availability of much more data (Haghighi & Castillo, 2021). In the first subsection the relationship between predicting recidivism and bias will be discussed. In the second subsection Machine Learning Models will be discussed that are used to predict recidivism. The next subsection will be about bias mitigation methods for Machine Learning Models that predict recidivism. Finally in the subsection 2.4, the research gaps will be discussed.

2.1 *Predicting Recidivism and Bias*

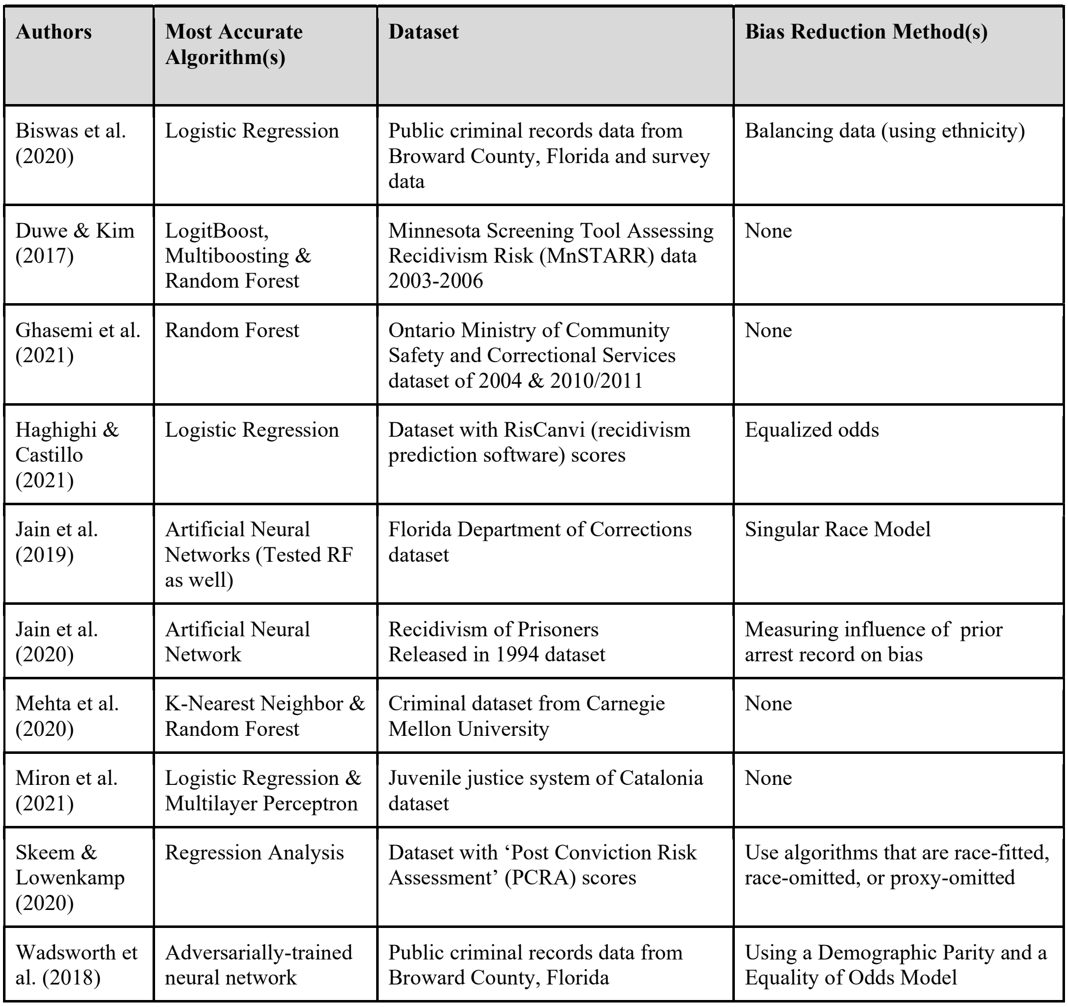
In an article about machine bias in 2016 in ProPublica (Angwin, Larson, Mattu, & Kirchner, 2016), the authors showed that the COMPAS software could be ethnically biased. The software predicted that black offenders were more likely to reoffend by giving white offenders lower risk scores in general. This same article showed that a higher percentage of the black offenders from this dataset received a high risk assessment, but did not reoffend again At the same time white offenders with a low risk assessment had a higher percentage of reoffenders. For this thesis the evaluation of the bias is important. Bias can have several meanings, but in the context of this thesis it is a fairness metric. Fairness can be defined in two ways that is relevant for this thesis: 1) As predictive parity which looks at the True Positive Rate (TPR) for the race and gender categories or 2) As equalized odds which focuses on the False Positive Rate (FPR) for these two categories (Biddle, 2020). A high TPR would be desirable because this means more correctly labeled offenders and a low FPR would be desirable because this means fewer wrongly labeled offenders. A ethnic bias however, can occur even before the use of software such as COMPAS or the use of a ML-model. (Danks & London, 2017) distinguish three stages in which bias can occur:

1. Bias in the training data (e.g when mining data)
2. Bias in the algorithmic focus (e.g. feature selection)
3. Bias in the algorithmic processing (e.g. choosing a algorithm and parameters)

A bias in stage 1 could be present in the datasets used for this thesis. When making arrests, the police can be more active in neighbourhoods crime is expected to happen, which leads to more patrolling in that neighbourhood, which leads to more arrests, which leads to more patrolling and so on: a phenomenon called a 'runaway feedback loop' ((Ensign, Friedler, Neville, Scheidegger, & Venkatasubramanian, 2018)). This could lead to a disproportional amount of offenders from unsafe neighbourhoods which can be an ethnic minorities. Due to time constraints, this thesis will not investigate this type of bias. It would take a considerable amount of time to research how much bias has been introduced in stage 1 of one dataset, let alone three. Stage 2 is for example which characteristics are used as input feature to assess whether or not a offender will commit recidivism. The inclusion of race as a feature can directly lead to a bias, but indirect information connected to race such as the police precinct where the arrest happened can be biased if that happens to be a precinct in a neighbourhood of mostly ethnic minorities. For this thesis, stage 2 is also mostly ignored because of the use of 3 publicly available datasets and not making a choice which questions are asked after a arrest (i.e. determining which features will be featured in the dataset). All the features from the datasets will be used as input for the algorithms, aside from the features that are directly linked to whether or not a offender committed recidivism as using these features would give the model the target label as an input. And features that were generated with risk assessment software will also be removed because the performance of the ML-models would otherwise be diluted, Unfortunately this means that biases in stage 1 and stage 2 will not be mitigated in this thesis, as it will solely focus on stage 3. In the next subsection it will be explained which algorithms are chosen for this thesis and why. In section 4.1 more will be explained about the parameters of the models used in this thesis.

2.2 *Predicting Recidivism with Machine Learning Models*

Several papers have been published in the field of predicting recidivism with ML-models. Figure 1 shows an overview of 10 papers from the past five years. A distinction can be made between papers that aim to generate a high accuracy when predicting which criminals are going to reoffend (these papers have 'None' mentioned in the column of 'Bias Reduction Method(s)') and papers that make predictions, but focus mainly on the reduction of bias. There are six papers that focus on reducing bias, three of which utilize simpler algorithms such as Logistic Regression and Regression Analysis (Biswas, Kolczynska, Rantanen, & Rozenshtein, 2020; Haghighi & Castillo, 2021; Skeem & Lowenkamp, 2020).



Authors	Most Accurate Algorithm(s)	Dataset	Bias Reduction Method(s)
Biswas et al. (2020)	Logistic Regression	Public criminal records data from Broward County, Florida and survey data	Balancing data (using ethnicity)
Duwe & Kim (2017)	LogitBoost, Multiboosting & Random Forest	Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR) data 2003-2006	None
Ghasemi et al. (2021)	Random Forest	Ontario Ministry of Community Safety and Correctional Services dataset of 2004 & 2010/2011	None
Haghighi & Castillo (2021)	Logistic Regression	Dataset with RisCanvi (recidivism prediction software) scores	Equalized odds
Jain et al. (2019)	Artificial Neural Networks (Tested RF as well)	Florida Department of Corrections dataset	Singular Race Model
Jain et al. (2020)	Artificial Neural Network	Recidivism of Prisoners Released in 1994 dataset	Measuring influence of prior arrest record on bias
Mehta et al. (2020)	K-Nearest Neighbor & Random Forest	Criminal dataset from Carnegie Mellon University	None
Miron et al. (2021)	Logistic Regression & Multilayer Perceptron	Juvenile justice system of Catalonia dataset	None
Skeem & Lowenkamp (2020)	Regression Analysis	Dataset with 'Post Conviction Risk Assessment' (PCRA) scores	Use algorithms that are race-fitted, race-omitted, or proxy-omitted
Wadsworth et al. (2018)	Adversarially-trained neural network	Public criminal records data from Broward County, Florida	Using a Demographic Parity and a Equality of Odds Model

Figure 1: Ten papers on predicting recidivism with Machine Learning from 2017 and onwards

When looking at the four other papers mainly focused on accuracy. three of them achieved the highest prediction accuracy using a Random Forest (RF) model (Duwe & Kim, 2017; Ghasemi et al., 2021; Mehta, Shah, Patel, & Kanani, 2020; Miron, Tolan, Gutiérrez, & Castillo, 2020), whilst only one out of the six papers focused on bias have tried reducing the bias in a Random Forest model (Jain, Huber, Fegaras, & Elmasri, 2019). In a paper focused on testing multiple algorithms, the authors (Duwe & Kim, 2017) used 12 supervised learning models on a dataset with offenders from Minnesota. More traditional methods such as Logistic Regression performed more poorly in comparison to newer techniques in this research. LogitBoost and Random Forest obtained the highest accuracy scores in this paper. LogitBoost is a method that 'boosts' weaker learners or models. The writers do state that the difference in accuracy between the best and the worst models is 'relatively modest'.

In a research on a Canadian dataset with offenders from Ontario, Canada (Ghasemi et al., 2021) the authors tested the prediction accuracy of 3 ML-models: Decision Trees (DT), RF and Support Vector Machines (SVM). RF scored the best out of these three. In a paper focused on getting a high accuracy when predicting recidivism (Mehta et al., 2020), the researchers tested K-Nearest Neighbor (KNN) RF and LR on a dataset with offenders from the United States of America. The highest accuracy was achieved by RF again. In another paper (Miron et al., 2020) the results of 7 ML-models were compared to the results of software called 'Structured Assessment of Violence Risk in Youth' (SAVRY). The models used in this research were LR, Multi-Layer Perceptron (MLP), SVM, KNN, RF, DT and Naive Bayes (NB). In this research by Miron et al. the best results were achieved by LR, MLP and SVM. It should be noted that this research used Area Under the Curve instead of accuracy and made big changes in the dataset by encoding sensitive information and by 'equalizing the base rates' (p. 125). Overall RF performed the best when used to predict recidivism on certain datasets.

2.3 *Predicting Recidivism with Machine Learning Models and Bias Reduction*

The main focus of this thesis is not the accuracy of ML-models when predicting recidivism, but on reducing the bias in ML-models. In recent years authors have focused on this part of machine learning.

In a research focused on balancing data (Biswas et al., 2020) between black and white offenders, Logistic Regression was used on a dataset in which the ethnicities were balanced during the data processing. The datasets were seperated in race. The higher amount of offenders in the categories was reduced to the same amount of number as the category with less offenders at random. The results were compared to the scores of a LR model on an unchanged dataset. The results in this paper show that the use of a balanced dataset can lead to an increase in 'equalized odds' (see Section 3.2), but can also lead to a decrease in accuracy.

In another research (Haghighi & Castillo, 2021) the Area Under the Curve (AUC) was compared of the recidivism prediction software RisCanvi and 2 ML-models (MLP and SVM). The results were that both ML-models outperformed RiskCanvi on AUC-scores, but that the higher accuracy could also lead to a higher bias. The models were calibrated to reduce the bias by using a relaxation method.

In a paper focused on using a singular race model (Jain et al., 2019), the authors attempted to increase the fairness when predicting recidivism while also maintaining the accuracy. The algorithms used in this research were Artificial Neural Networks (ANN), KNN, RF, AdaBoost (an algorithm that uses weighted sums to increase the predictive power of weak learners), DT and SVM. They separated the races during the data processing and trained and tested the models on a single race and compared these results to models tested on an unchanged dataset. The results were that all singular race models yielded a higher accuracy than that of the baseline model tested on all ethnicities. The increase in accuracy could however sometimes also lead to an increase in bias.

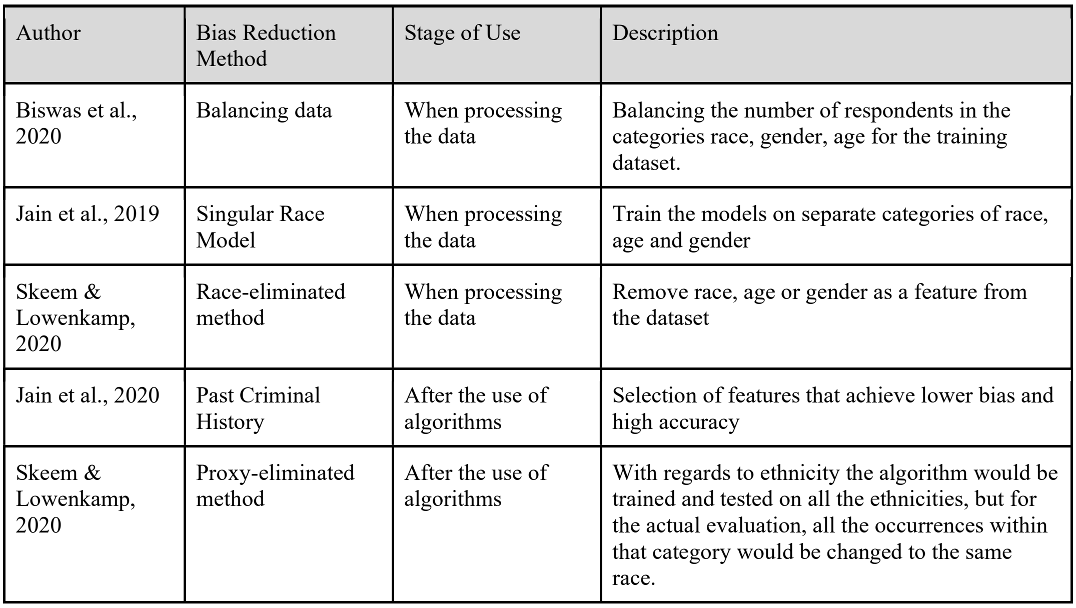
In a research also focused on bias mitigation, the authors (Jain, Huber, Elmasri, & Fegaras, 2020) used several models that use the past criminal information of the offenders, but each model selected different features based on different number of arrest cycles. The accuracy and the bias of the different models are then compared and the best performing model is picked.

Three ways of debiasing were tested by Skeem and Lowenkamp (2020) in their paper:

1. The "proxy eliminated" algorithm in which the algorithm is trained with all the features, but afterward the variation in the sample is removed for the feature that has to be debiased (for example the model is trained on a train set containing all the races, but then all races are turned into the dominant race for the test set)

2. The "race eliminated" algorithm in which the feature that has to be debiased is removed

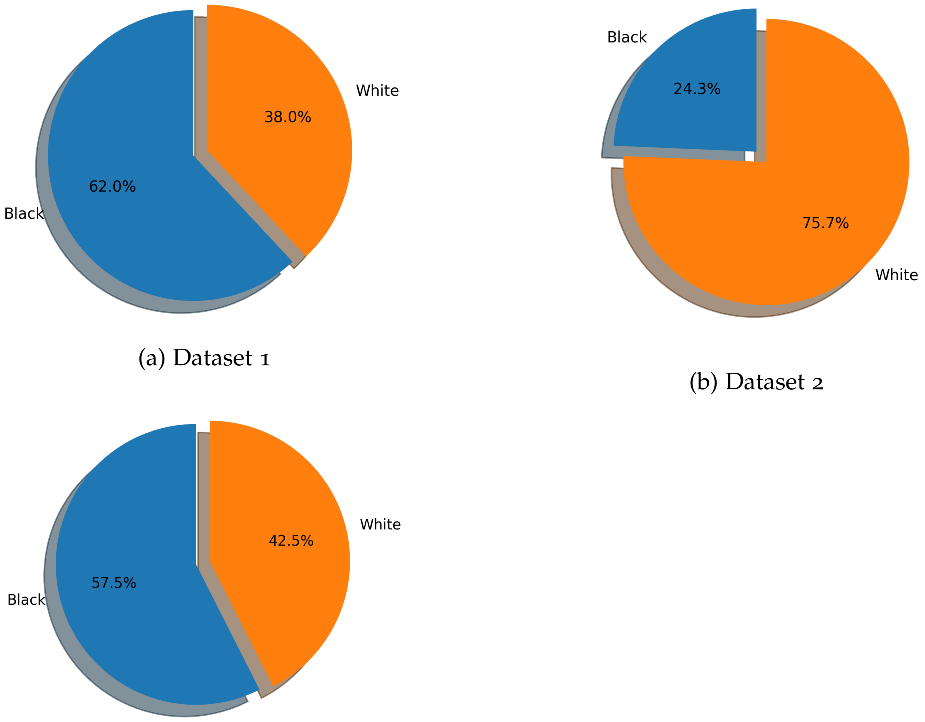
3. The "criminal history discount" in which the scores of the group prone to bias are reduced to lower the chance of a bias in the algorithm.



Author	Bias Reduction Method	Stage of Use	Description
Biswas et al., 2020	Balancing data	When processing the data	Balancing the number of respondents in the categories race, gender, age for the training dataset.
Jain et al., 2019	Singular Race Model	When processing the data	Train the models on separate categories of race, age and gender
Skeem & Lowenkamp, 2020	Race-eliminated method	When processing the data	Remove race, age or gender as a feature from the dataset
Jain et al., 2020	Past Criminal History	After the use of algorithms	Selection of features that achieve lower bias and high accuracy
Skeem & Lowenkamp, 2020	Proxy-eliminated method	After the use of algorithms	With regards to ethnicity the algorithm would be trained and tested on all the ethnicities, but for the actual evaluation, all the occurrences within that category would be changed to the same race.

Figure 2: 5 bias reduction methods used in this thesis

Other authors (Wadsworth, Vera, & Piech, 2018) used two adversarial networks: a Demographic Parity and a Equality of Odds Model to reduce the bias generated in the prediction model. They used the same COMPAS dataset (dataset 1: Florida) that will be used for this thesis. From of all these bias mitigation methods, 5 were suited for both the LR-model and the RF-model and the datasets (classification based) used in this thesis. Figure 2 gives an overview of the 5 methods of bias reduction that will be used in this thesis and in which stage the method will be used.



(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

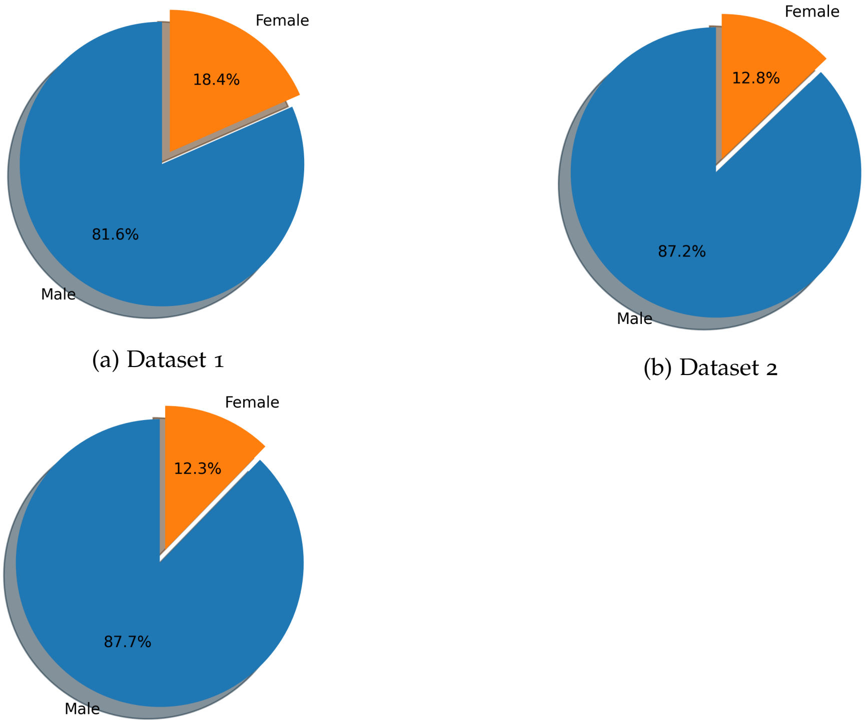
Figure 3: Race distribution

2.4 *Research Gaps in Related Work*

Most papers did their research on only 1 dataset. This can make it more difficult to reproduce results on other datasets because of differences in the distribution of races or genders. Furthermore the features used to make the predictions can vary greatly. Before data processing, two datasets used in this thesis had more than 50 features in comparison to the other dataset that had only 17 features. One can imagine that a model can produce different results when testing on a dataset with a lot of features compared to a dataset with very few features. For this reason, this thesis will use 3 datasets that have a different distribution of races (see Figure 3), some differences in the distribution of genders (see Figure 4) and differences in the distribution of recidivists (see Figure 5).

Another issue is that the papers focused on getting a high accuracy had high prediction scores with newer models such as RF. The papers in which bias mitigation techniques were tested usually used simpler models such as LR. This thesis will test the mitigation techniques on a simpler and a more advanced model, so a conclusion can be drawn whether the mitigation techniques can work on both ML-models. Ideally one ML-model would score higher on prediction accuracy and lower on bias making it more usable in practice. This thesis will attempt to breach this gap in research.

To reduce the bias, the reduction methods shown in Figure 2 will be used. A selection has been made of the reduction methods mentioned in Figure 1 that are suitable for the three recidivism datasets used in this thesis and for use with a LR and

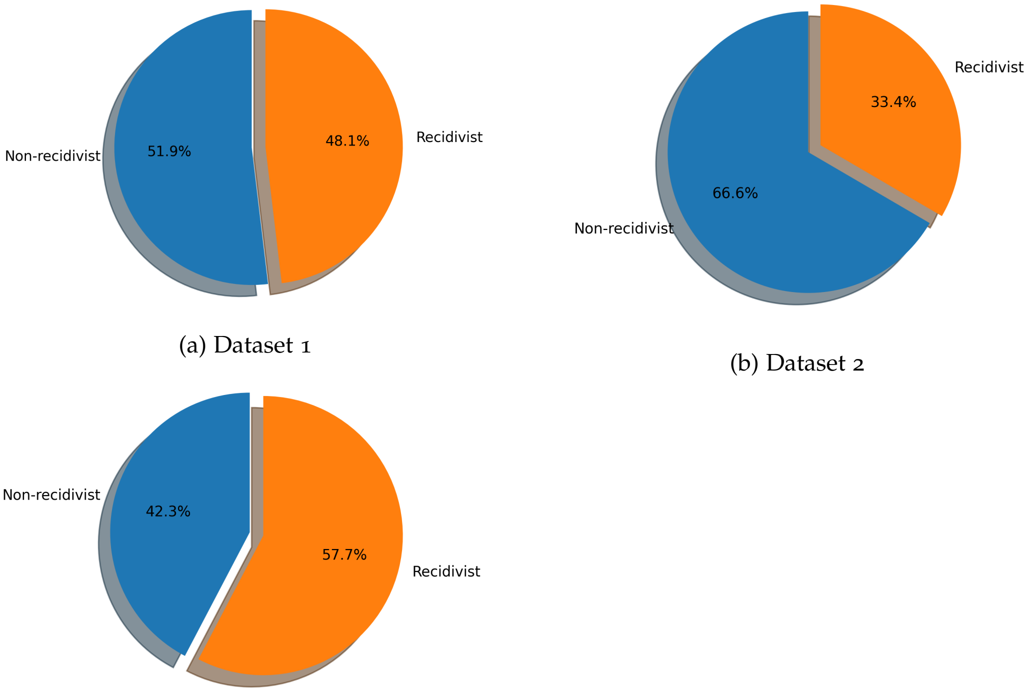


(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

Figure 4: Gender distribution



(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

Figure 5: Recidivist distribution

RF algorithms. A distinction is made in the stage of use of the bias reduction methods to formulate the research questions.

