Universiteit
Leiden
The Netherlands

Opleiding Informatica & Economie

Fairness in the use of

AI-based algorithms in justice systems

Jort Visser

Supervisors:
Rüya Koçer & Lu Cao

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

25/06/2024

Abstract

Machine learning algorithms have gained a large role within decision-making processes in recent years. Within criminal justice, algorithms like COMPAS are used. This software tool is made to predict recidivism, but turned out to be discriminatory towards African American people. In this thesis, the landscape of this unfairness is studied. Furthermore, algorithms that are similar to COMPAS are implemented to investigate how the unfairness emerges and for which demographic groups. The results show that several groups are very dependent on which choices are made in the development of the algorithm. Besides this, some groups seem to be advantaged or disadvantaged with all the prediction models. Groups that contain one or more features like young, male, and African American were disadvantaged mostly. Features like old, female, and white caused a disadvantage for certain groups.

Contents

1 Introduction

Should AI-based algorithms influence decision-making processes in our justice system? For the last couple of years, AI has had a massive impact on our daily lives. Algorithms are commonly used for recommendations of food, movies and who to date [11]. Besides this, these algorithms are also used for more important things, such as recruiting new people for jobs and deciding whether to give a mortgage [4, 13]. These developments have established because the efficiency of these decision-making models are relatively high compared to the methods that were used before. However, these decision-making models could have a huge impact on people's daily lives, and therefore, these algorithms should make fair predictions. Researchers have discovered that there have been many cases where AI-based prediction models are biased and do not make fair predictions [2]. A 'fair algorithm' is an algorithm that is not biased. This means that the algorithm does not cause advantages or disadvantages for specific groups based on sensitive characteristics, such as gender, race and age. Unfortunately, there are decision-making models that are 'unfair'. For instance, a software tool was built to judge attendees of a beauty contest, but it was biased against dark tinted attendees. A facial recognition algorithm in digital cameras is also a known example, because it predicts many Asian people as blinking [11]. There are also discriminatory algorithms that produce unfair predictions regarding criminal justice, which could have a huge impact on people's lives [5]. Researchers are working on this problem by determining how to measure if an algorithm is fair or not. This is difficult because there is not just one way to define fairness. S. Verma, et al. [16] explained twenty notions of definitions of fairness and compared them with each other by demonstrating them on a single unifying case-study. However, computer scientists could still not agree on which definition is the best one.

In this bachelor thesis, the computational ethics of the use of AI-based algorithms in justice systems will be studied. This thesis is mainly focused on the nature of the unfairness in algorithms within justice systems. A demonstration of the challenges from this will be given, which will be done by focusing on a prominent case named COMPAS. An overview of COMPAS will be given, after which an experiment with a dataset of COMPAS follows. After this, an investigation will follow to determine who will be affected by unfairness in AI-based models. Furthermore, a discussion will be conducted regarding the options that arise during the development of such a model and the influences these options have on the outcome of the model. The goal of this thesis is to give an overview of the ongoing debate of this problem and to discuss my results of the experiments. An analytical exposition of the problem will be provided, but this does not includes a discussion of the possible solutions for this. The research question of this thesis is formulated as follows.

How and for whom may the unfairness emerge in AI-based algorithms used in decision-making processes in justice systems?

1.1 Thesis overview

Section 2 includes the theory; Section 3 discusses COMPAS and explains the case-study; Section 4 describes the the methodology of the experiments and section 5 show their outcome; These outcomes will be discussed in section 6. Section 7 concludes.

2 Theory

2.1 Bias and Fairness in AI-based algorithms

Decision-making models are very dependent on data, because they are trained by it. This means that the training data has a big impact on the functionality of these algorithms. When the training data contains biases, algorithms will learn these biases too. Moreover, the models can amplify the biases in training data. When the data that is generated from biased algorithms is used to train new algorithms, these biased data will result in more biased algorithms. In addition, algorithms can even show biased actions and results when the training data is not biased [11]. A fair algorithm is an algorithm that does not discriminates based on sensitive attributes. When an algorithm is biased, the algorithm will make unfair predictions. This unfair algorithm makes calculations based on sensitive characteristics and historical patterns that contains bias. As a result, some specific demographic groups can be disadvantaged when these unfair algorithms make important predictions about people. Predictions about the ability to pay back loans or whether someone is a criminal, can have huge impacts on people's lives and especially when these predictions are false [10]. Thus, different types of biases can lead to discriminatory algorithms [12].

2.1.1 Biases

There are different kinds of biases that could lead to bias in algorithms. The most important ones that can cause discriminatory outcomes are listed as follows.

- **Omitted Variable Bias**. This bias arises when a certain feature is not taken into account when building the model, even though the feature affects the target variable Y [6].
- **Measurement Bias**. This bias appears when people measure features in an inaccurate way, which causes inaccurate data. [11].
- **Representation Bias**. Representation bias occurs when the sampling of a population during the data collection process is not representative enough [2].
- **Historical Bias**. Historical bias is the already existing bias that can slip into algorithms even if the data is perfectly sampled and measured [15].