This thesis will address the research gap of not using bias reduction methods on the most accurate ML-model for predicting recidivism. Furthermore, the amount of bias reduction methods used in the individual papers from Figure 1 were limited. By trying a multitude of methods, this thesis can determine what reduction methods can achieve a high accuracy and a low bias when predicting recidivism.

3 EXPERIMENTAL SETUP

This section is divided into two subsections. In the first subsection 3.1 Data, the three datasets will be described and what data processing has been done per dataset. In the second part 3.2 Methods and Models, the ML-models will be explained, what parameter tuning has been done and how the results will be evaluated.

3.1 *Data*

For this thesis three datasets will be used from the United States that all contain offenders and whether or not they recommitted a crime within 3 years (recidivism). Recidivism will be the target label for all three datasets. None of the datasets were specifically mined for this thesis, but are publicly available online.

In the category gender, all the datasets only included male or female offenders. In the category race however, the datasets were more complex. Dataset 1 has 9791 black offenders, 6086 white and 2439 non-white/non-black races. The other races include 1451 Hispanic offenders, 860 'other', 71 Asian and 57 Native American. All the non-white/non-black races account for 13.31% of the total. When removing the 'other' category, these races only account for 8.62% of the total. Dataset 2 has 19118 white offenders, 6148 black and 724 other races (Native American, Asian or Alaska Native) that account for only 2.78% of the total. Dataset 3 only includes black and white offenders. Given the low amount of non-white/non-black offenders in dataset 1 and 2 and none in dataset 3, the choice was made to only include black and white offenders for this thesis.

1) Florida dataset: The first dataset is a csv file that consists of conviction records from Broward County, Florida (USA) and contains 18,316 offenders with 52 features before data processing. The amount of features is high in this dataset, because it contains the scores from the recidivism prediction software COMPAS as well as characteristics of the offenders such as juvenile record, gender, age and crime committed. The data was collected from offenders that committed their first crime in 2013 or 2014. The dataset is publicly available on Kaggle and Github. From this dataset 30 features were removed. This includes the name, features related to the scoring system of the COMPAS software and characteristics on reoffending (because this is directly linked to whether or not an offender recommitted a crime). Categorical NA values were changed to the string 'None'. Numerical NA values were changed to the mean value of that specific column in this dataset. After data processing 61.96% of the offenders is black in this dataset, 81,60% is male and 48.10% committed recidivism from all the offenders.

After data processing this dataset contains 17.496 offenders with 21 features.

2) Iowa dataset: The second dataset used for this thesis is a CSV file from the Department of Corrections (DOC) from Iowa (USA). It has 26,020 offenders with 17 features before data processing such as release type, main supervising district and what offense was committed. 8,681 out of these offenders have committed recidivism within 3 years of their first offense. The dataset contains offenders that committed their first offense in 2013 to 2015. This dataset is publicly available on the website of the DOC Iowa. From this dataset 6 features were removed. This includes features on reoffending, because this is directly linked to whether or not an offender recommitted a crime. Categorical NA values were changed to the string 'None'. The Numerical NA values were changed to the mean value of that specific column in this dataset. After data processing 75.66% of the offenders is black in this dataset, 87.16% is male and 33.40% committed recidivism from all the offenders. After data processing this dataset contains 25,987 offenders with 11 features.

3) Georgia dataset: The third dataset is a CSV file from the Georgia Department of Community Supervision and contains 25,835 offenders and has 51 features before data processing. This is the most comprehensive dataset when taking features into account, because the first dataset included COMPAS scores which will mostly be discarded for this thesis. Example of features that are not in the other datasets are education level, results of mandatory drug tests (weed, cocaine and methamphetamine), employment info, information on prior arrests, if the offender is gang affiliated etc. The data was collected from offenders that committed their first crime in 2013 to 2015. From this dataset 3 features were removed because they were directly linked to whether or not an offender recommitted a crime. Categorical NA values were changed to the string 'None'. The Numerical NA values were changed to a the mean value of of that specific column in this dataset. After data processing 57.46% of the offenders is black in this dataset, 87.74% is male and 57.68% committed recidivism from all the offenders. After data processing this dataset contains 25,835 with 50 features.

For a complete overview of all the features of the three datasets, a short description and the datatype per feature, see Appendix A.

3.2 *Method and Models*

In the background section of this proposal, Random Forest yielded the highest accuracy most often when used for predicting recidivism. This algorithm will be compared to the relatively older method of Logistic Regression. LR has been used as a baseline ML-model that can outperform existing software specifically designed for predicting recidivism (Dressel & Farid, 2018). Logistic Regression will also be the baseline model in this thesis and will be compared to a Random Forest model. The accuracy and the fairness will be measured in these two models. To run the models, the programming language Python will be used. The Numpy (Harris et al., 2020) and Pandas (McKinney et al., 2010) packages for Python will be used for data processing. The Logistic Regression and Random Forest models will be used from the Scikit-Learn package (Pedregosa et al., 2011) to analyse the datasets. The confusion matrices (see Appendix C) were generated using the Matplotlib package (Hunter, 2007).

When evaluating the final results, the accuracy of the ML-model will be taken into account as well because a model will have limited practical use when it is unable to predict recidivism accurately. The bias reduction methods that lead to no or to a limited decrease in accuracy, but also cause an increase in fairness are deemed the most successful. To determine how much a 'limited decrease in accuracy' is, all the results on accuracy and bias of the reduction methods will have to be compared.

For the Logistic Regression models, there was limited parameter tuning. As stated in another article, the possibility of tuning hyperparameters is limited for LR (Brownlee, 2019). The only tuning that was done, was changing the number of iterations until the maximum accuracy was reached for the base model.

To optimize the performance of the Random Forest model, three parameters were tuned that are associated with an improvement in predicting power: 1) max_features 2) n_estimators 3) min_sample_leaf (Srivastava, 2015). But as cited in another paper (Probst, Wright, & Boulesteix, 2019), the basic model of RF already performs well and it is hard to gain a significant increase in performance with tuning. The tuning of max_features yielded the best scores in accuracy and TPR and FPR with lower numbers. The tuning of max_features started at 0.8, and ended at 0.05. For the min_sample_leaf the starting number 50 as was recommended (Srivastava, 2015), but the results were far below the baseline scores. Turning it down to 10 and eventually 1 improved all the three metrics, but still well below the baseline, with an accuracy drop of 7% and a TPR drop of about 10%. For this thesis the main increase in performance for Random Forest comes from adjusting the number of trees in the model, which was changed from 100 to 150 in the models.

For this thesis the evaluation of the bias is important. Bias can have several meanings, but in the context of this thesis it is a fairness metric. Fairness can be defined in two ways that is relevant for this thesis: 1) As predictive parity which looks at the True Positive Rate (TPR) for the race and gender categories or 2) As equalized odds which focuses on the False Positive Rate (FPR) for these two categories (Biddle, 2020). Both these definitions will be used to evaluate the fairness of the models.The accuracy in predicting the binary target label of recidivism will be used as an output as well. To measure the accuracy, the amount of correctly labeled offenders is divided by the total number of predictions. This way of calculating the accuracy has been used in past research as well (Ghasemi et al., 2021). This means that the output of the LR and RF-models in subsections 4.1 and 4.2 will be shown with three metrics: 1) The accuracy: refers to the percentage of correctly predicted labels 2) The True Positive Rate (TPR): refers to the percentage of offenders that correctly have received the label recidivist out of all the recidivists 3) The False Positive Rate (FPR): rate refers to the percentage that were wrongly labeled as recidivist from all the offenders that were given the label of recidivist. The accuracy is an important metric to evaluate a model. For fairness, looking at points two and three are also very important. It cannot be understated how important FPR and TPR are in real life. TPR represents people that should have been treated with care as they recommitted a crime. FPR represents the part of the population that is falsely believed to be recommitting a crime within 3 years. This explanation hopefully shows the importance of using the accuracy next to the other two metrics when it comes to predicting recidivism and measuring the bias. To divide the datasets in a train and test sets, a division of 70% to 30% was used.

	Accuracy	**TPR**	**FPR**
Dataset 1: LR Base Model	75.29%	77.39%	26.90%
Dataset 1: RF Base Model	85.50%	86.31%	15.34%
Dataset 2: LR Base Model	67.66%	93.42%	83.87%
Dataset 2: RF Base Model	65.12%	80.28%	65.20%
Dataset 3: LR Base Model	71.47%	58.90%	19.19%
Dataset 3: RF Base Model	73.20%	58.47%	15.86%

Table 1: Results of the base models

In the next section, the results of the bias mitigation methods will be shown. The trade-off between accuracy and bias are inherent to these mitigation methods (Kleinberg, Mullainathan, & Raghavan, 2016; Pleiss, Raghavan, Wu, Kleinberg, & Weinberger, 2017; Skeem & Lowenkamp, 2020), so it is up to the people involved in the assessment of recidivism how to handle the tradeoffs. For this thesis the results will be displayed and the most effective methods will be pointed out. Any decisions on sacrificing accuracy for bias or vice versa will not be made.

4 RESULTS

In this paragraph the results of the bias mitigation methods will be compared to the results of the base models for both the LR- and the RF-models. Each bias mitigation method has been tested on both race and gender on all three datasets. For the results, a high accuracy and a high TPR are considered good, while a high FPR is considered to be bad since this means that more offenders have been mislabeled as a potential recidivist.

Three Bias mitigation methods had separate outputs for both genders and races, so the mean was taken of these scores for the tables used in this subsection 4.2. This regards the Singular Race Model, the Past Criminal History Model and the Proxy-Eliminated Model. For the separate scores on race and gender, the full percentages per bias mitigation method are included in Appendix B.

In this section, the results of the base models will be discussed first. Then the general results on each of the three datasets will be discussed in subsection 4.2. After this, the difference in performance between the bias mitigation methods will be discussed in subsection 4.3. Finally the difference in results of the bias mitigation methods on the LR-model and the RF-model will be discussed in 4.4.

4.1 *Base Models*

The highest score in accuracy, TPR or FPR between the LR and RF model is colored green. When looking at the results of the two base models (See Table 1), there is a big difference performance. RF is almost 10% more accurate than LR on the first dataset and also scores about 9% higher on the True Positive Rate. The False Positive Rate is almost 11% higher in the LR-model. The second dataset has considerably less features

	Accuracy	TPR	FPR
Base Model	**75.29%**	**77.39%**	**26.90%**
Singular Race Model	75.06%	74.31%	25.67%
Singular Gender Model	75.52%	78.95%	30.36%
Balanced Race Model	74.29%	73.86%	25.21%
Balanced Gender Model	76.24%	78.65%	26.77%
Race Eliminated Model	75.25%	77.35%	26.94%
Gender Eliminated Model	75.29%	77.32%	26.83%
Past Criminal History Race Model	74.83%	74.07%	25.78%
Past Criminal History Gender Model	75.11%	78.56%	30.81%
Proxy-Eliminated Race Model	75.65%	72.93%	21.47%
Proxy-Eliminated Gender Model	74.66%	78.06%	28.98%

Table 2: Dataset 1: Logistic Regression results

after data processing (11) compared to the first dataset (21) and the accuracy of both models is much worse. LR does score high on TPR, but has a much worse FPR than RF because a low number on this metric is better. The results on dataset 3 are much closer with RF scoring slightly better on accuracy and FPR and only 0.43% worse on TPR. It is noticeable that the FPR is very high for both ML-models on dataset 2.

It seems that RF scores better on two datasets than LR and also scores better on FPR on all three datasets. LR scores better on TPR on the 2nd dataset and also has a higher accuracy on this dataset. In total Random Forest scores better on 6 out of 9 outputs in comparison to LR that outscores only on 3 out of 9 outputs. This would make RF the better performing model, at least in this research setting. The results of these baseline models will be compared in the following subsection to the bias mitigated models for race and for gender.

4.2 *General Results by Dataset*

In this subsections the tables will have a similar design. Each table starts with the scores of the base model in bold text. If a bias mitigation method outperforms the base model on any metric, the score is displayed in green. This subsection contains a table for each dataset and ML-model with the scores shown per mitigation method.

Looking at the scores of the bias mitigation methods on the first dataset using LR in Table 2, the Balanced Gender Model scores better than the base model on all fronts, which is an impressive result. The Singular Gender Model, the Gender Eliminated Model and the Proxy-Eliminated Race Model outscore the base model on 2 of the 3 metrics. In all three of these cases, the overall accuracy improves but the score in the third metric is worse. In total the bias mitigation methods outscore the base LR-model 14 out of 30 times.

Looking at the scores of the bias mitigation methods on the first dataset using RF in table 3, the Proxy-Eliminated Race Model and the Proxy-Eliminated Gender Model outscore the base model on all fronts. Only the Singular Gender Model managed to outscore the base model on 2 out of three metrics, but at the cost of a worse FPR. The bias mitigation methods outscore the base RF-model 11 out of 30 times in total.

	Accuracy	**TPR**	**FPR**
Base Model	**85.50%**	**86.31%**	**15.34%**
Singular Race Model	84.93%	66.15%	16.90%
Singular Gender Model	85.54%	88.03%	18.83%
Balanced Race Model	83.25%	83.94%	17.52%
Balanced Gender Model	80.64%	81.73%	20.72%
Race Eliminated Model	84.35%	85.52%	16.51%
Gender Eliminated Model	85.36%	86.49%	15.80%
Past Criminal History Race Model	84.92%	84.08%	15.11%
Past Criminal History Gender Model	84.87%	87.89%	20.31%
Proxy-Eliminated Race Model	88.45%	88.31%	11.41%
Proxy-Eliminated Gender Model	89.52%	90.56%	11.59%

Table 3: Dataset 1: Random Forest results

	Accuracy	**TPR**	**FPR**
Base Model	**67.66%**	**93.42%**	**83.87%**
Singular Race Model	65.87%	94.21%	87.91%
Singular Gender Model	68.76%	89.04%	89.69%
Balanced Race Model	66.68%	91.51%	81.76%
Balanced Gender Model	69.88%	95.33%	85.69%
Race Eliminated Model	67.61%	93.19%	83.56%
Gender Eliminated Model	75.29%	77.32%	26.83%
Past Criminal History Race Model	65.87%	94.21%	87.91%
Past Criminal History Gender Model	69.47%	96.99%	94.03%
Proxy-Eliminated Race Model	66.40%	95.01%	87.23%
Proxy-Eliminated Gender Model	66.95%	92.50%	83.05%

Table 4: Dataset 2: Logistic Regression results

	Accuracy	TPR	FPR
Base Model	**65.12%**	**80.28%**	**65.20%**
Singular Race Model	63.18%	67.91%	78.31%
Singular Gender Model	65.78%	80.90%	69.58%
Balanced Race Model	64.10%	78.76%	64.48%
Balanced Gender Model	64.98%	81.06%	70.11%
Race Eliminated Model	64.90%	79.82%	64.93%
Gender Eliminated Model	65.42%	80.63%	65.01%
Past Criminal History Race Model	63.14%	80.32%	69.43%
Past Criminal History Gender Model	66.76%	82.64%	70.17%
Proxy-Eliminated Race Model	70.66%	87.18%	60.30%
Proxy-Eliminated Gender Model	73.06%	85.33%	50.94%

Table 5: Dataset 2: Random Forest results

	Accuracy	TPR	FPR
Base Model	**71.47%**	**58.90%**	**19.19%**
Singular Race Model	70.13%	55.69%	19.33%
Singular Gender Model	68.78%	62.07%	27.27%
Balanced Race Model	67.84%	50.92%	19.60%
Balanced Gender Model	69.38%	66.70%	28.15%
Race Eliminated Model	70.95%	59.93%	20.85%
Gender Eliminated Model	75.29%	77.32%	26.83%
Past Criminal History Race Model	71.37%	58.48%	19.19%
Past Criminal History Gender Model	69.67%	63.01%	26.67%
Proxy-Eliminated Race Model	66.62%	48.77%	21.24%
Proxy-Eliminated Gender Model	67.56%	51.13%	19.83%

Table 6: Dataset 3: Logistic Regression results

Looking at the scores of the bias mitigation methods on the second dataset using LR in table 4, no model managed to outperform the base model on all three metrics. There are however 3 models that managed to outscore the base model on two out of three metrics: the Balanced Gender Model, the Gender Eliminated Model and the Past Criminal History Gender Model. The accuracy of the Gender Eliminated Model is even an impressive 7.63% higher, the FPR is improved by 57.04%, although the TPR suffers a great drop of 16.10%. In total the bias mitigation methods outscore the base LR-model 13 out of 30 times.

Looking at the scores of the bias mitigation methods on the second dataset using RF in table 5, the Gender Eliminated Model, the Proxy-Eliminated Race Model and the Proxy-Eliminated Gender Model outperform the base model on all fronts. Singular Gender Model and the Past Criminal History Gender Model both outperform the base model on two out of three metrics. The increase in accuracy of the the Proxy-Eliminated Race Model and the Proxy-Eliminated Gender Model is remarkable with an increase of 5.54& and 7.94% respectively. In total the bias mitigation methods outscore the base RF-model 17 out of 30 times which is the highest number out of all tables.

Looking at the scores of the bias mitigation methods on the third dataset using LR in table 6, no model managed to outperform the base model on all three metrics. Only the Gender Eliminated Model outperformed the base model on 2 metrics. In

	Accuracy	TPR	FPR
Base Model	**73.20%**	**58.47%**	**15.86%**
Singular Race Model	72.55%	58.95%	13.60%
Singular Gender Model	70.14%	61.81%	24.92%
Balanced Race Model	73.09%	57.72%	15.50%
Balanced Gender Model	70.22%	63.18%	23.30%
Race Eliminated Model	73.38%	58.54%	15.59%
Gender Eliminated Model	73.46%	58.44%	15.39%
Past Criminal History Race Model	72.71%	57.16%	15.89%
Past Criminal History Gender Model	70.97%	62.95%	24.07%
Proxy-Eliminated Race Model	70.84%	55.62%	17.87%
Proxy-Eliminated Gender Model	86.49%	76.70%	06.00%

Table 7: Dataset 3: Random Forest results

comparison to the other tables, the bias mitigation methods only outscored the base LR-model 7 out of 30 times in total which is a rather low amount of times.

Looking at the scores of the bias mitigation methods on the third dataset using RF in table 7, both the Race Eliminated Model and the Proxy-Eliminated Gender Model outperformed the the base model on all fronts. The accuracy improved by an impressive 13.29%, the TPR with 18.23% and the FPR with 9.86%. In table 3 and 5 the Proxy eliminated models both outscored the base RF-model, but in table 7 the Race Model performs worse than the base model on all fronts. The Singular Race Model and the Gender Eliminated Model manage to outperform the base model on 2 fronts. The bias mitigation methods outscore the base RF-model 14 out of 30 times in total.

4.3 *Overview Results per Dataset*

Table 8 puts together the information shown in the previous subsection and shows which method worked best on which dataset. Dataset 2 had the lowest amount of features with only 11 after data processing and the bias reduction methods seem to work best on that dataset outscoring the base model 30 times. Dataset 3 has the most features with 50 after data processing and the mitigation methods seem to work worst on this dataset with metrics only improving 21 times. The Gender Eliminated and the Proxy-Eliminated Gender Models score higher than the base model the most often. Another interesting point is that this table shows that the bias mitigation methods show the most promise on the gender based models by scoring high 48 times versus the race based models only scoring high 28 times. This might have something to do with the greater disparity in gender (see Figure 4) versus race (see Figure 3).

	Singular Race	**Singular Gender**	**Balanced Race**	**Balanced Gender**	**Race Eliminated**	**Gender Eliminated**	**Past Criminal History Race**	**Past Criminal History Gender Model**	**Proxy-Eliminated Race**	**Proxy-Eliminated Gender**	**Total**
Dataset 1	1	4	2	3	0	3	2	2	5	4	26
Dataset 2	1	3	2	3	2	5	2	4	4	4	30
Dataset 3	2	2	1	2	4	4	1	2	0	3	21
Total	4	9	5	8	6	12	5	8	9	11	**77**

Table 8: Number of times a bias mitigated model outperformed the base model per dataset (including accuracy)

	Singular Race	**Singular Gender**	**Balanced Race**	**Balanced Gender**	**Race Eliminated**	**Gender Eliminated**	**Past Criminal History Race**	**Past Criminal History Gender Model**	**Proxy-Eliminated Race**	**Proxy-Eliminated Gender**	**Total**
Dataset 1	1	2	2	2	0	2	2	2	3	3	19
Dataset 2	1	1	2	2	2	3	2	2	3	3	21
Dataset 3	2	2	1	1	3	2	1	2	0	2	16
Total	4	5	5	5	5	7	5	6	6	8	**56**

Table 9: Number of times a bias mitigated model outperformed the base model per dataset (excluding accuracy)

Table 9 shows the same information as Table 8, but does not take an improvement in accuracy into account. It shows the results when purely focused on an increase in TPR and/or a decrease in FPR. At the top end, the same methods keep scoring the highest as in table 8, with the Gender Eliminated and the Gender Eliminated Model scoring the highest in both tables. The Past Criminal History Gender and the Proxy-Eliminated Gender Model keep scoring high as well in both Table 8 and 9. The Singular Gender and Balanced Gender drop off a bit in table 9 compared to table 8, although both still outperformed the base model 5 times. The difference in results between these two tables is not that big. Looking at both tables the most successful models would be the Gender Eliminated, the Proxy-eliminated gender and the Gender Eliminated Model. All three are gender based models. The race models outperformed the base model 24 times in total, the gender models outperformed 31 times in total. This tendency can also be seen in Table 9. When not accounting for the type of ML-model used and the features in the dataset, the best performing bias mitigated models for gender and race are shown below.

The number behind the model stands for the number of times the model outscored the base model. The best best performing mitigation models for gender are:

1. The Gender Eliminated Model (12)
2. The Proxy-Eliminated Gender Model (11)
3. Singular Gender Model (9)

And the best performing mitigation models for race are:

1. The Proxy-Eliminated Race Model (9)
2. The Race Eliminated Model (6)
3. Balanced Race Model/Past Criminal History Race Model (5)

The overall best performing bias mitigation methods form the same list as the list for the best performing gender models, but with one inclusion of a race based model:

1. The Gender Eliminated Model (12)
2. The Proxy-Eliminated Gender Model (11)
3. Singular Gender Model (9)
4. The Proxy-Eliminated Race Model (9)

The gender and race scores can also be combined to create a general overview of the mitigation methods per dataset, which is shown in Table 10. Here the Proxy-Eliminated method is the most successful, with the Race/Gender eliminated method as the second most successful. The other three methods are together tied for third place. Note that table 10 is less useful in practice given the vast differences in the performance of bias mitigation models per ML-model, which can be seen in the next subsection.

	Dataset 1	Dataset 2	Dataset 3	Total
Singular Race/Gender	5	4	4	13
Balanced Race/Gender	5	5	3	13
Race/Gender Eliminated	3	7	8	18
Past Criminal History Race/Gender	4	6	3	13
Proxy-Eliminated Race/Gender	9	8	3	20
Total	26	30	21	77

Table 10: Number of times a method outperformed the base model per dataset

	Accuracy	TPR	FPR	Total
Singular Race Model	0	1	1	2
Singular Gender Model	2	2	0	4
Balanced Race Model	0	0	2	2
Balanced Gender Model	2	3	1	6
Race Eliminated Model	0	1	1	2
Gender Eliminated Model	3	1	2	6
Past Criminal History Race Model	0	1	2	3
Past Criminal History Gender Model	1	3	0	4
Proxy-Eliminated Race Model	1	1	1	3
Proxy-Eliminated Gender Model	0	1	1	2
Total	9	14	11	**34**

Table 11: Number of times a model outperformed the base LR-Model

4.4 *Overview Results Logistic Regression and Random Forest*

Table 11 shows the results per mitigation method for LR. The two best performing models when using LR are the Balanced Gender and Gender Eliminated Model, that both outperformed the base model on six out of six scores. The Singular Gender and the Past Criminal History Gender Model both outperformed the base model on 4 scores taking the 2nd place. The 4 least performing models only managed to improve the base model score 2 times. The LR-model most often had an improvement in TPR and least often in accuracy.

	Accuracy	TPR	FPR	Total
Singular Race Model	0	1	1	2
Singular Gender Model	2	3	0	5
Balanced Race Model	0	0	2	2
Balanced Gender Model	0	2	0	2
Race Eliminated Model	1	1	2	4
Gender Eliminated Model	2	2	2	6
Past Criminal History Race Model	0	1	1	2
Past Criminal History Gender Model	1	3	1	5
Proxy-Eliminated Race Model	2	2	2	6
Proxy-Eliminated Gender Model	3	3	3	9
Total	11	18	14	**43**

Table 12: Number of times a model outperformed the base RF-Model

	LR-model	RF-model	Total
Singular Race Model	2	2	4
Singular Gender Model	4	5	9
Balanced Race Model	2	2	4
Balanced Gender Model	6	2	8
Race Eliminated Model	2	4	6
Gender Eliminated Model	6	6	12
Past Criminal History Race Model	3	2	5
Past Criminal History Gender Model	4	5	9
Proxy-Eliminated Race Model	3	6	9
Proxy-Eliminated Gender Model	2	9	11
Total	34	43	**77**

Table 13: Number of times bias mitigated models outperformed the base models in total

Table 12 shows the results per mitigation method for RF. The Proxy-Eliminated Gender Model scores the highest by far by outperforming the base model on 9 out of 9 scores. The Gender Eliminated and the Proxy-Eliminated Race Model both take second place by outperforming 6 scores. The Singular Gender and the Past Criminal History Gender Model both outperformed on 5 scores as well. Just like with the LR models, the 4 worst performing bias mitigation methods only outscored 2 times. For both RF and LR, the Singular Race and the Balanced Race Model scored the worst. The RF-model also most often had an improvement in TPR and least often in accuracy, just as the LR-model.

Table 13 shows the total amount that the bias mitigated models outperformed the base model in total for the LR- and the RF-model. Overall the mitigation methods work considerably more often on RF: outperforming the base model on 43 scores, versus the 34 times of the LR-model. Interestingly enough, one the two best performing models for LR (Balanced Gender Model) is among the worst performing models for RF: outscoring 6 times for LR and only twice for RF. Similarly, the absolute best scoring model for RF (Proxy-Eliminated Gender Model) outperformed the base model every time on all fronts (9 times in total), but only managed to do so twice for LR, putting it among the worst performing models there. The Gender Eliminated Model is the only one among the highest scoring ones for both ML-models.

The best performing models for LR are:

1. Balanced Gender Model (6)
2. Gender Eliminated Model (6)
3. Singular Gender Model (4)
4. Past Criminal History Gender Model (4)

The best performing models for RF are:

1. Proxy-Eliminated Gender Model (9)
2. Gender Eliminated Model (6)

3. Proxy-Eliminated Race Model (6)

5 DISCUSSION

The baseline Random Forest outperformed the baseline model Logistic Regression on two of the three datasets and also more often on TPR and FPR. RF does seem like the ML-model that is better in general for predicting recidivism while also having a high TPR and a low FPR. LR however is still useful as a predictive tool, it seems to do better on datasets with a low amount of features (dataset 2) where it outperformed the RF-model. Both models have their own mitigation methods that are more effective. Logistic Regression had the best results, with the Balanced Gender Model and the Gender Eliminated Model, both outperforming the base model on 6 scores. The results of balancing the dataset and testing it on a LR-model also showed it could make an algorithm more fair in the original paper (Biswas et al., 2020). Random Forest had the best results with the Proxy-Eliminated Gender Model outperforming the base model on 9 out of 9 scores. The Gender Eliminated Model and the Proxy-Eliminated Race Model also did well outperforming the base model on 6 out of 9 scores. The effectiveness of the proxy-Eliminated model has also been confirmed (Skeem & Lowenkamp, 2020) in favor of methods that completely eliminate race.

Interestingly enough, the best performing mitigation methods for RF differ from the ones scoring the highest for LR, with the exception being the Gender Eliminated Model which scores high for both models. The best mitigation model for RF (Proxy-Eliminated Gender Model) actually scores among the worst for LR and vice versa one of the best performing mitigation model for LR (Balanced Gender Model) scores among the worst for the RF-model. The bias mitigation methods also seem to have more effect on the Random Forest Model by outscoring the base model 43 times in total after their use, while the mitigation methods only allowed the LR-model to outperform the base model 34 times.

Another distinction that can be made is between race and gender. The mitigation models based on race outperformed the base model only 28 times, whilst the mitigation models based on gender outscored the base model 48 times. It seems a lot harder to lower the bias when accounting for race, than to lower the bias when accounting for gender. This can have something to do with the greater disparity between genders than the disparity between races. The final distinction that can be made is in the dataset used. The characteristics of a dataset also influenced how effective the bias mitigation methods were. They had the most impact on dataset 2 which has the least features (11) and the least impact on dataset 3 which had the most features (50).

Given the big differences in the effectiveness of the bias mitigation methods when regarding the amount of features in the dataset, the type of ML-model used and the disparity in gender/race, it seems wise to choose different mitigation methods accordingly. If society wants to have an accurate risk assessment of whether a convict will commit recidivism and a low ethnic/gender bias, this thesis shows that Random Forest is generally the better choice of model in terms of accuracy as was the case in the literature from section 2.2 (Duwe & Kim, 2017; Ghasemi et al., 2021; Jain et al., 2019; Mehta et al., 2020) and it even works better with mitigation methods in general in comparison to Logistic Regression. The performance however is influenced heavily by the characteristics of the dataset. Random Forest in this thesis is best combined with the Proxy-eliminated method to mitigate a gender bias. This is also the best bias reduction method for RF to use to reduce a racial bias.

Logistic Regression still seems to have some merit, in this thesis at least when used on a smaller dataset and it is best combined with the Past Criminal History Race Model and the Proxy-Eliminated Race Model to mitigate the bias for race. To mitigate a gender based bias with LR, the best performing models were the Balanced Gender and the Gender Eliminated Model which both outscored the base model the same number of times.

There are however quite some limitations in this thesis. In one paper on predicting recidivism discussed in this thesis (Miron et al., 2020) one of the best performing ML-models was Logistic Regression, but also a Multilayer Perceptron (MLP). Due to time constraints and this being the only paper with this model performing so well, MLP was not included in this thesis. It would be interesting to see how this model would fare on multiple datasets compared to RF and LR.

Furthermore, for this thesis 3 datasets from the USA were used, which mostly consisted of black and male offenders. A dataset with a larger variety of races and/or genders or a different distribution of races and/or genders could have very different results. Differences in the quality and the amount of features can also impact which method is more suited. In this thesis LR outperformed RF on a dataset with fewer features, but it could also be due to the quality and not the quantity of the features. It would be interesting to see a study performed on a plethora of datasets with even bigger disparities using these ML-models.

This thesis did not research the combination of multiple bias mitigation methods. The Past Criminal History method for example can be combined with any of the other 4 methods and had decent results on its own scoring third highest overall when used to mitigate bias. Future research on this could be valuable in finding a combined method that is even more successful in maintaining accuracy and lowering bias, although the success is again dependent on the ML-model chosen, the characteristics of the dataset and the type of bias that is to be mitigated.

6 CONCLUSION

The main research question was 'To what extent can the bias be lowered while maintaining the accuracy when predicting recidivism with a Random Forest and a Logistic Regression model?'. By answering the three subquestions mentioned in the introduction, an answer will be given to this question.

How accurate can Logistic Regression and Random Forest predict recidivism in the three datasets? The Random Forest Model can predict recidivism with a higher accuracy than the Logistic Regression Model for dataset 1 and 3, but struggles more with dataset 2. After implementing the bias reduction methods, the RF-model outscored the base model more often than the LR-model.

How do the five bias reduction methods individually influence the prediction accuracy and bias of the Logistic Regression and Random Forest models? When using table 10, the Proxy-Eliminated Race/Gender methods worked best on dataset 1 and also on dataset 2. On dataset 3, the Race/Gender Eliminated method scored the best. The results from table 9 however might give a better indication, given the vast difference in results of the race-based models and the gender-based models.

What bias reduction methods had the best results on which model? And on which dataset? The best performing models for LR when mitigating the bias for race were the Past Criminal History Race Model and the Proxy-Eliminated Race Model. To mitigate the bias of gender with LR, the best performing models were the Balanced Gender Model and the Gender Eliminated Model. Random Forest in this thesis is best combined with the Proxy-eliminated method to mitigate a racial bias. This is also the best bias reduction method for RF to use to mitigate a bias in gender. The Proxy-Eliminated Race Model outscored the base model the most often on dataset 1, the Gender Eliminated Model on dataset 2 and the Race Eliminated Model and the Gender Eliminated Model both outscored the most often on dataset 3.

Different mitigation methods impact different models. As the results have shown, there are models that after using a bias mitigation method, perform better than their baseline counterparts. Usually though there is a tradeoff between accuracy and either FPR or TPR. What metric is more important could also impact the decision what bias mitigation method is the most desirable for a lawmaker, a Department of Justice or a judge while maintaining a high accuracy. Future research could focus on adding the MLP model or test on datasets from other regions or with a different distribution. Using a combination of mitigation methods is also a promising, although time consuming endeavour. With a plethora of bias mitigation methods available and probably even more in the future, a research testing other new bias mitigation methods on these ML-models is also most welcome.

REFERENCES

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Biddle, J. B. (2020). On predicting recidivism: Epistemic risk, tradeoffs and values in machine learning. *Canadian Journal of Philosophy*, 1-21. Retrieved from http://doi.org/10.1017/can.2020.27 doi: 10.1017/can.2020.27

Biswas, A., Kolczynska, M., Rantanen, S., & Rozenshtein, P. (2020). The role of in-group bias and balanced data: A comparison of human and machine recidivism risk predictions. *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies. Association for Computing Machinery, New York, NY, USA*, 97–104. Retrieved from https://doi.org/10.1145/3378393.3402507 doi: 10.1145/3378393.3402507

Brownlee, J. (2019). *Tune hyperparameters for classification machine learning algorithms.* Retrieved from https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/ (Online; accessed 18-05-2015)

Danks, D., & London, A. (2017, 08). Algorithmic bias in autonomous systems. In (p. 4691-4697). doi: 10.24963/ijcai.2017/654

Dressel, J., & Farid, H. (2018). The role of in-group bias and balanced data: A comparison of human and machine recidivism risk predictions. *Science Advances*, *4*(1). Retrieved from https://doi.org/10.1145/3378393.3402507 doi: 10.1145/3378393.3402507

Duwe, G., & Kim, K. (2017). Out with the old and in with the new? an empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Review*, *28*(6), 570–600. Retrieved from https://doi.org/10.1177/0887403415604899 doi: 10.1177/0887403415604899

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018, 02). Runaway feedback loops in predictive policing. In *Proceedings of the 1st conference on fairness, accountability and transparency* (pp. 160–171). Retrieved from https://arxiv.org/abs/1706.09847 doi: 10.48550/ARXIV.1706.09847

Ghasemi, M., Anvari, D., Atapour, M., Wormith, J. S., Stockdale, K. C., & Spiteri, R. J. (2021). The application of machine learning to a general risk–need assessment instrument in the prediction of criminal recidivism. *Criminal Justice and Behavior*, *48*(4), 518–538. Retrieved from https://doi.org/10.1177/0093854820969753 doi: 10.1177/0093854820969753

Haghighi, M. K., & Castillo, C. (2021). Enhancing a recidivism prediction tool with machine learning: effectiveness and algorithmic fairness. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. Association for Computing Machinery, New York, NY, USA*, 210–214. Retrieved from https://doi.org/10.1145/3462757.3466150 doi: 10.1145/3462757.3466150

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. doi: 10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(3), 90–95.

Jain, B., Huber, M., Elmasri, R. A., & Fegaras, L. (2020). Reducing race-based bias and increasing recidivism prediction accuracy by using past criminal history details. *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '20). Association for Computing Machinery, New York, NY, USA*, 1–8. Retrieved from https://doi.org/10.1145/3389189.3397990 doi: 10.1145/3389189.3397990

Jain, B., Huber, M., Fegaras, L., & Elmasri, R. A. (2019). Singular race models: addressing bias and accuracy in predicting prisoner recidivism. *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '19). Association for Computing Machinery, New York, NY, USA*, 599–607. Retrieved from https://doi.org/10.1145/3316782.3322787 doi: 10.1145/3316782.3322787

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv*, 1-23. Retrieved from https://arxiv.org/abs/1609.05807 doi: 10.48550/ARXIV.1609.05807

McKinney, W., et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56).

Mehta, H., Shah, S., Patel, N., & Kanani, P. (2020). Classification of criminal recidivism using machine learning techniques. *International Journal of Advanced Science and*

Technology, *29*(04), 5110 - 5122. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/24940

Miron, M., Tolan, S., Gutiérrez, E. G., & Castillo, C. (2020). Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law*, *29*(2), 5110 - 5122. Retrieved from https://doi.org/10.1007/s10506-020-09268-y doi: 10.1007/s10506-020-09268-y

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . others (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. (2017). On fairness and calibration.

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3). Retrieved from https://doi.org/10.1002%2Fwidm.1301 doi: 10.1002/widm.1301

Skeem, J., & Lowenkamp, C. (2020). Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & the Law*, *38*(3), 259–278. Retrieved from https://doi.org/10.1002/bsl.2465 doi: 10.1002/bsl.2465

Srivastava, T. (2015). *Tuning the parameters of your random forest model.* Retrieved from https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/ (Online; accessed 18-05-2015)

Wadsworth, C., Vera, F., & Piech, C. (2018). Achieving fairness through adversarial learning: An application to recidivism prediction. *arXiv preprint*. Retrieved from https://doi.org/10.48550/arXiv.1807.00199 doi: 10.48550/arXiv.1807.00199

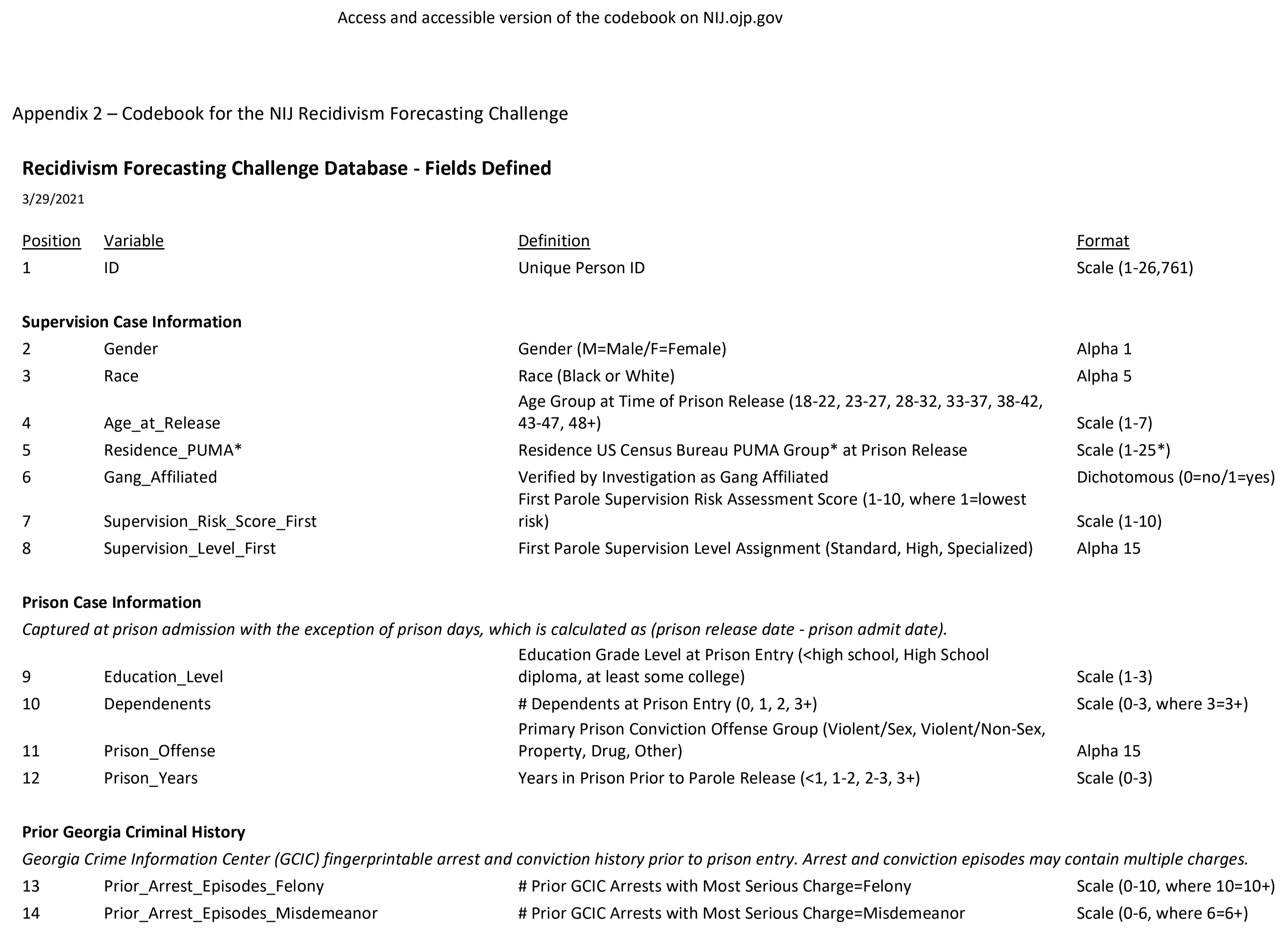
APPENDIX A: FEATURES OF THE DATASETS

	Feature	Description	Type
1	id	unique number	float64
2	name	full name	object
3	first	first name	object
4	last	last name	object
5	compas_screening_date	date COMPAS was used	object
6	sex	gender	object
7	dob	date of birth	object
8	age	age	int64
9	age_cat	age category	object
10	race	race	object
11	juv_fel_count	number of juvenile felonies	int64
12	decile_score	decile score COMPAS	int64
13	juv_misd_count	juvenile misdemeanor count	int64
14	juv_other_count	juvenile other count	int64
15	priors_count	number of priors	int64
16	days_b_screening_arrest	days before screening	float64
17	c_jail_in	time when the defendant was jailed.	object
18	c_jail_out	time when the defendant was released from the jail.	object
19	c_case_number	criminal case number	object
20	c_offense_date	criminal offense date	object
21	c_arrest_date	crime arrest date	object
22	c_days_from_compas	days from arrest to COMPAS score	float64
23	c_charge_degree	charge degree	object
24	c_charge_desc	charge description	object
25	is_recid	Is defendant a recidivist	int64
26	r_case_number	case number of reoffence	object
27	r_charge_degree	reoffence charge degree	object
28	r_days_from_arrest	reoffence days from arrest	float64
29	r_offense_date	reoffence date	object
30	r_charge_desc	reoffence description	object
31	r_jail_in	time when put in jail for reoffence	object
32	r_jail_out	time let out of jail for reoffence	object
33	violent_recid	whether reoffence is deemed violent during arrest	float64
34	is_violent_recid	whether reoffence is deemed violent in court	int64
35	vr_case_number	violent reoffence case number	object
36	vr_charge_degree	violent reoffence charge degree	object
37	vr_offense_date	violent reoffence offense date	object
38	vr_charge_desc	violent reoffence charge description	object
39	type_of_assessment	assessment type	object
40	decile_score.1	decile score 2	int64
41	score_text	text of decile score	object
42	screening_date	screening date	object
43	v_type_of_assessment	violent offense type of assessment	object
44	v_decile_score	violent offense decile score	int64
46	v_score_text	violent offense score text	object
47	v_screening_date	violent offense screening date	object
48	in_custody	time when put in custody	object
49	out_custody	time when out of custody	object
50	priors_count.1	number of priors 2	int64
51	start	start	int64
52	end	end	int64
53	event	event	int64

Table 14: Features Dataset 1

	Column Name	Description	Type
1	Fiscal Year Released	Fiscal year (year ending June 30) for which the offender was released from prison.	Number
2	Recidivism Reporting Year	Fiscal year (year ending June 30) that marks the end of the 3-year tracking period. For example, offenders exited prison in FY 2012 are found in recidivism reporting year FY 2015.	Number
3	Main Supervising District	The Judicial District supervising the offender for the longest time during the tracking period.	Plain Text
4	Release Type	Reasoning for Offender's release from prison.	Plain Text
5	Race - Ethnicity	Offender's Race and Ethnicity	Plain Text
6	Age At Release	Offender's age group at release from prison.	Plain Text
7	Sex	Gender of our offender	Plain Text
8	Offense Classification	Maximum penalties: A Felony=Life; B Felony=25 or 50 years; C Felony=10 years; D Felony=5 years; Aggravated Misdemeanor=2 years; Serious Misdemeanor=1 year; Simple Misdemeanor=30 days	Plain Text
9	Offense Type	General category for the most serious offense for which the offender was placed in prison.	Plain Text
10	Offense Subtype	Further classification of the most serious offense for which the offender was placed in prison.	Plain Text
11	Return to Prison	No = Did not return to prison within the three year tracking period; Yes = Admitted to prison for any reason within the three year tracking period	Plain Text
12	Days to Return	Number of days it took before the offender returned to prison.	Number
13	Recidivism Type	Indicates the reason for return to prison.	Plain Text
14	New Offense Classification	New conviction maximum penalties: A Felony=Life; B Felony=25 or 50 years; C Felony=10 years; D Felony=5 years; Aggravated Misdemeanor=2 years; Serious Misdemeanor=1 year; Simple Misdemeanor=30 days	Plain Text
15	New Offense Type	General category for the new conviction while the offender is out of prison.	Plain Text
16	New Offense Sub Type	Further classification of the new conviction.	Plain Text
17	Target Population	The Department of Corrections has undertaken specific strategies to reduce recidivism rates for prisoners who are on parole.	Plain Text

Table 15: Features Dataset 2



Access and accessible version of the codebook on NIJ.ojp.gov

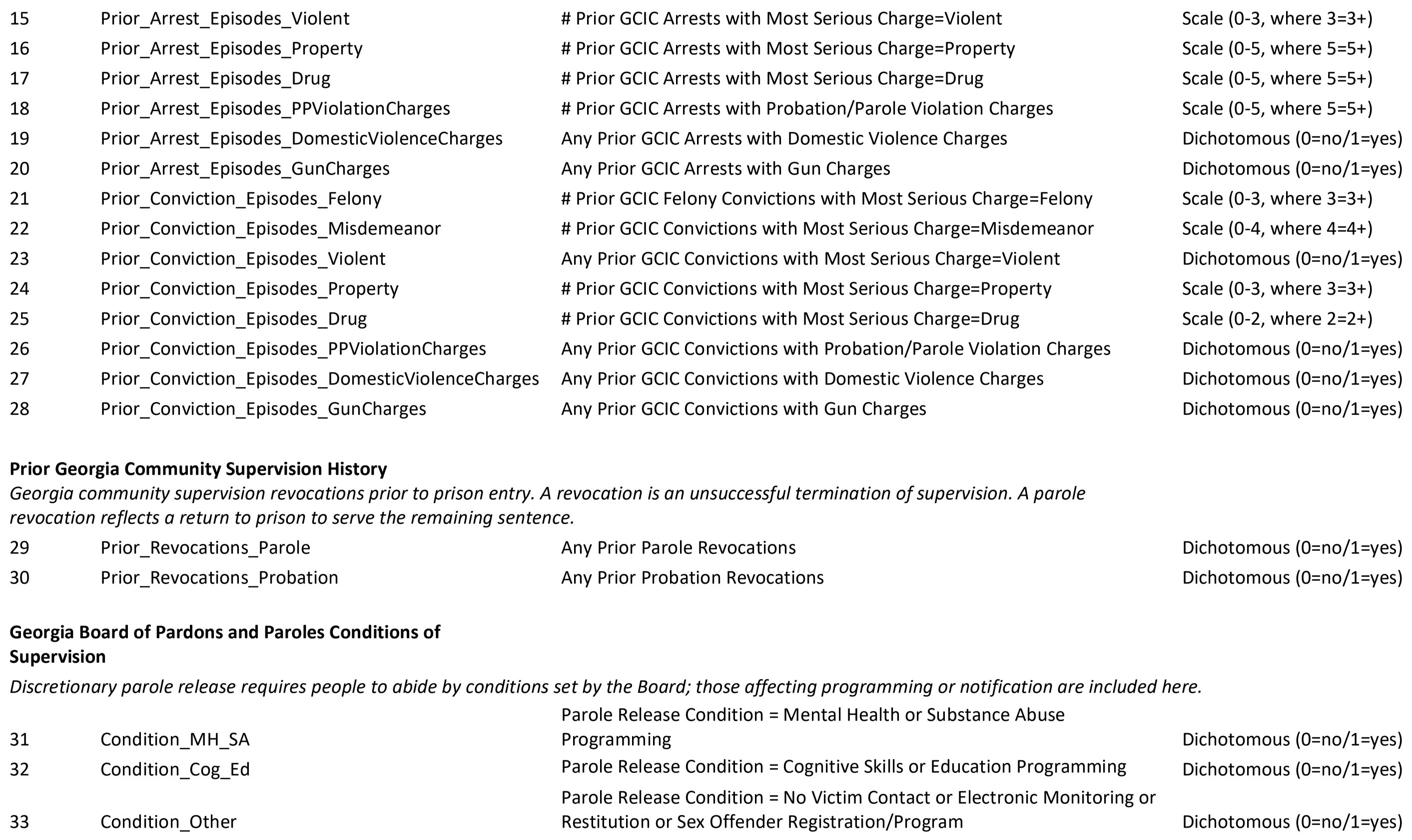
Appendix 2 – Codebook for the NIJ Recidivism Forecasting Challenge

Recidivism Forecasting Challenge Database - Fields Defined

3/29/2021

Position	Variable	Definition	Format
1	ID	Unique Person ID	Scale (1-26,761)
Supervision Case Information			
2	Gender	Gender (M=Male/F=Female)	Alpha 1
3	Race	Race (Black or White)	Alpha 5
4	Age_at_Release	Age Group at Time of Prison Release (18-22, 23-27, 28-32, 33-37, 38-42, 43-47, 48+)	Scale (1-7)
5	Residence_PUMA*	Residence US Census Bureau PUMA Group* at Prison Release	Scale (1-25*)
6	Gang_Affiliated	Verified by Investigation as Gang Affiliated	Dichotomous (0=no/1=yes)
7	Supervision_Risk_Score_First	First Parole Supervision Risk Assessment Score (1-10, where 1=lowest risk)	Scale (1-10)
8	Supervision_Level_First	First Parole Supervision Level Assignment (Standard, High, Specialized)	Alpha 15
Prison Case Information			
Captured at prison admission with the exception of prison days, which is calculated as (prison release date - prison admit date).			
9	Education_Level	Education Grade Level at Prison Entry (<high school, High School diploma, at least some college)	Scale (1-3)
10	Dependenents	# Dependents at Prison Entry (0, 1, 2, 3+)	Scale (0-3, where 3=3+)
11	Prison_Offense	Primary Prison Conviction Offense Group (Violent/Sex, Violent/Non-Sex, Property, Drug, Other)	Alpha 15
12	Prison_Years	Years in Prison Prior to Parole Release (<1, 1-2, 2-3, 3+)	Scale (0-3)
Prior Georgia Criminal History			
Georgia Crime Information Center (GCIC) fingerprintable arrest and conviction history prior to prison entry. Arrest and conviction episodes may contain multiple charges.			
13	Prior_Arrest_Episodes_Felony	# Prior GCIC Arrests with Most Serious Charge=Felony	Scale (0-10, where 10=10+)
14	Prior_Arrest_Episodes_Misdemeanor	# Prior GCIC Arrests with Most Serious Charge=Misdemeanor	Scale (0-6, where 6=6+)

Figure 6: Features Dataset 3-1, retrieved on 21-06-2022 from https://nij.ojp.gov/funding/recidivism-forecasting-challenge-appendix-2-codebook.pdf



15	Prior_Arrest_Episodes_Violent	# Prior GCIC Arrests with Most Serious Charge=Violent	Scale (0-3, where 3=3+)
16	Prior_Arrest_Episodes_Property	# Prior GCIC Arrests with Most Serious Charge=Property	Scale (0-5, where 5=5+)
17	Prior_Arrest_Episodes_Drug	# Prior GCIC Arrests with Most Serious Charge=Drug	Scale (0-5, where 5=5+)
18	Prior_Arrest_Episodes_PPViolationCharges	# Prior GCIC Arrests with Probation/Parole Violation Charges	Scale (0-5, where 5=5+)
19	Prior_Arrest_Episodes_DomesticViolenceCharges	Any Prior GCIC Arrests with Domestic Violence Charges	Dichotomous (0=no/1=yes)
20	Prior_Arrest_Episodes_GunCharges	Any Prior GCIC Arrests with Gun Charges	Dichotomous (0=no/1=yes)
21	Prior_Conviction_Episodes_Felony	# Prior GCIC Felony Convictions with Most Serious Charge=Felony	Scale (0-3, where 3=3+)
22	Prior_Conviction_Episodes_Misdemeanor	# Prior GCIC Convictions with Most Serious Charge=Misdemeanor	Scale (0-4, where 4=4+)
23	Prior_Conviction_Episodes_Violent	Any Prior GCIC Convictions with Most Serious Charge=Violent	Dichotomous (0=no/1=yes)
24	Prior_Conviction_Episodes_Property	# Prior GCIC Convictions with Most Serious Charge=Property	Scale (0-3, where 3=3+)
25	Prior_Conviction_Episodes_Drug	# Prior GCIC Convictions with Most Serious Charge=Drug	Scale (0-2, where 2=2+)
26	Prior_Conviction_Episodes_PPViolationCharges	Any Prior GCIC Convictions with Probation/Parole Violation Charges	Dichotomous (0=no/1=yes)
27	Prior_Conviction_Episodes_DomesticViolenceCharges	Any Prior GCIC Convictions with Domestic Violence Charges	Dichotomous (0=no/1=yes)
28	Prior_Conviction_Episodes_GunCharges	Any Prior GCIC Convictions with Gun Charges	Dichotomous (0=no/1=yes)

Prior Georgia Community Supervision History

Georgia community supervision revocations prior to prison entry. A revocation is an unsuccessful termination of supervision. A parole revocation reflects a return to prison to serve the remaining sentence.

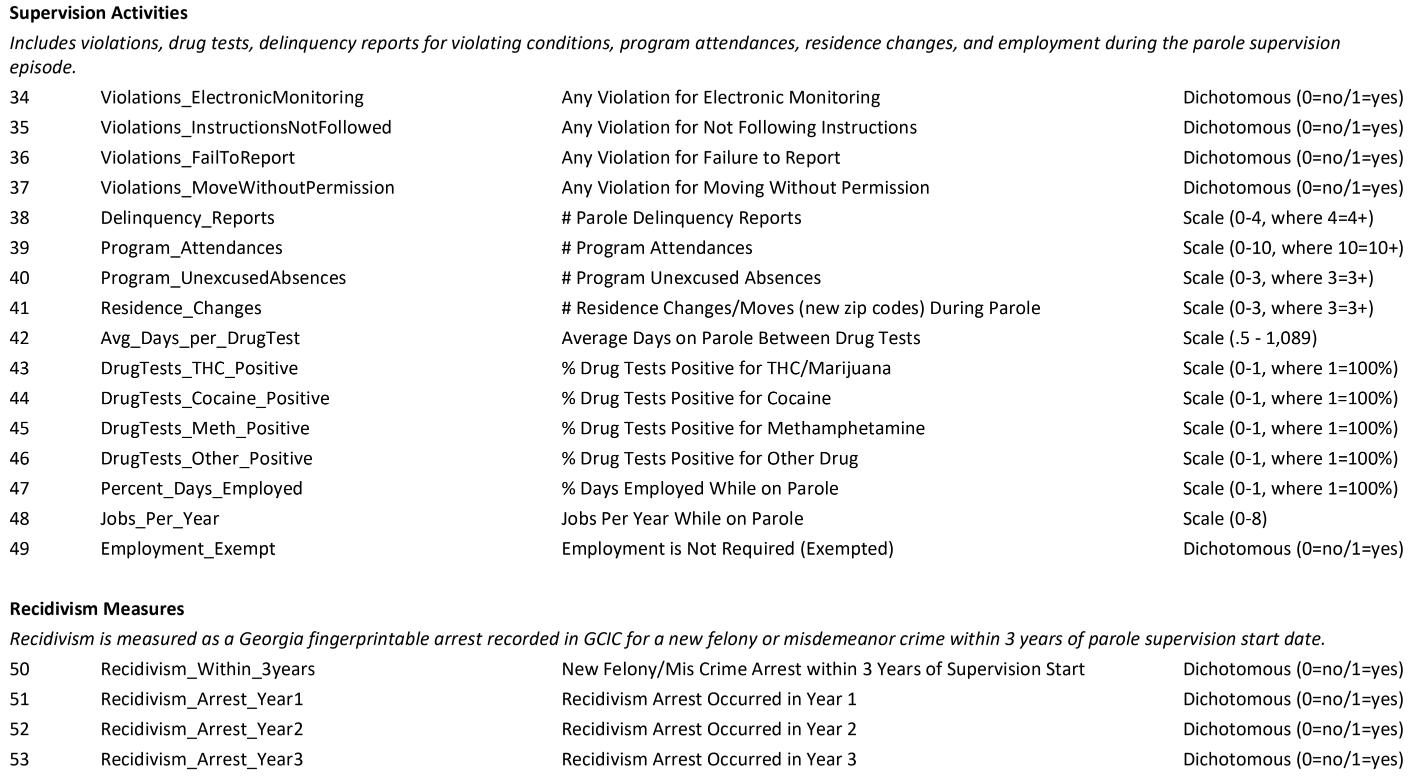
29	Prior_Revocations_Parole	Any Prior Parole Revocations	Dichotomous (0=no/1=yes)
30	Prior_Revocations_Probation	Any Prior Probation Revocations	Dichotomous (0=no/1=yes)

Georgia Board of Pardons and Paroles Conditions of Supervision

Discretionary parole release requires people to abide by conditions set by the Board; those affecting programming or notification are included here.

31	Condition_MH_SA	Parole Release Condition = Mental Health or Substance Abuse Programming	Dichotomous (0=no/1=yes)
32	Condition_Cog_Ed	Parole Release Condition = Cognitive Skills or Education Programming	Dichotomous (0=no/1=yes)
33	Condition_Other	Parole Release Condition = No Victim Contact or Electronic Monitoring or Restitution or Sex Offender Registration/Program	Dichotomous (0=no/1=yes)

Figure 7: Features Dataset 3-2, retrieved on 21-06-2022 from https://nij.ojp.gov/funding/recidivism-forecasting-challenge-appendix-2-codebook.pdf



Supervision Activities

Includes violations, drug tests, delinquency reports for violating conditions, program attendances, residence changes, and employment during the parole supervision episode.

34	Violations_ElectronicMonitoring	Any Violation for Electronic Monitoring	Dichotomous (0=no/1=yes)
35	Violations_InstructionsNotFollowed	Any Violation for Not Following Instructions	Dichotomous (0=no/1=yes)
36	Violations_FailToReport	Any Violation for Failure to Report	Dichotomous (0=no/1=yes)
37	Violations_MoveWithoutPermission	Any Violation for Moving Without Permission	Dichotomous (0=no/1=yes)
38	Delinquency_Reports	# Parole Delinquency Reports	Scale (0-4, where 4=4+)
39	Program_Attendances	# Program Attendances	Scale (0-10, where 10=10+)
40	Program_UnexcusedAbsences	# Program Unexcused Absences	Scale (0-3, where 3=3+)
41	Residence_Changes	# Residence Changes/Moves (new zip codes) During Parole	Scale (0-3, where 3=3+)
42	Avg_Days_per_DrugTest	Average Days on Parole Between Drug Tests	Scale (.5 - 1,089)
43	DrugTests_THC_Positive	% Drug Tests Positive for THC/Marijuana	Scale (0-1, where 1=100%)
44	DrugTests_Cocaine_Positive	% Drug Tests Positive for Cocaine	Scale (0-1, where 1=100%)
45	DrugTests_Meth_Positive	% Drug Tests Positive for Methamphetamine	Scale (0-1, where 1=100%)
46	DrugTests_Other_Positive	% Drug Tests Positive for Other Drug	Scale (0-1, where 1=100%)
47	Percent_Days_Employed	% Days Employed While on Parole	Scale (0-1, where 1=100%)
48	Jobs_Per_Year	Jobs Per Year While on Parole	Scale (0-8)
49	Employment_Exempt	Employment is Not Required (Exempted)	Dichotomous (0=no/1=yes)

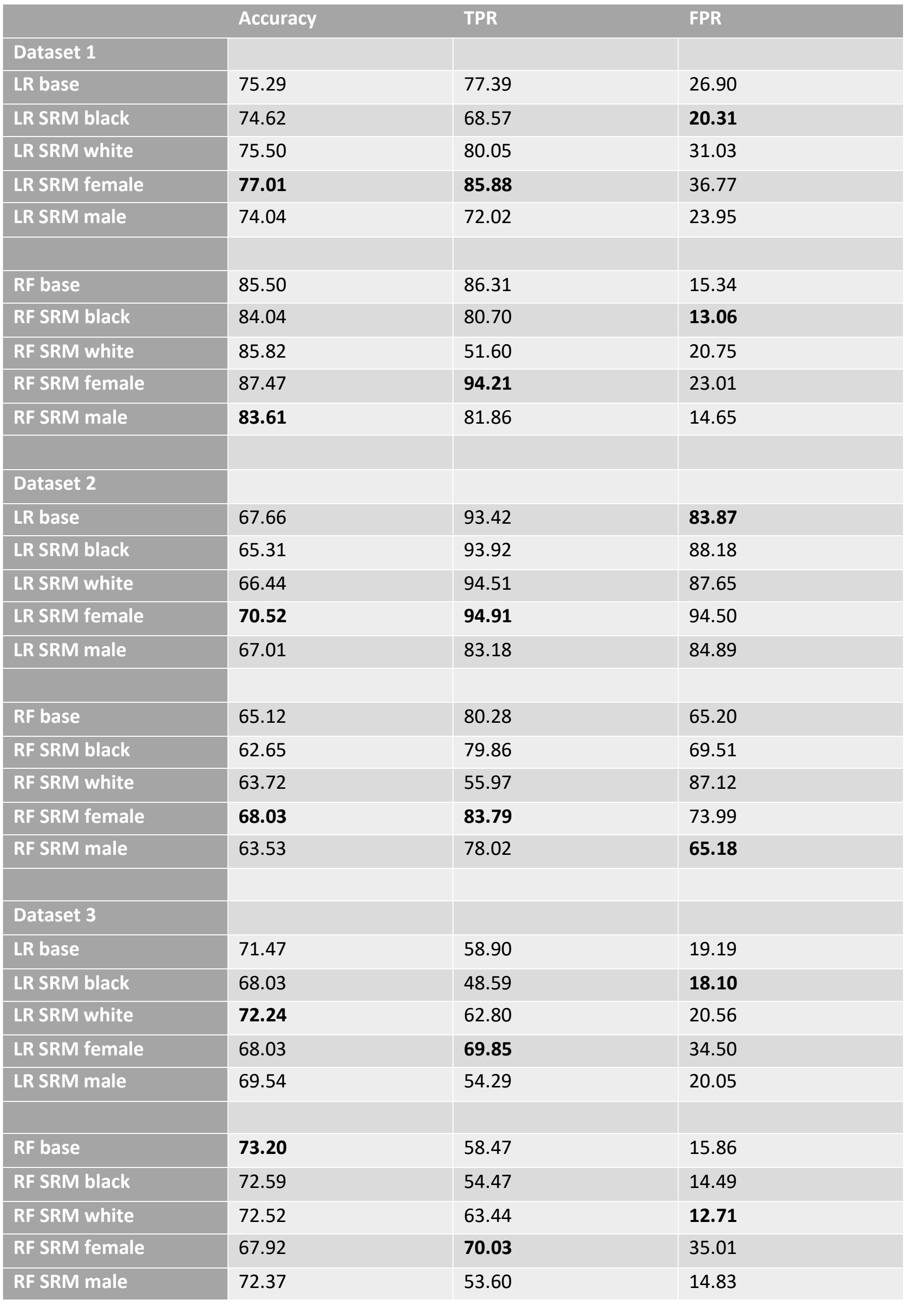
Recidivism Measures

Recidivism is measured as a Georgia fingerprintable arrest recorded in GCIC for a new felony or misdemeanor crime within 3 years of parole supervision start date.

50	Recidivism_Within_3years	New Felony/Mis Crime Arrest within 3 Years of Supervision Start	Dichotomous (0=no/1=yes)
51	Recidivism_Arrest_Year1	Recidivism Arrest Occurred in Year 1	Dichotomous (0=no/1=yes)
52	Recidivism_Arrest_Year2	Recidivism Arrest Occurred in Year 2	Dichotomous (0=no/1=yes)
53	Recidivism_Arrest_Year3	Recidivism Arrest Occurred in Year 3	Dichotomous (0=no/1=yes)

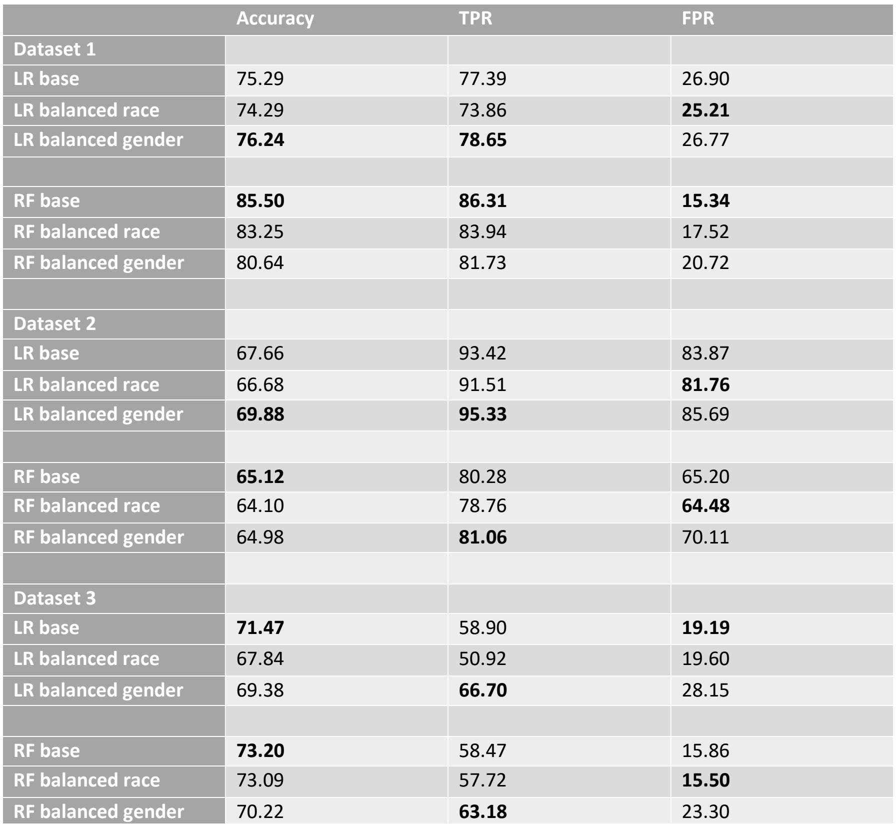
Figure 8: Features Dataset 3-3, retrieved on 21-06-2022 from https://nij.ojp.gov/funding/recidivism-forecasting-challenge-appendix-2-codebook.pdf

APPENDIX B: ACCURACY, TPR AND FPR PER BIAS MITIGATED MODEL



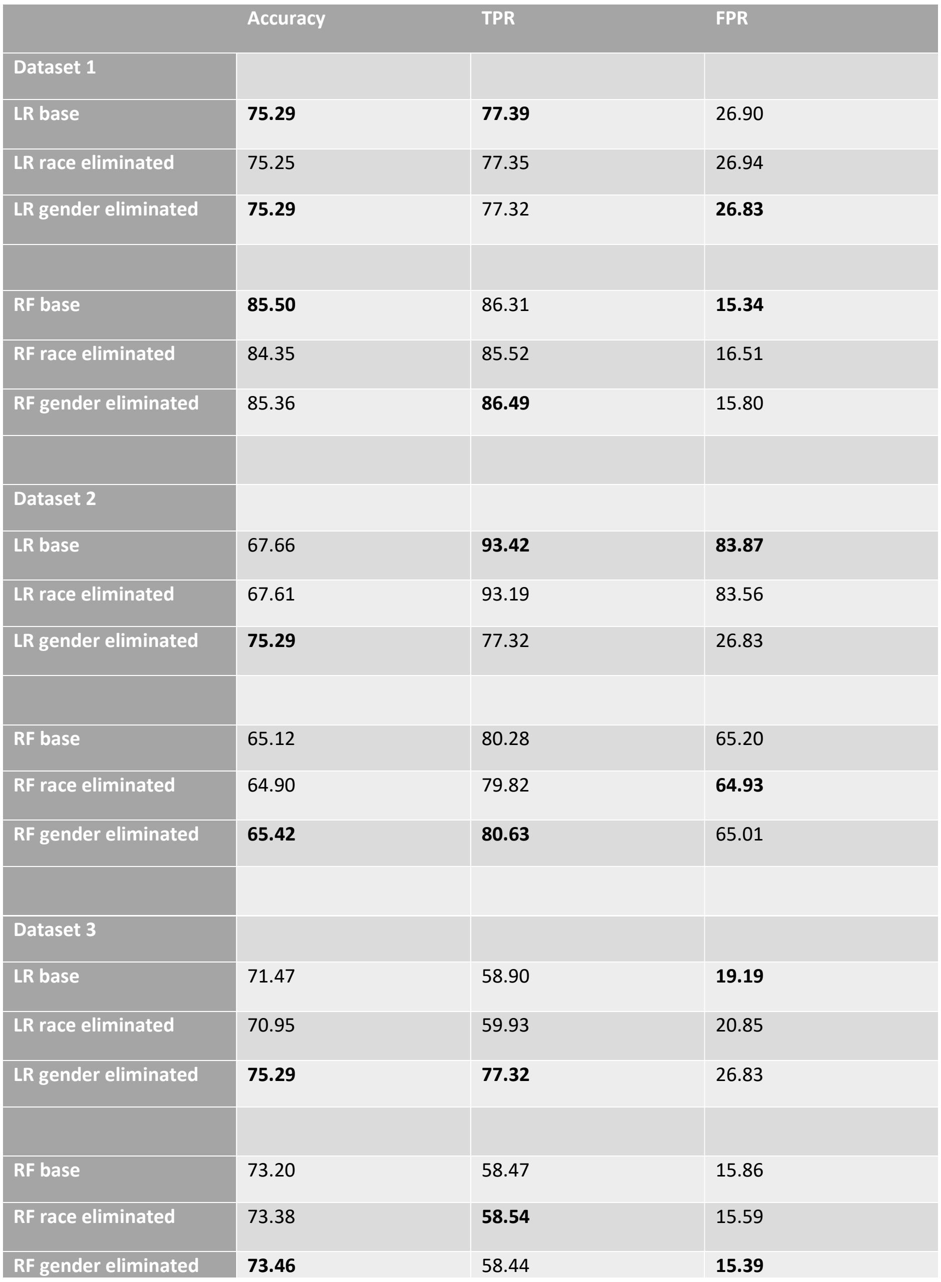
	Accuracy	TPR	FPR
Dataset 1			
LR base	75.29	77.39	26.90
LR SRM black	74.62	68.57	**20.31**
LR SRM white	75.50	80.05	31.03
LR SRM female	**77.01**	**85.88**	36.77
LR SRM male	74.04	72.02	23.95
RF base	85.50	86.31	15.34
RF SRM black	84.04	80.70	**13.06**
RF SRM white	85.82	51.60	20.75
RF SRM female	87.47	**94.21**	23.01
RF SRM male	**83.61**	81.86	14.65
Dataset 2			
LR base	67.66	93.42	**83.87**
LR SRM black	65.31	93.92	88.18
LR SRM white	66.44	94.51	87.65
LR SRM female	**70.52**	**94.91**	94.50
LR SRM male	67.01	83.18	84.89
RF base	65.12	80.28	65.20
RF SRM black	62.65	79.86	69.51
RF SRM white	63.72	55.97	87.12
RF SRM female	**68.03**	**83.79**	73.99
RF SRM male	63.53	78.02	**65.18**
Dataset 3			
LR base	71.47	58.90	19.19
LR SRM black	68.03	48.59	**18.10**
LR SRM white	**72.24**	62.80	20.56
LR SRM female	68.03	**69.85**	34.50
LR SRM male	69.54	54.29	20.05
RF base	**73.20**	58.47	15.86
RF SRM black	72.59	54.47	14.49
RF SRM white	72.52	63.44	**12.71**
RF SRM female	67.92	**70.03**	35.01
RF SRM male	72.37	53.60	14.83

Figure 9: Results from the Singular Race/Gender Models



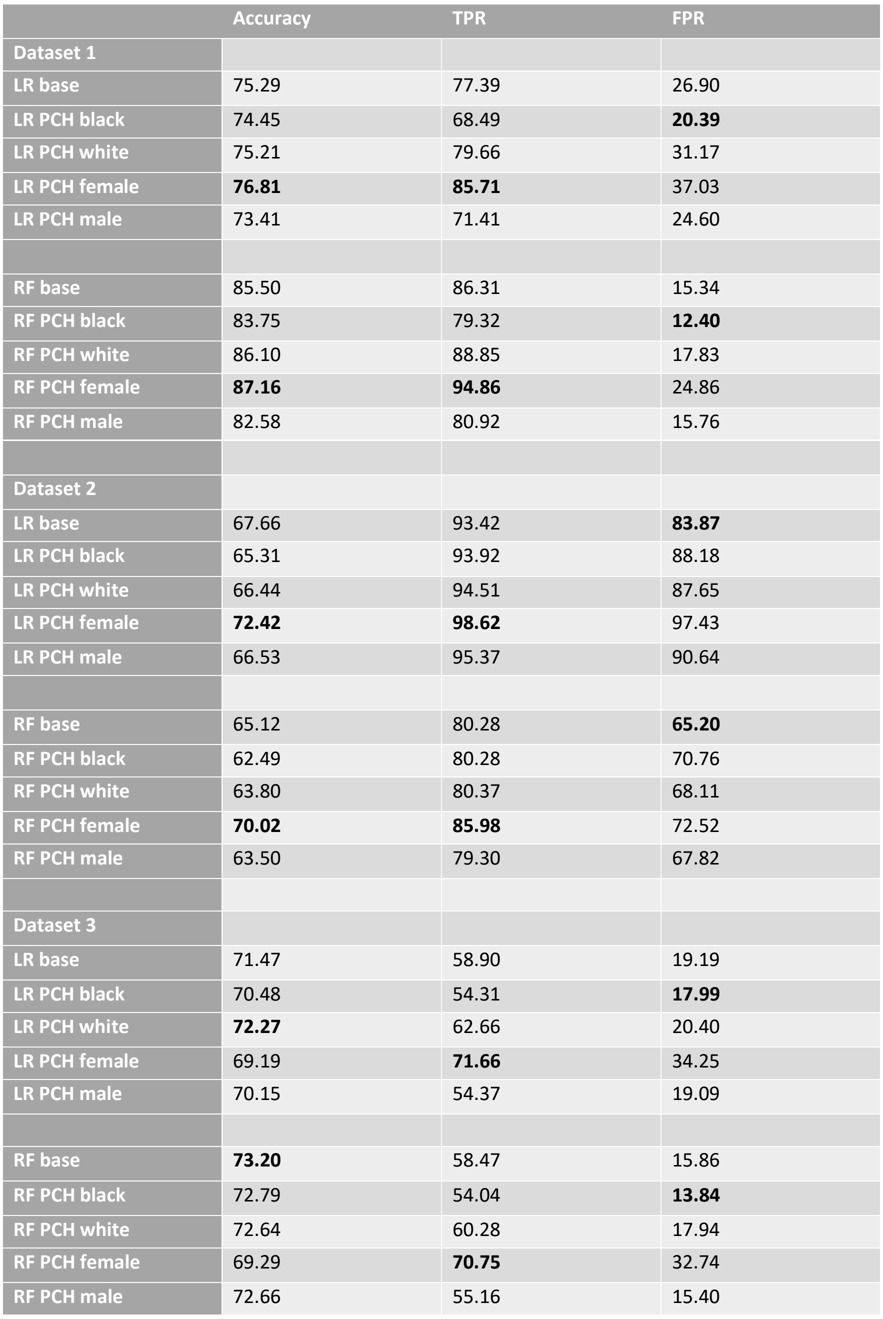
	Accuracy	TPR	FPR
Dataset 1			
LR base	75.29	77.39	26.90
LR balanced race	74.29	73.86	**25.21**
LR balanced gender	**76.24**	**78.65**	26.77
RF base	**85.50**	**86.31**	**15.34**
RF balanced race	83.25	83.94	17.52
RF balanced gender	80.64	81.73	20.72
Dataset 2			
LR base	67.66	93.42	83.87
LR balanced race	66.68	91.51	**81.76**
LR balanced gender	**69.88**	**95.33**	85.69
RF base	**65.12**	80.28	65.20
RF balanced race	64.10	78.76	**64.48**
RF balanced gender	64.98	**81.06**	70.11
Dataset 3			
LR base	**71.47**	58.90	**19.19**
LR balanced race	67.84	50.92	19.60
LR balanced gender	69.38	**66.70**	28.15
RF base	**73.20**	58.47	15.86
RF balanced race	73.09	57.72	**15.50**
RF balanced gender	70.22	**63.18**	23.30

Figure 10: Results from the Balanced Race/Gender Models



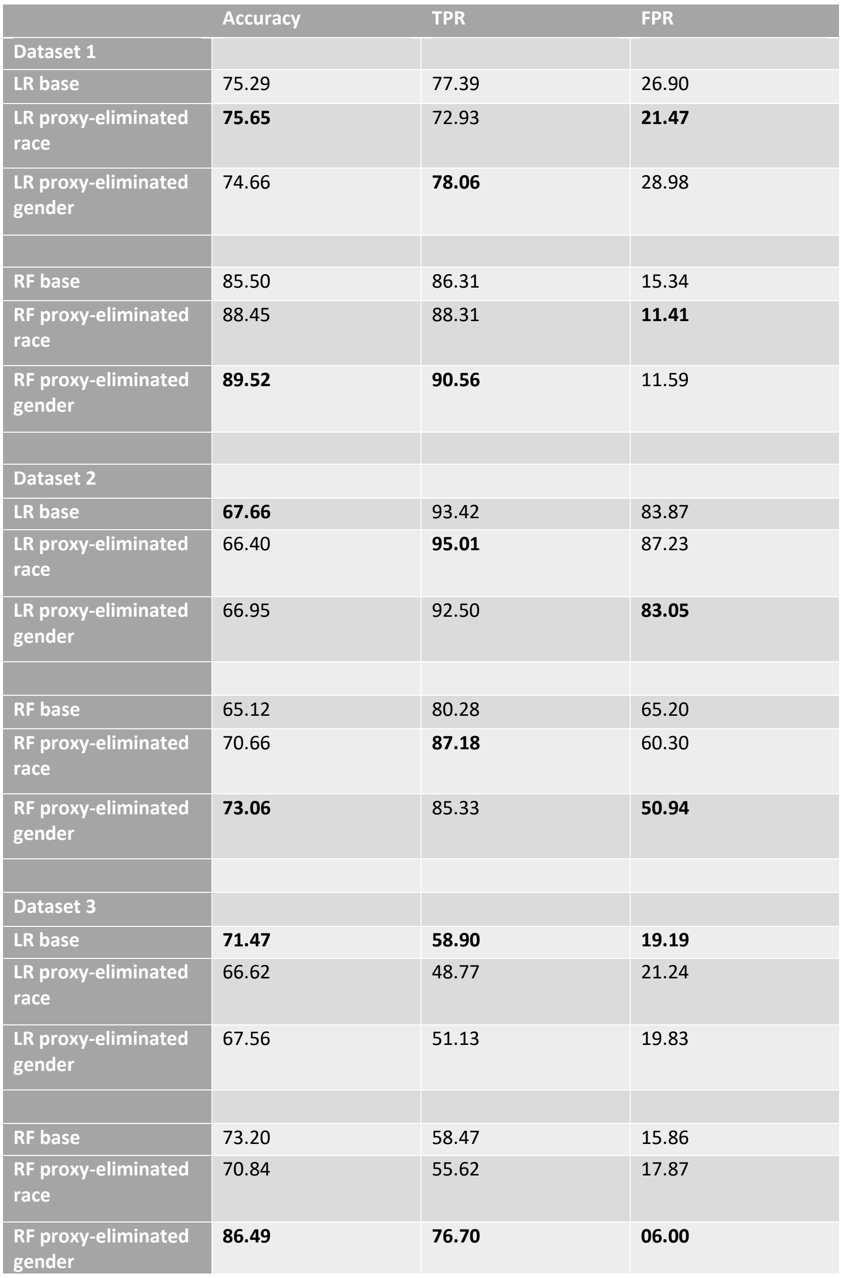
	Accuracy	TPR	FPR
Dataset 1			
LR base	**75.29**	**77.39**	26.90
LR race eliminated	75.25	77.35	26.94
LR gender eliminated	**75.29**	77.32	**26.83**
RF base	**85.50**	86.31	**15.34**
RF race eliminated	84.35	85.52	16.51
RF gender eliminated	85.36	**86.49**	15.80
Dataset 2			
LR base	67.66	**93.42**	**83.87**
LR race eliminated	67.61	93.19	83.56
LR gender eliminated	**75.29**	77.32	26.83
RF base	65.12	80.28	65.20
RF race eliminated	64.90	79.82	**64.93**
RF gender eliminated	**65.42**	**80.63**	65.01
Dataset 3			
LR base	71.47	58.90	**19.19**
LR race eliminated	70.95	59.93	20.85
LR gender eliminated	**75.29**	**77.32**	26.83
RF base	73.20	58.47	15.86
RF race eliminated	73.38	**58.54**	15.59
RF gender eliminated	**73.46**	58.44	**15.39**

Figure 11: Results from the Race/Gender Eliminated models



	Accuracy	TPR	FPR
Dataset 1			
LR base	75.29	77.39	26.90
LR PCH black	74.45	68.49	**20.39**
LR PCH white	75.21	79.66	31.17
LR PCH female	**76.81**	**85.71**	37.03
LR PCH male	73.41	71.41	24.60
RF base	85.50	86.31	15.34
RF PCH black	83.75	79.32	**12.40**
RF PCH white	86.10	88.85	17.83
RF PCH female	**87.16**	**94.86**	24.86
RF PCH male	82.58	80.92	15.76
Dataset 2			
LR base	67.66	93.42	**83.87**
LR PCH black	65.31	93.92	88.18
LR PCH white	66.44	94.51	87.65
LR PCH female	**72.42**	**98.62**	97.43
LR PCH male	66.53	95.37	90.64
RF base	65.12	80.28	**65.20**
RF PCH black	62.49	80.28	70.76
RF PCH white	63.80	80.37	68.11
RF PCH female	**70.02**	**85.98**	72.52
RF PCH male	63.50	79.30	67.82
Dataset 3			
LR base	71.47	58.90	19.19
LR PCH black	70.48	54.31	**17.99**
LR PCH white	**72.27**	62.66	20.40
LR PCH female	69.19	**71.66**	34.25
LR PCH male	70.15	54.37	19.09
RF base	**73.20**	58.47	15.86
RF PCH black	72.79	54.04	**13.84**
RF PCH white	72.64	60.28	17.94
RF PCH female	69.29	**70.75**	32.74
RF PCH male	72.66	55.16	15.40

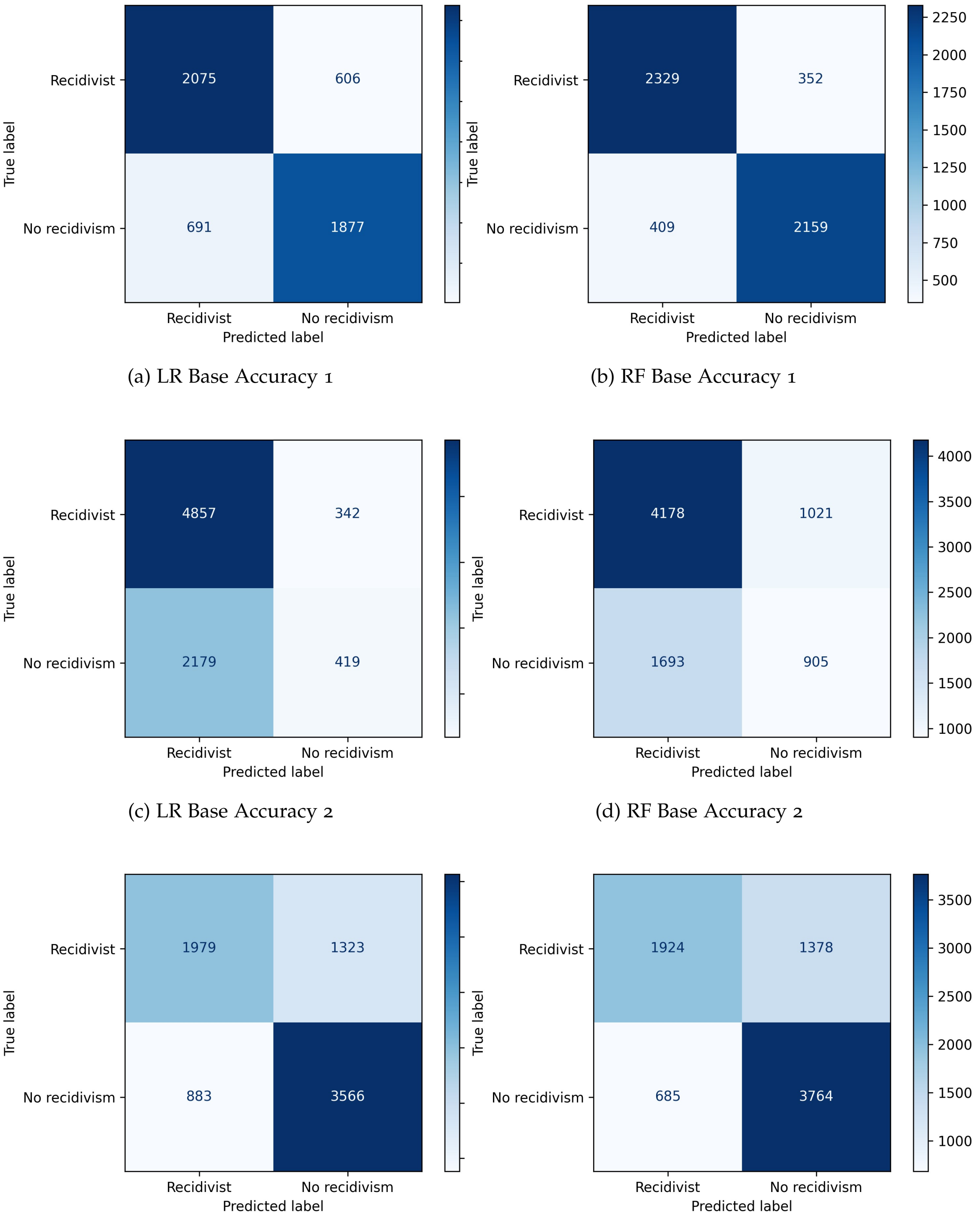
Figure 12: Results from the Past Criminal History Models



	Accuracy	TPR	FPR
Dataset 1			
LR base	75.29	77.39	26.90
LR proxy-eliminated race	**75.65**	72.93	**21.47**
LR proxy-eliminated gender	74.66	**78.06**	28.98
RF base	85.50	86.31	15.34
RF proxy-eliminated race	88.45	88.31	**11.41**
RF proxy-eliminated gender	**89.52**	**90.56**	11.59
Dataset 2			
LR base	**67.66**	93.42	83.87
LR proxy-eliminated race	66.40	**95.01**	87.23
LR proxy-eliminated gender	66.95	92.50	**83.05**
RF base	65.12	80.28	65.20
RF proxy-eliminated race	70.66	**87.18**	60.30
RF proxy-eliminated gender	**73.06**	85.33	**50.94**
Dataset 3			
LR base	**71.47**	**58.90**	**19.19**
LR proxy-eliminated race	66.62	48.77	21.24
LR proxy-eliminated gender	67.56	51.13	19.83
RF base	73.20	58.47	15.86
RF proxy-eliminated race	70.84	55.62	17.87
RF proxy-eliminated gender	**86.49**	**76.70**	**06.00**

Figure 13: Results from the Proxy-Eliminated Models

APPENDIX C: CONFUSION MATRICES OF BASE AND BIAS MITIGATED MODELS



(a) LR Base Accuracy 1

(b) RF Base Accuracy 1

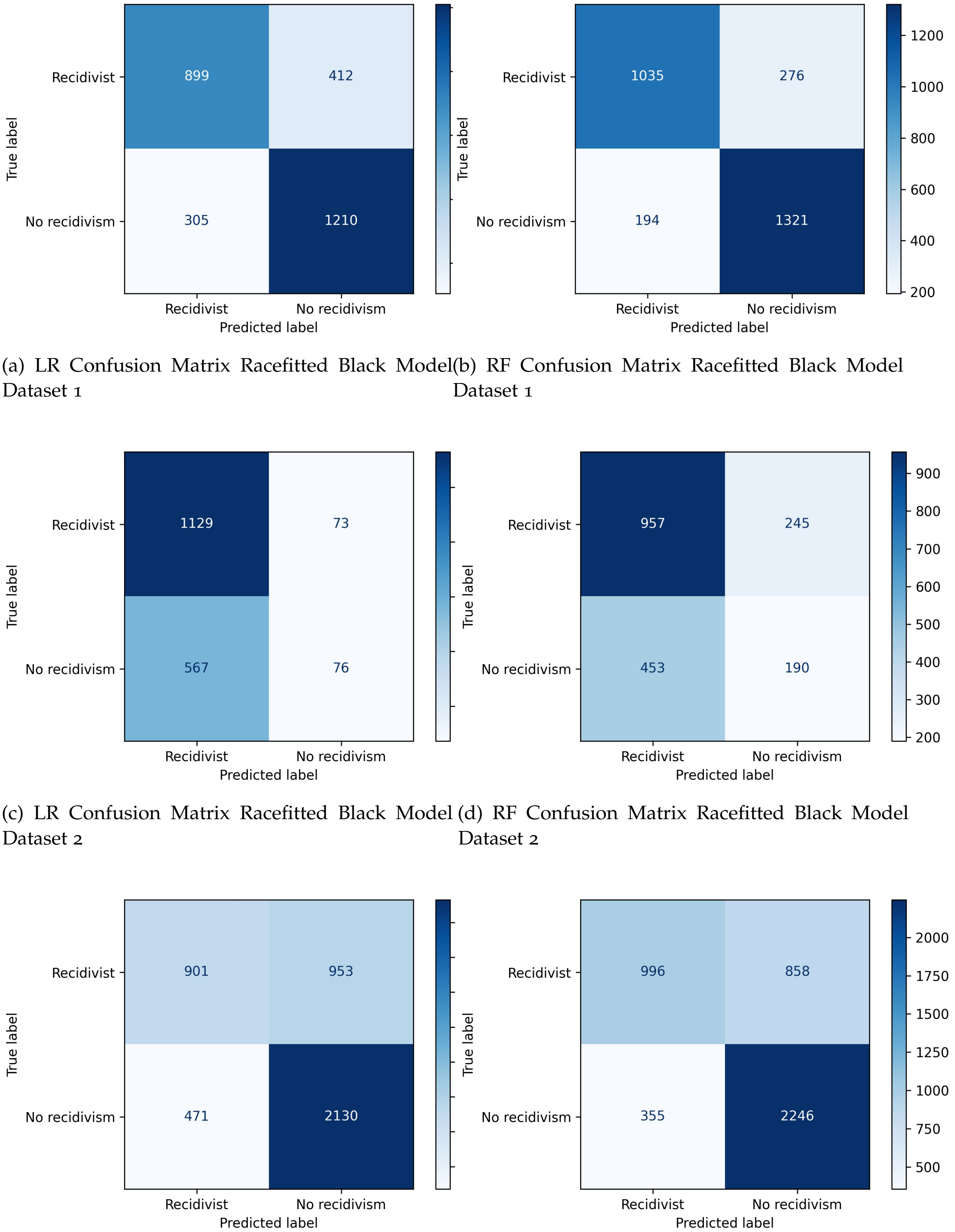
(c) LR Base Accuracy 2

(d) RF Base Accuracy 2

(e) RF Base Accuracy 2

(f) RF Base Accuracy 3

Figure 14: Base Models Confusion Matrices



(a) LR Confusion Matrix Racefitted Black Model Dataset 1

(b) RF Confusion Matrix Racefitted Black Model Dataset 1

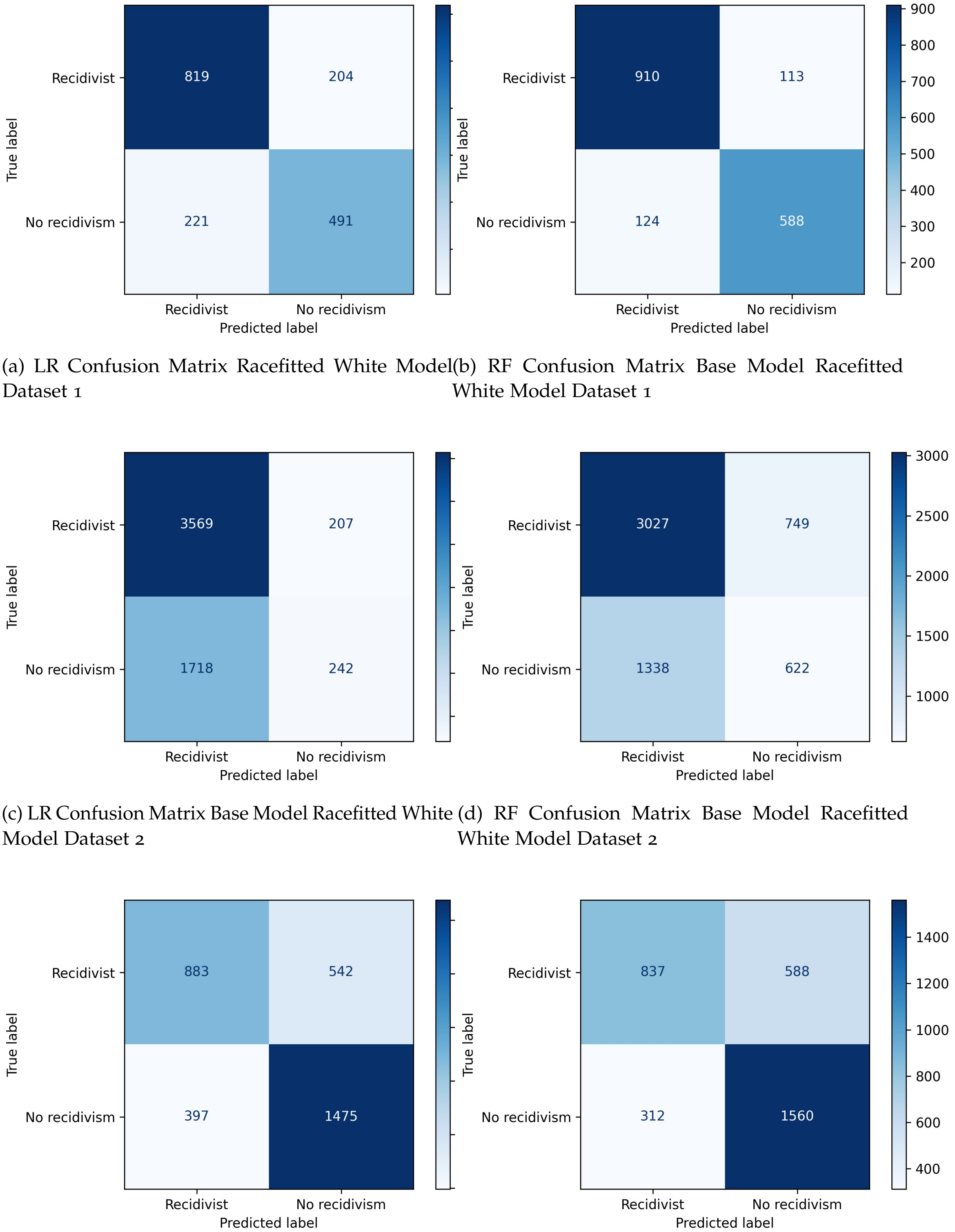
(c) LR Confusion Matrix Racefitted Black Model Dataset 2

(d) RF Confusion Matrix Racefitted Black Model Dataset 2

(e) LR Confusion Matrix Racefitted Black Model Dataset 3

(f) RF Confusion Matrix Racefitted Black Model Dataset 3

Figure 15: Race Fitted Black Models Confusion Matrices



(a) LR Confusion Matrix Racefitted White Model Dataset 1

(b) RF Confusion Matrix Base Model Racefitted White Model Dataset 1

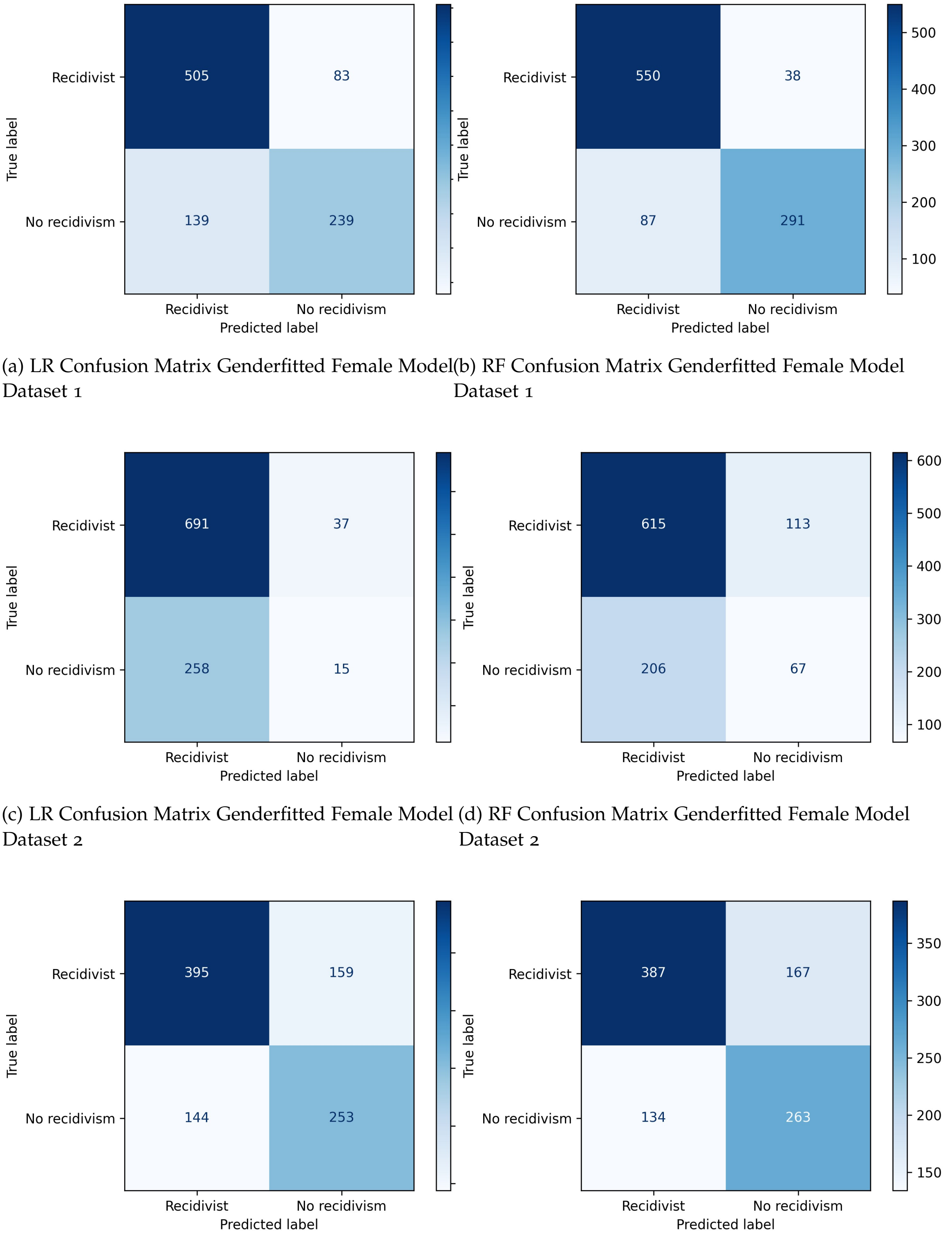
(c) LR Confusion Matrix Base Model Racefitted White Model Dataset 2

(d) RF Confusion Matrix Base Model Racefitted White Model Dataset 2

(e) LR Confusion Matrix Base Model Racefitted White Model Dataset 3

(f) RF Confusion Matrix Base Model Racefitted White Model Dataset 3

Figure 16: Race Fitted White Models Confusion Matrices



(a) LR Confusion Matrix Genderfitted Female Model Dataset 1

(b) RF Confusion Matrix Genderfitted Female Model Dataset 1

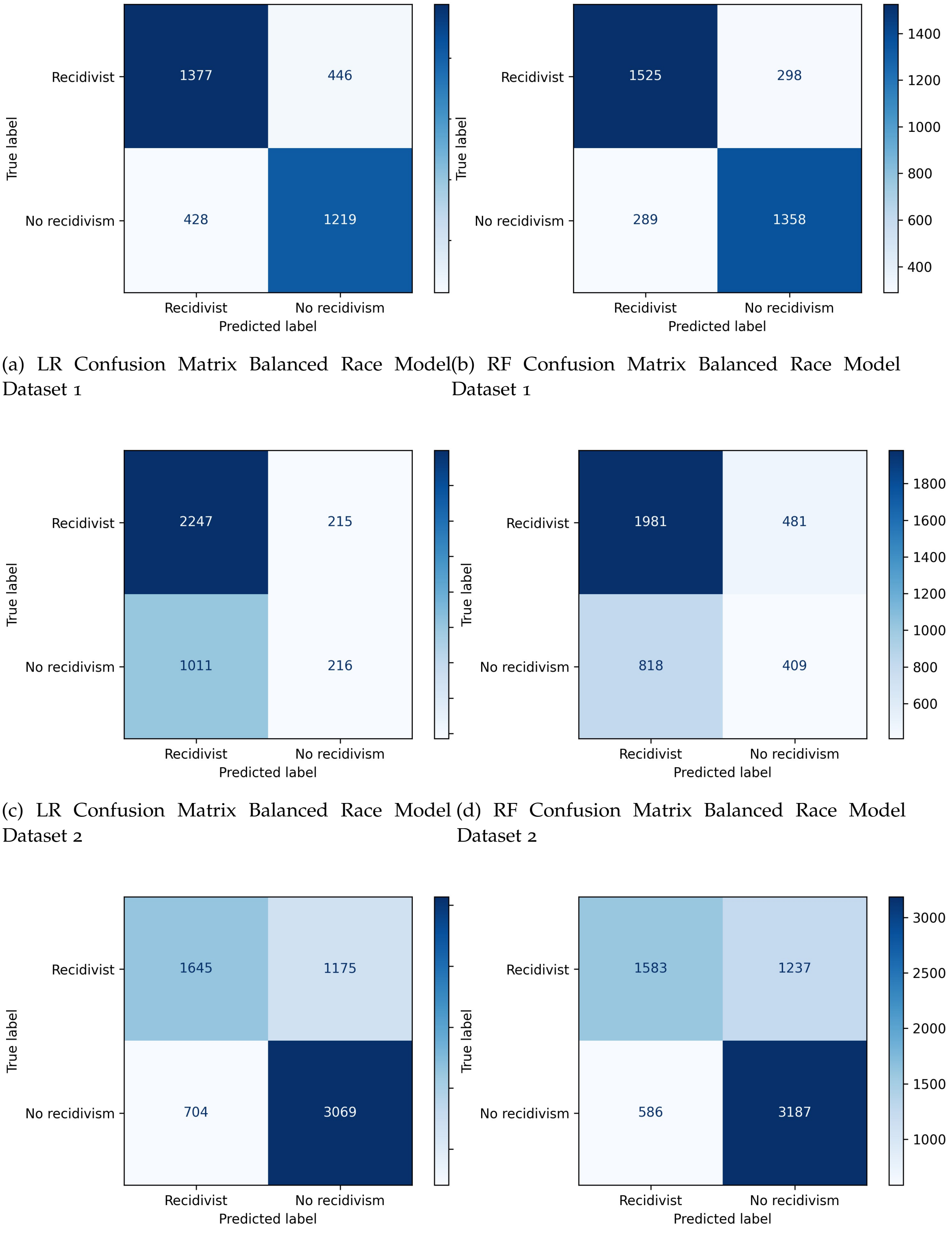
(c) LR Confusion Matrix Genderfitted Female Model Dataset 2

(d) RF Confusion Matrix Genderfitted Female Model Dataset 2

(e) LR Confusion Matrix Genderfitted Female Model Dataset 3

(f) RF Confusion Matrix Genderfitted Female Model Dataset 3

Figure 17: Gender Fitted Female Models Confusion Matrices



(a) LR Confusion Matrix Balanced Race Model Dataset 1

(b) RF Confusion Matrix Balanced Race Model Dataset 1

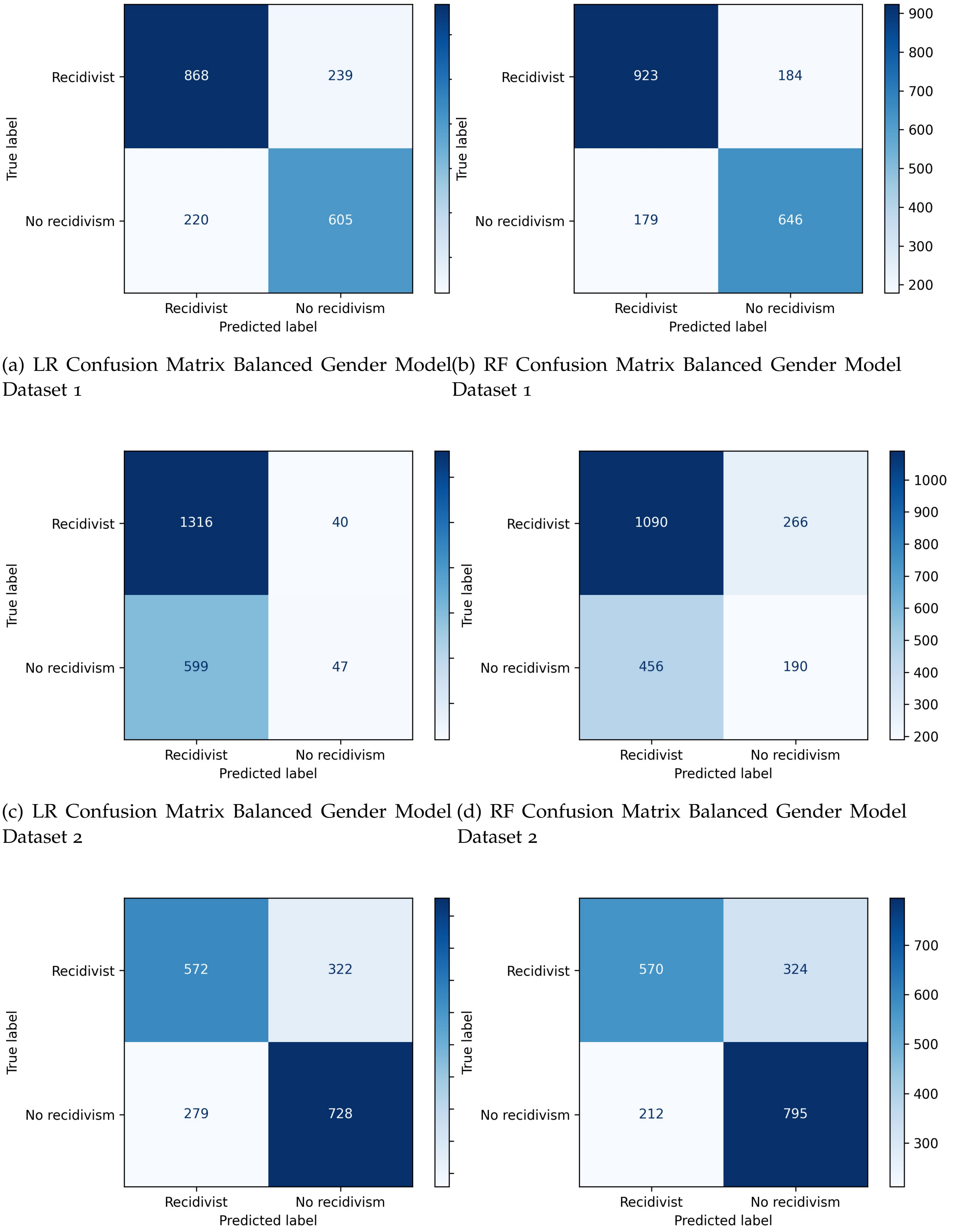
(c) LR Confusion Matrix Balanced Race Model Dataset 2

(d) RF Confusion Matrix Balanced Race Model Dataset 2

(e) LR Confusion Matrix Balanced Race Model Dataset 3

(f) RF Confusion Matrix Balanced Race Model Dataset 3

Figure 18: Balanced Race Models Confusion Matrices



(a) LR Confusion Matrix Balanced Gender Model Dataset 1

(b) RF Confusion Matrix Balanced Gender Model Dataset 1

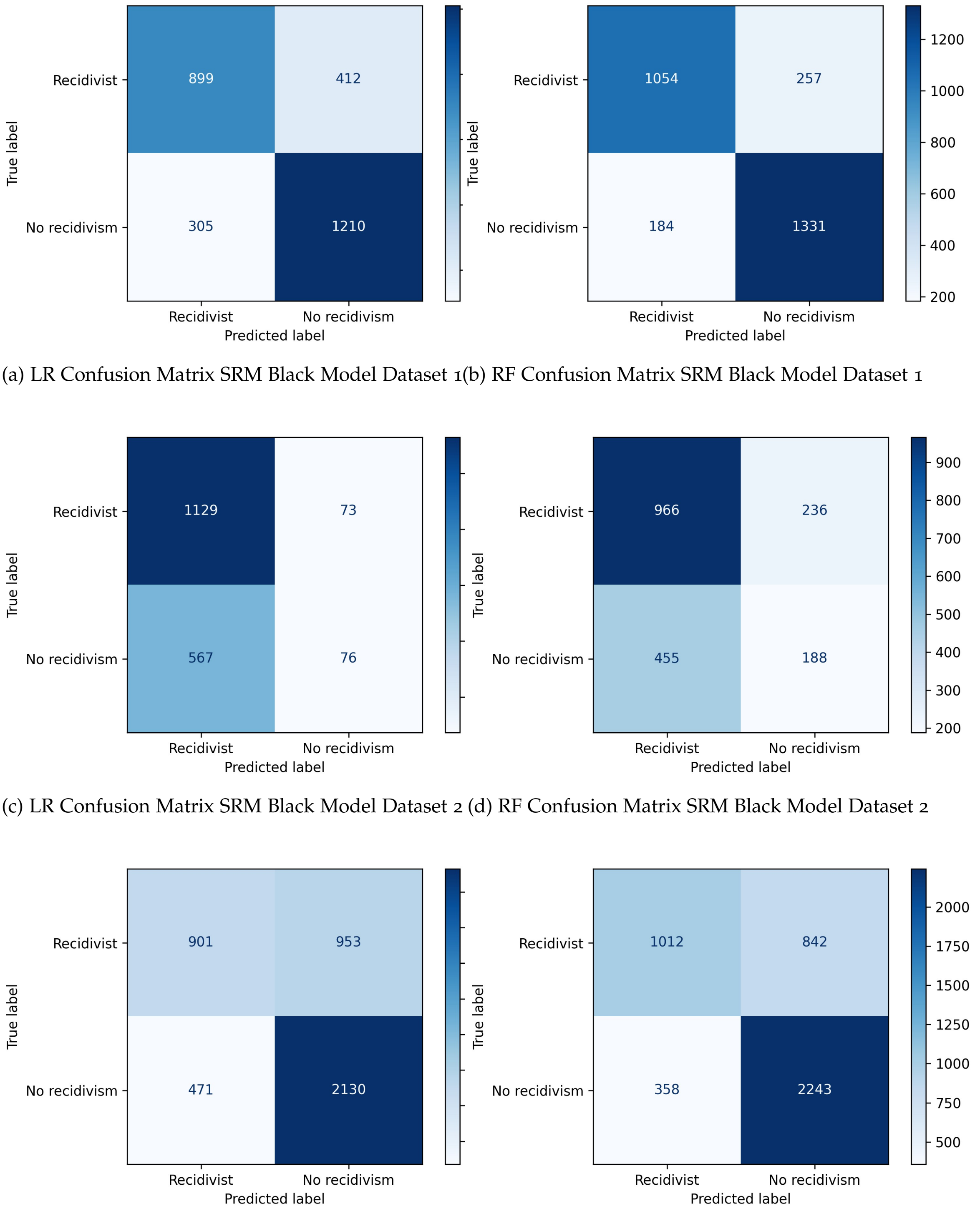
(c) LR Confusion Matrix Balanced Gender Model Dataset 2

(d) RF Confusion Matrix Balanced Gender Model Dataset 2

(e) LR Confusion Matrix Balanced Gender Model Dataset 3

(f) RF Confusion Matrix Balanced Gender Model Dataset 3

Figure 19: Balanced Gender Models Confusion Matrices

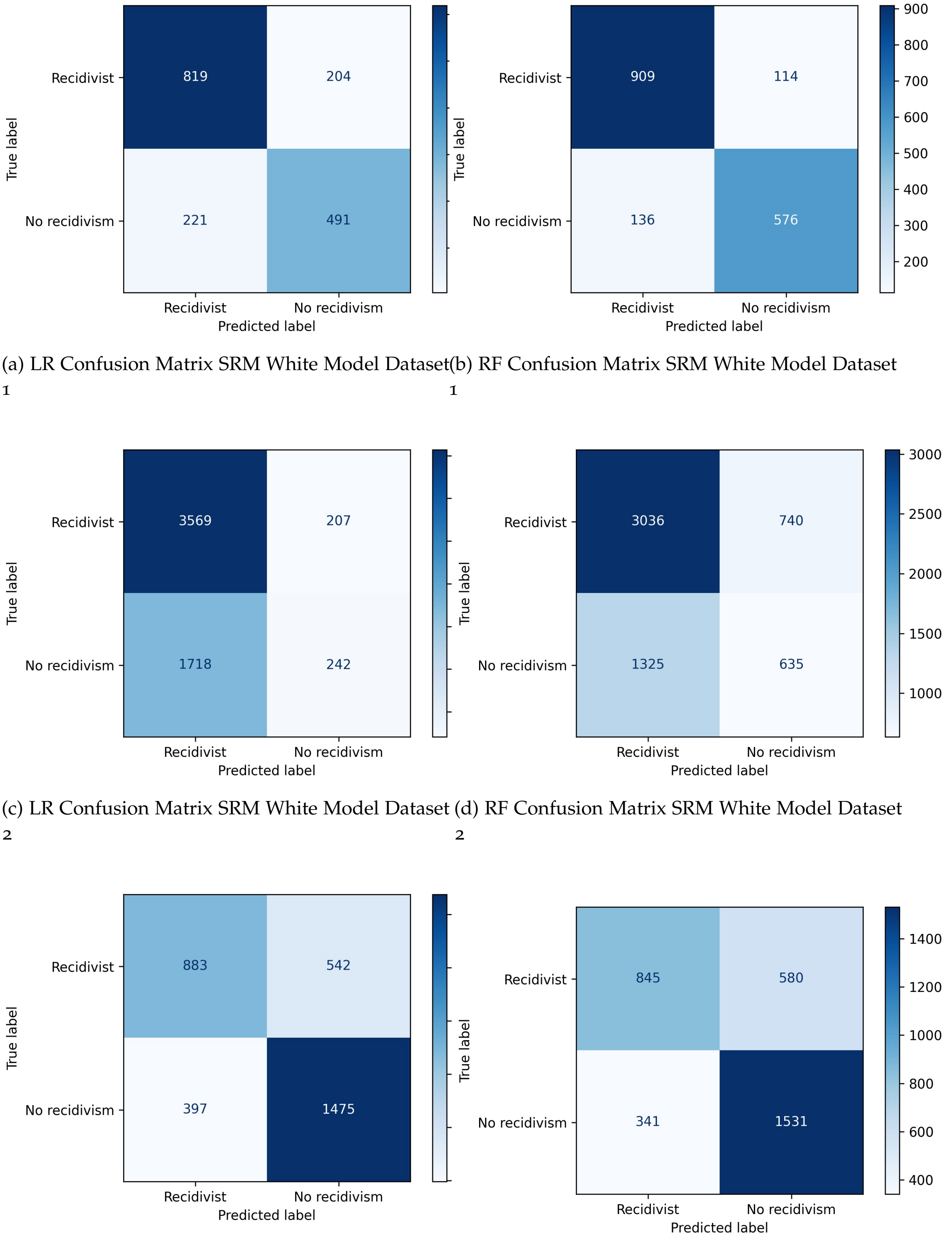


(a) LR Confusion Matrix SRM Black Model Dataset 1(b) RF Confusion Matrix SRM Black Model Dataset 1

(c) LR Confusion Matrix SRM Black Model Dataset 2 (d) RF Confusion Matrix SRM Black Model Dataset 2

(e) LR Confusion Matrix SRM Black Model Dataset 3 (f) RF Confusion Matrix SRM Black Model Dataset 3

Figure 20: Single Race Method Black Models Confusion Matrices



(a) LR Confusion Matrix SRM White Model Dataset 1

(b) RF Confusion Matrix SRM White Model Dataset 1

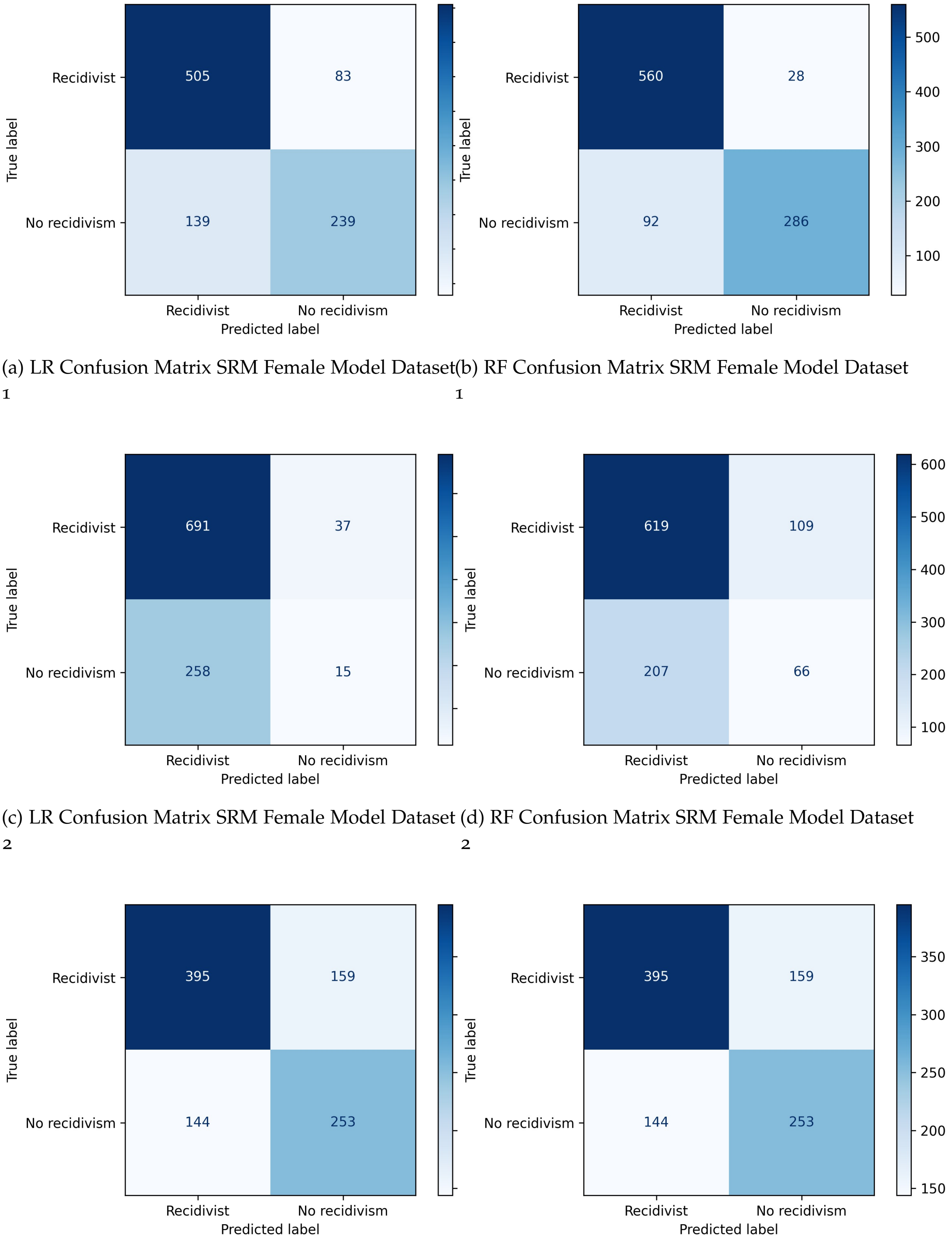
(c) LR Confusion Matrix SRM White Model Dataset 2

(d) RF Confusion Matrix SRM White Model Dataset 2

(e) LR Confusion Matrix SRM White Model Dataset 3

(f) RF Confusion Matrix SRM White Model Dataset 3

Figure 21: Single Race Method White Models Confusion Matrices



(a) LR Confusion Matrix SRM Female Model Dataset 1

(b) RF Confusion Matrix SRM Female Model Dataset 1

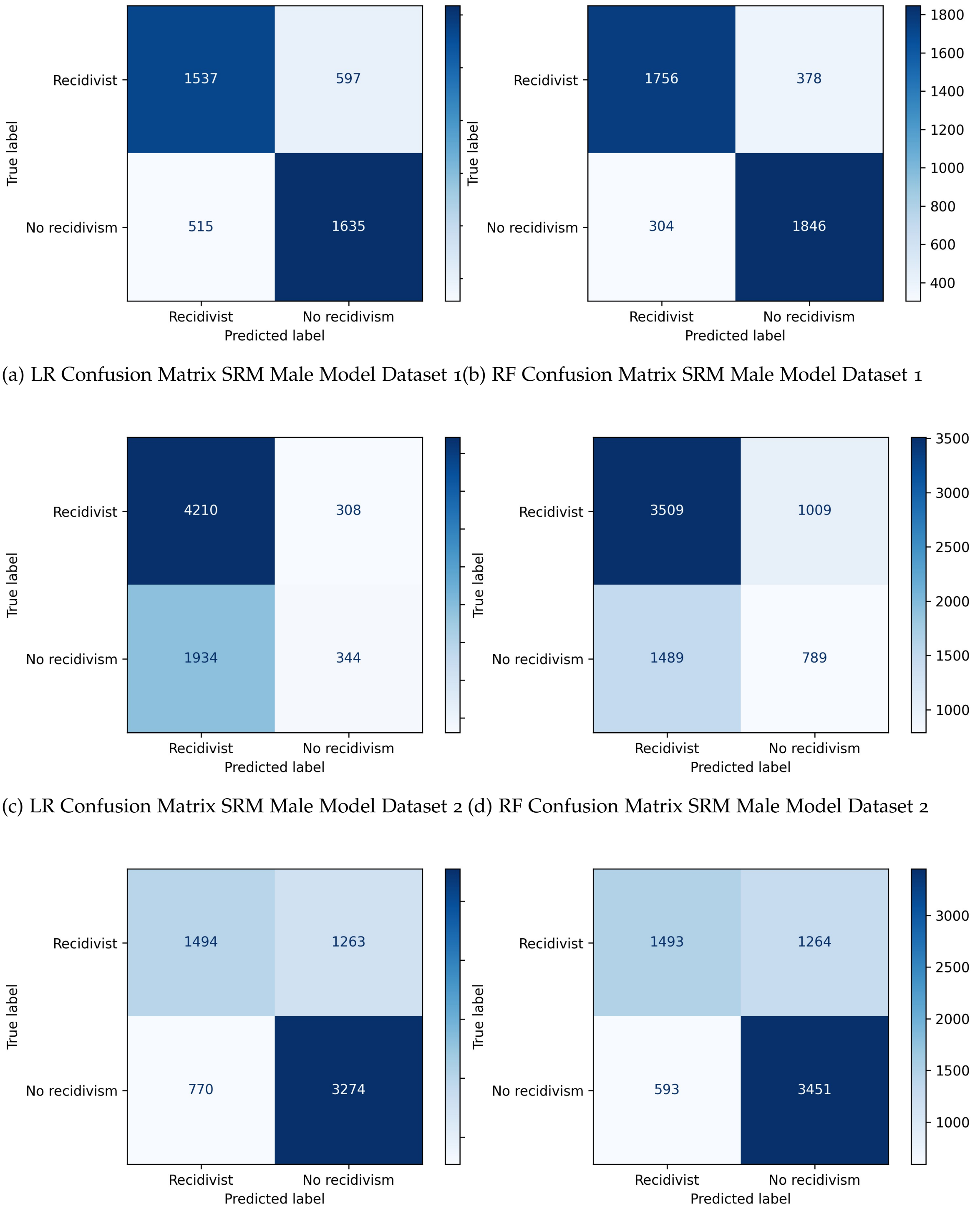
(c) LR Confusion Matrix SRM Female Model Dataset 2

(d) RF Confusion Matrix SRM Female Model Dataset 2

(e) LR Confusion Matrix SRM Female Model Dataset 3

(f) RF Confusion Matrix SRM Female Model Dataset 3

Figure 22: Single Race Method Female Models Confusion Matrices



(a) LR Confusion Matrix SRM Male Model Dataset 1

(b) RF Confusion Matrix SRM Male Model Dataset 1

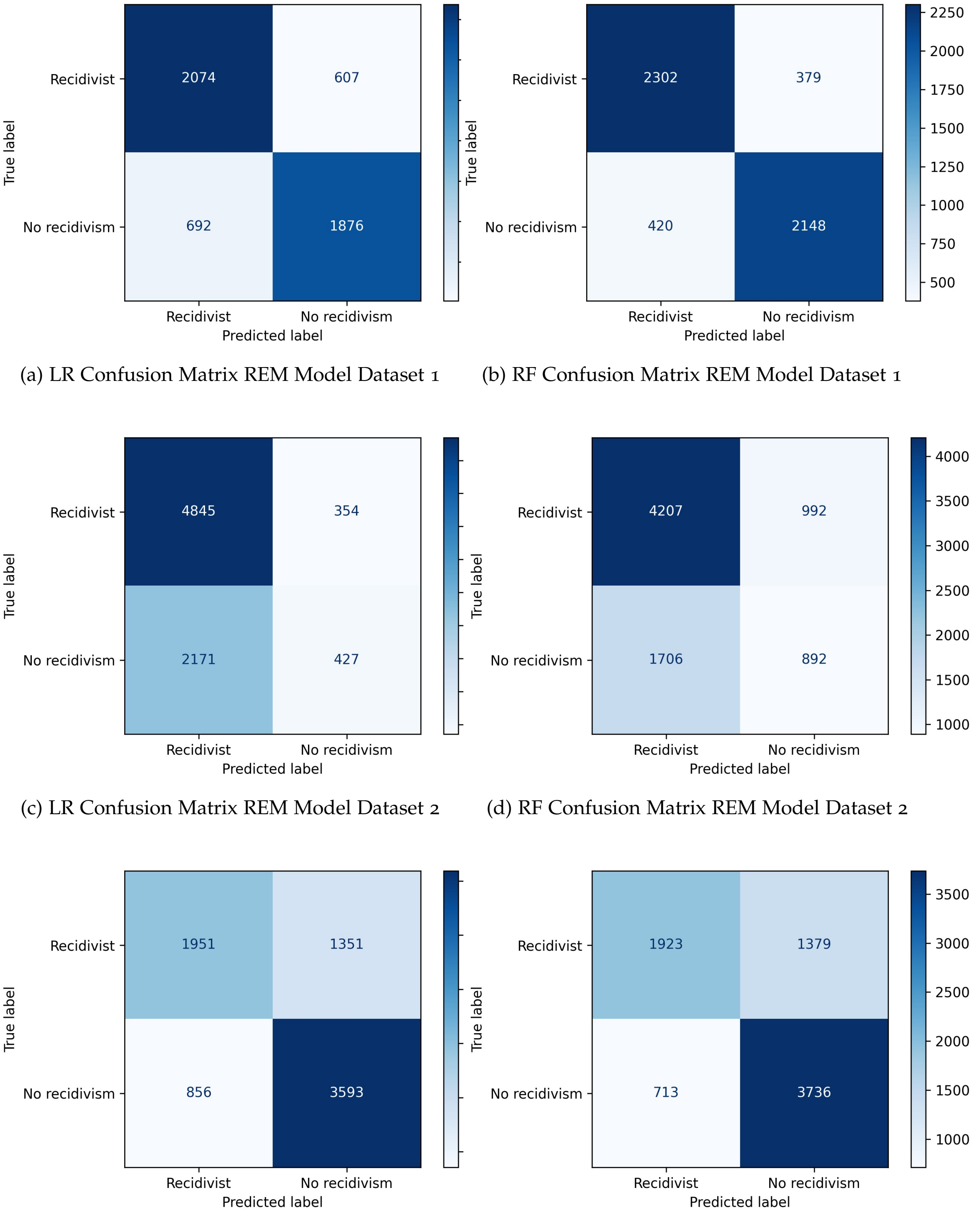
(c) LR Confusion Matrix SRM Male Model Dataset 2

(d) RF Confusion Matrix SRM Male Model Dataset 2

(e) LR Confusion Matrix SRM Male Model Dataset 3

(f) RF Confusion Matrix SRM Male Model Dataset 3

Figure 23: Single Race Method Male Models Confusion Matrices



(a) LR Confusion Matrix REM Model Dataset 1

(b) RF Confusion Matrix REM Model Dataset 1

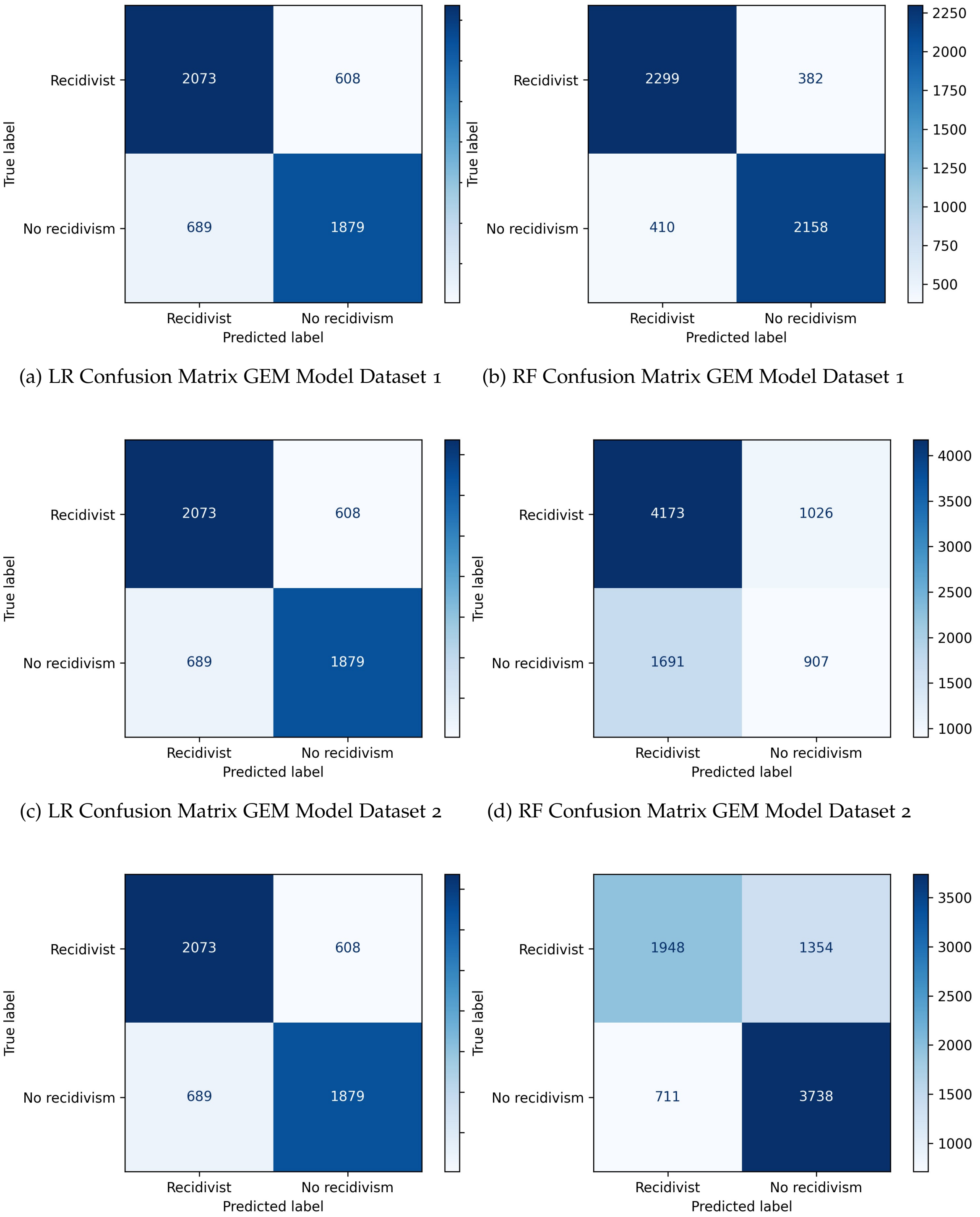
(c) LR Confusion Matrix REM Model Dataset 2

(d) RF Confusion Matrix REM Model Dataset 2

(e) LR Confusion Matrix REM Model Dataset 3

(f) RF Confusion Matrix REM Model Dataset 3

Figure 24: Race Eliminated Method Models Confusion Matrices



(a) LR Confusion Matrix GEM Model Dataset 1

(b) RF Confusion Matrix GEM Model Dataset 1

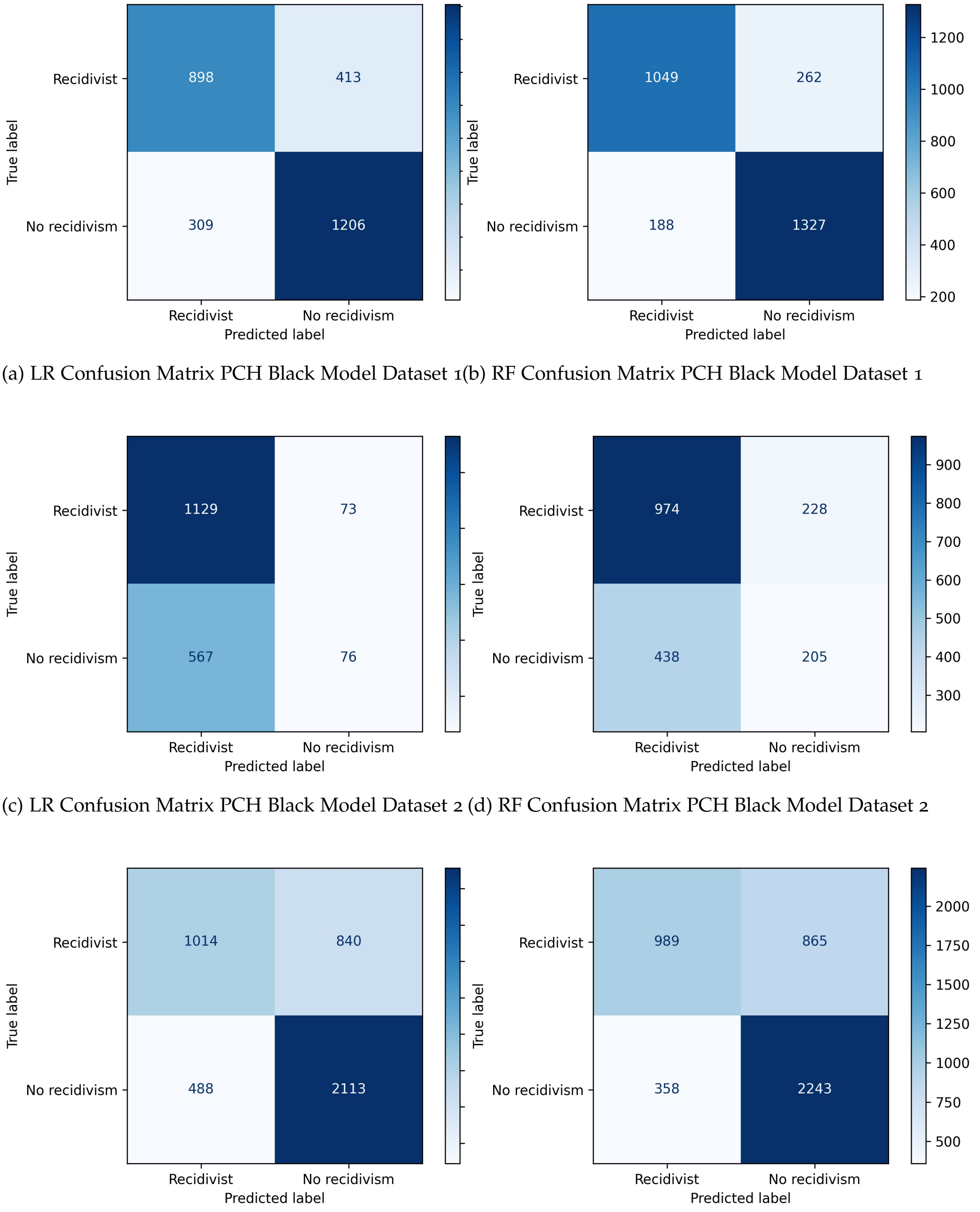
(c) LR Confusion Matrix GEM Model Dataset 2

(d) RF Confusion Matrix GEM Model Dataset 2

(e) LR Confusion Matrix GEM Model Dataset 3

(f) RF Confusion Matrix GEM Model Dataset 3

Figure 25: Gender Eliminated Method Models Confusion Matrices

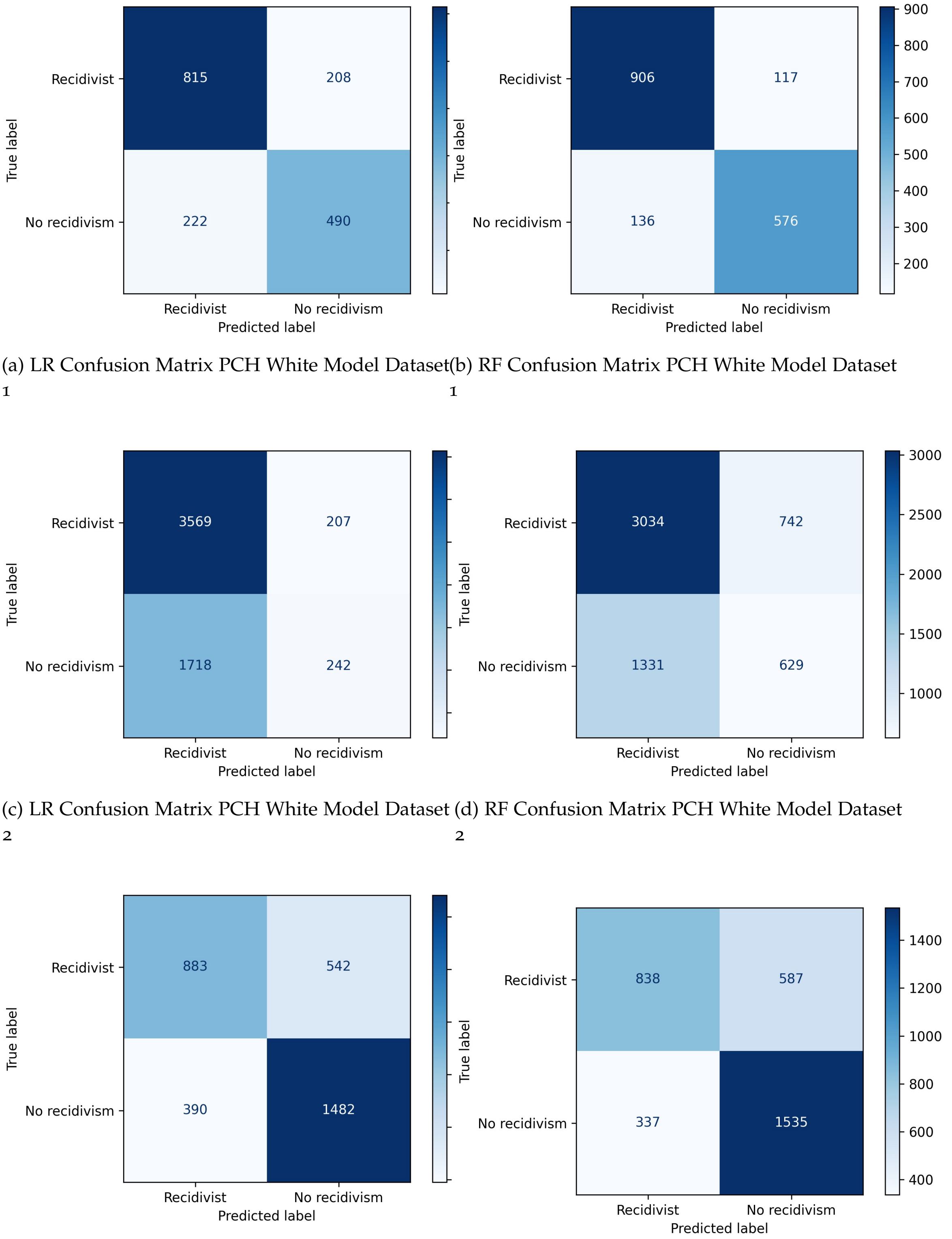


(a) LR Confusion Matrix PCH Black Model Dataset 1(b) RF Confusion Matrix PCH Black Model Dataset 1

(c) LR Confusion Matrix PCH Black Model Dataset 2 (d) RF Confusion Matrix PCH Black Model Dataset 2

(e) LR Confusion Matrix PCH Black Model Dataset 3 (f) RF Confusion Matrix PCH Black Model Dataset 3

Figure 26: Past Criminal History Black Models Confusion Matrices



(a) LR Confusion Matrix PCH White Model Dataset 1

(b) RF Confusion Matrix PCH White Model Dataset 1

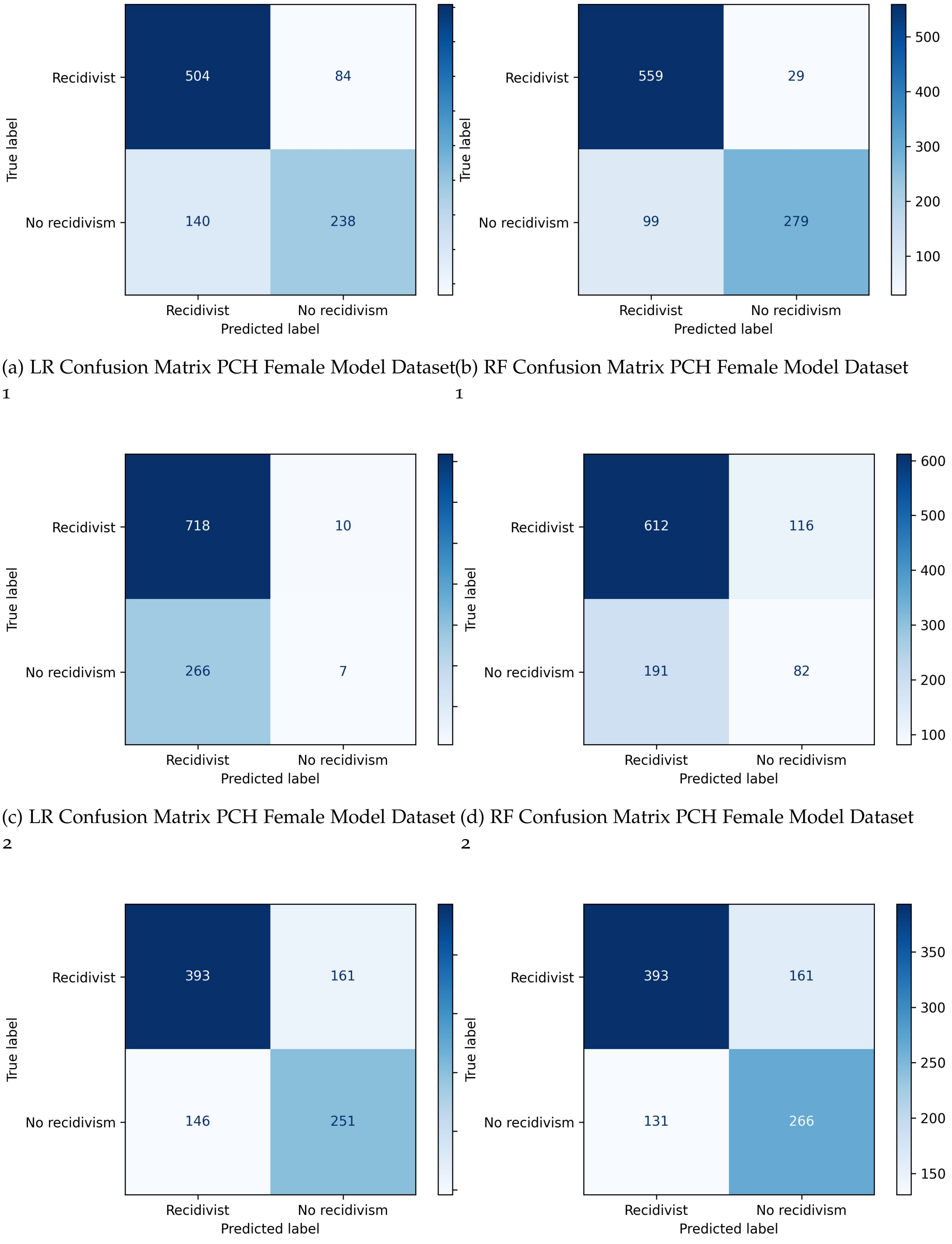
(c) LR Confusion Matrix PCH White Model Dataset 2

(d) RF Confusion Matrix PCH White Model Dataset 2

(e) LR Confusion Matrix PCH White Model Dataset 3

(f) RF Confusion Matrix PCH White Model Dataset 3

Figure 27: Past Criminal History White Models Confusion Matrices



(a) LR Confusion Matrix PCH Female Model Dataset 1

(b) RF Confusion Matrix PCH Female Model Dataset 1

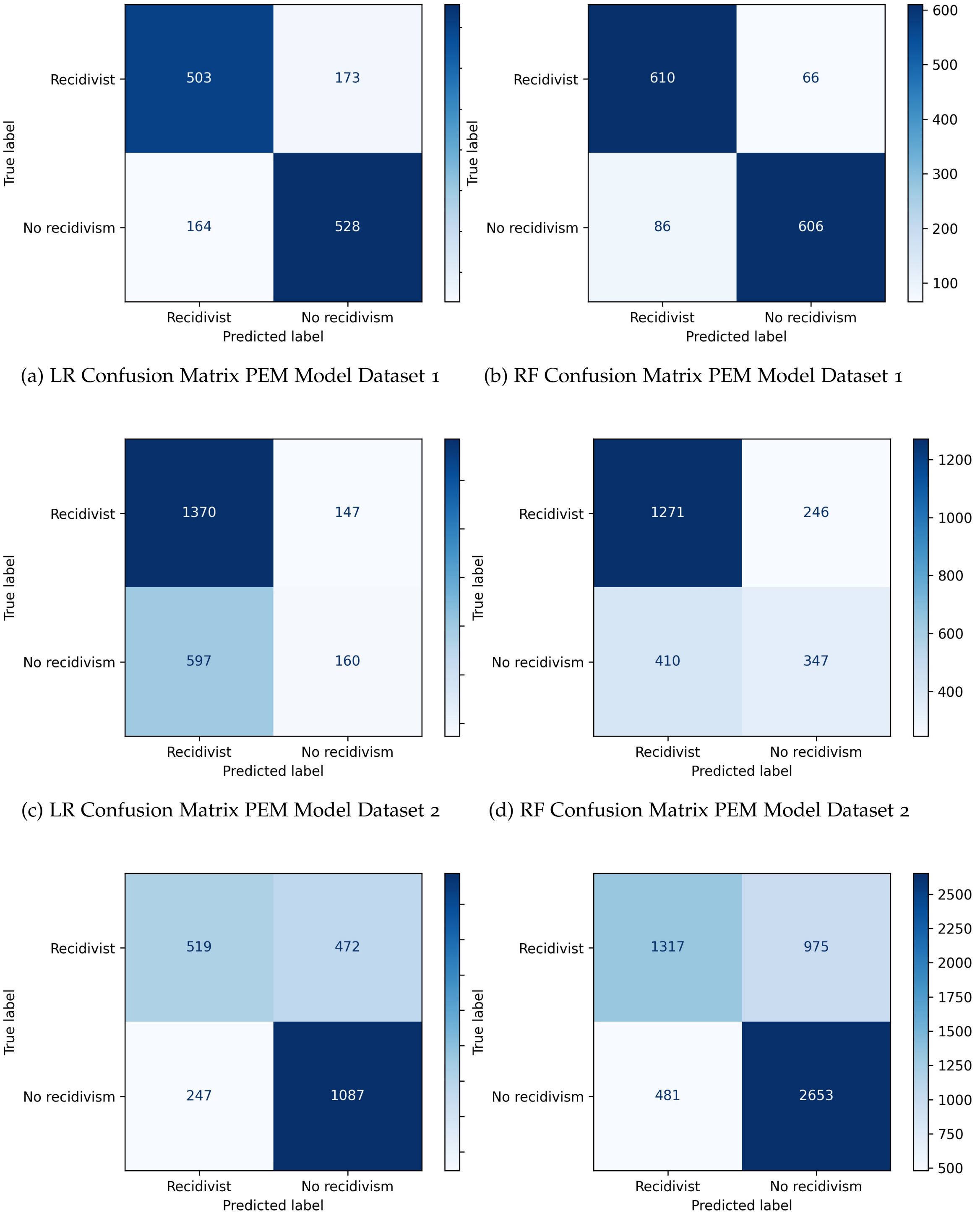
(c) LR Confusion Matrix PCH Female Model Dataset 2

(d) RF Confusion Matrix PCH Female Model Dataset 2

(e) LR Confusion Matrix PCH Female Model Dataset 3

(f) RF Confusion Matrix PCH Female Model Dataset 3

Figure 28: Past Criminal History Female Models Confusion Matrices



(a) LR Confusion Matrix PEM Model Dataset 1

(b) RF Confusion Matrix PEM Model Dataset 1

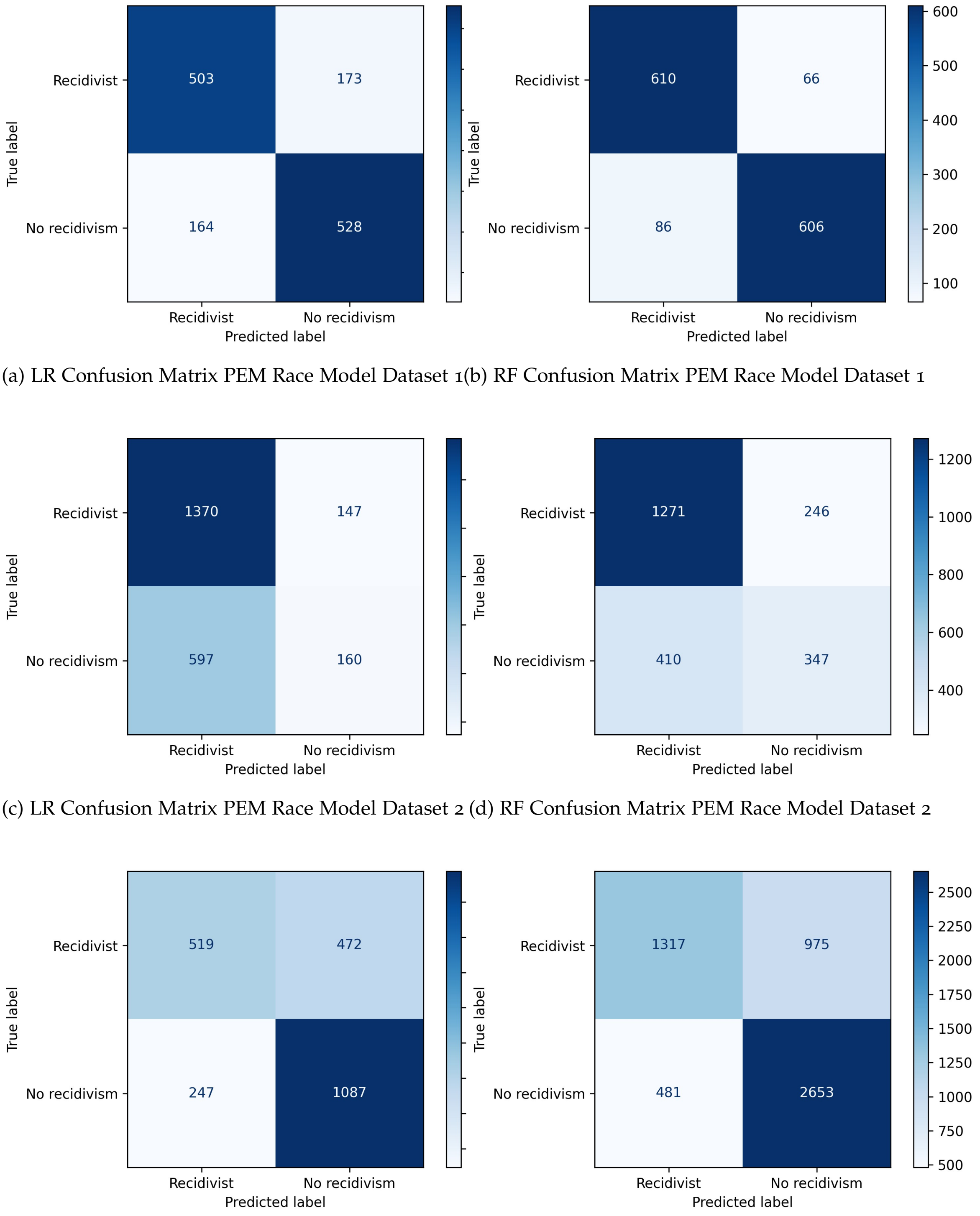
(c) LR Confusion Matrix PEM Model Dataset 2

(d) RF Confusion Matrix PEM Model Dataset 2

(e) LR Confusion Matrix PEM Model Dataset 3

(f) RF Confusion Matrix PEM Model Dataset 3

Figure 29: Proxy Eliminated Method Models Confusion Matrices

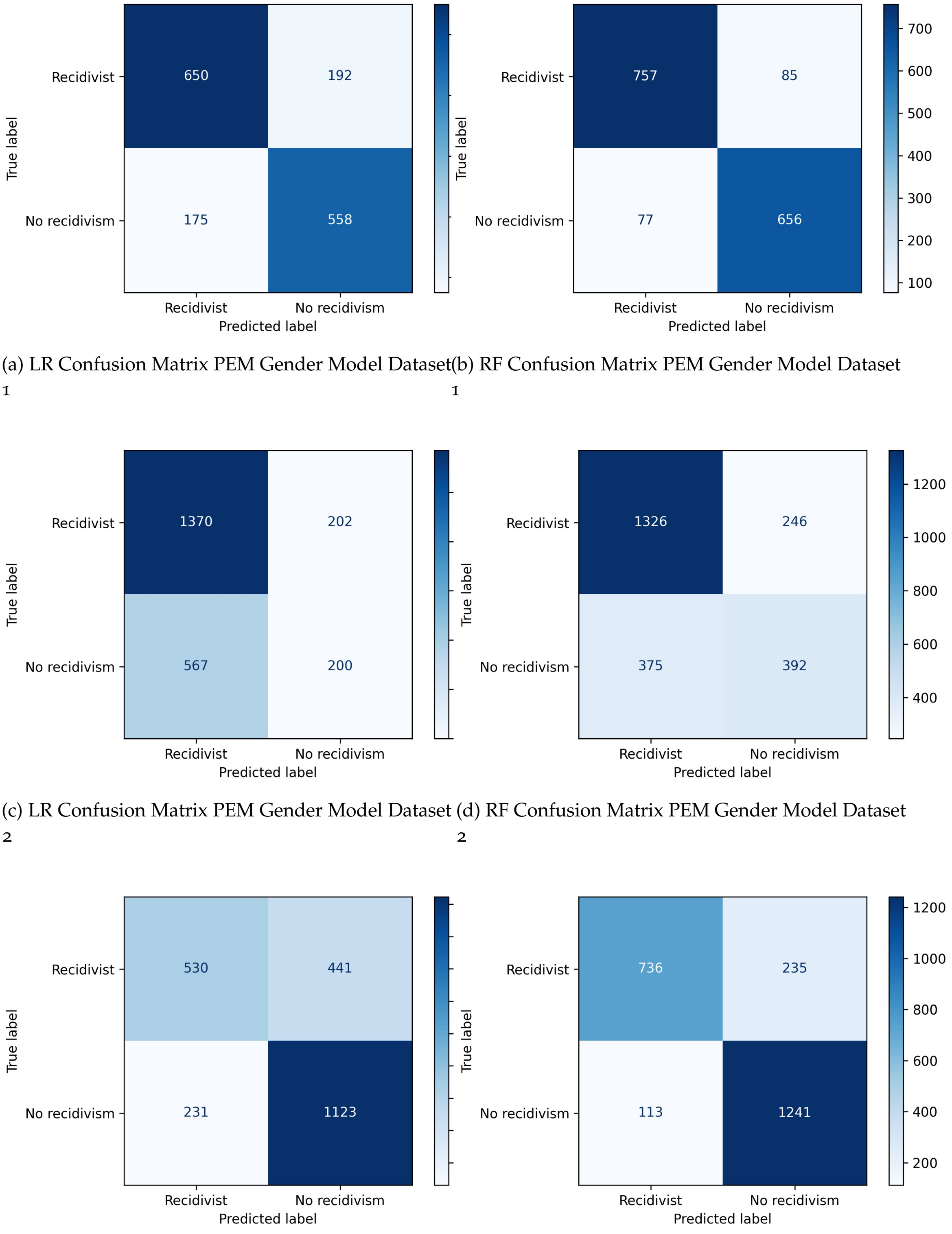


(a) LR Confusion Matrix PEM Race Model Dataset 1(b) RF Confusion Matrix PEM Race Model Dataset 1

(c) LR Confusion Matrix PEM Race Model Dataset 2 (d) RF Confusion Matrix PEM Race Model Dataset 2

(e) LR Confusion Matrix PEM Race Model Dataset 3 (f) RF Confusion Matrix PEM Race Model Dataset 3

Figure 30: Proxy Eliminated Method Race Models Confusion Matrices



(a) LR Confusion Matrix PEM Gender Model Dataset 1

(b) RF Confusion Matrix PEM Gender Model Dataset 1

(c) LR Confusion Matrix PEM Gender Model Dataset 2

(d) RF Confusion Matrix PEM Gender Model Dataset 2

(e) LR Confusion Matrix PEM Gender Model Dataset 3

(f) RF Confusion Matrix PEM Gender Model Dataset 3

Figure 31: Proxy Eliminated Method Gender Models Confusion Matrices