2.1.2 Definitions of fairness

As the concept of algorithmic fairness emerged, many researchers started discussing what definition of fairness they should apply for their algorithms [16]. Scientists from AI, software engineering and law communities interfered in this ongoing debate about what notion should be used for decision-making models for justice systems. The definitions in fairness differ on particular aspects. For example, some define fairness as treating every individual the same, others define it as treating different demographically groups the same [11].

N.A. Saxena, et al. [13] showed that the preference of people for a specific definition is dependent on the information of the case. Scientists presented three notions of fairness to participants and asked them which one is the fairest in the context on loan allocations. Participants began with

no knowledge about loan recipients, and then gained increasing knowledge about the gender and and race of the loan recipients. The definition of fairness that was most preferred was 'Calibrated fairness'. Calibrated fairness is based on the proportions of quality of individuals. This means that in this study, two persons with repayment rate r_1 and r_2 respectively will receive $r_1/(r_1 + r_2)$ and $r_2/(r_1 + r_2)$. Results also showed support for the principle of affirmative action, because participants chose relatively in favour of women and African American people when distributing loans. This could be because these people are part of historically disadvantaged groups.

Researchers have studied a lot of definitions of fairness and discussed them extensively. Several experiments have been done to compare each definition with each other, with the final goal to find out which definition is the best one. However, all studies have shown that this goal is almost impossible to reach. Each situation is different, because it contains different demographic groups and has different purposes. Besides, some of these notions exclude each other. Adhering to a certain definition would prevent the ability to adhere to others. Moreover, everyone has his own opinion on what fairness is and how it should be practiced. Nevertheless, the scientists ensure that more attention is paid to this problem. As a result, people will think more about the serious influence these algorithms have on minority groups and how the harming of this groups by algorithms can be tackled [11, 16, 12, 7].

2.1.3 Statistical measures

As mentioned earlier, a lot of definitions have been compared to each other. Each definition measures fairness in its own way, but the basis of this is measuring the correct or incorrect predictions of an algorithm. This is mostly done through a confusion matrix, as shown in figure 1.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Figure 1: A confusion matrix [16].

The notions from figure 1 that will be used in this thesis are summarized as follows [16].

- **True positive (TP)**. This is a true value which was predicted to be true.
- **False positive (FP)**. This is a false value which was predicted to be true, also known as a Type I Error.
- **True negative (TN)**. This is a false value which was predicted to be false.

- **False negative (FN)**. This is true value which was predicted to be false, also known as a Type II error.
- **True positive rate (TPR)**. The part of positive values that is correctly predicted out of all real positive values.
- **False positive rate (FPR)**. The part of negative values that is wrongly predicted out of all real negative values.

For the experiments in this thesis, only the notions TPR and FPR are used. The TPR can be stated as follows.

$$P[\hat{Y} = 1|Y = 1] \tag{1}$$

Here, P stands for the probability. $\hat{Y} = 1$ means here that the prediction is positive and $\hat{Y} = 0$ means that the prediction is negative. $Y = 1$ stands for a true positive value and $Y = 0$ stands for a true negative value. The FPR is formulated as follows.

$$P[\hat{Y} = 1|Y = 0] \tag{2}$$

Many of the definitions of fairness are formulated on the basis of the notions in truth table 1. For instance, some people think fairness is achieved when the PPV of the algorithm is the same for two different groups. An example of another definition is "Equalized odds". This definition states that fairness between two groups can be measured by calculating the difference between the FPR and the difference between the TPR of the two groups [12]. This notion is formally computed as follows:

$$|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S \neq 1, Y = 0]| \leq \varepsilon, \tag{3}$$

$$|P[\hat{Y} = 1|S = 1, Y = 1] - P[\hat{Y} = 1|S \neq 1, Y = 1]| \leq \varepsilon, \tag{4}$$

The formula above specifies that the absolute difference in FPR of two groups should be lower than ε, and the absolute difference in TPR of two groups should be lower than ε. $S = 1$ represents the privileged group and $S \neq 1$ represents the unprivileged group.

As explored in the following section, the two definitions 'False positive' and 'False negative' are crucial for criminal justice. In this field, a false positive means a person that is accused of being a criminal when he or she is not. A false negative here means a person who has been found innocent when he or she is actually not. With other words, a guilty person is acquitted.

2.2 Criminal justice policy preferences

As explained in the previous section, the choice on which definition of fairness to use for decision-making algorithms is very difficult. Computer scientists have to understand the meaning of each definition and compare them with each other, which is complicated. Something that computer scientists also cannot agree on, is the size of the consequences of the mistakes that their algorithms make [7, 14]. Every mistake has some sort of costs, but mistakes will also have personal and moral costs within the domain of criminal justice. Algorithms that make predictions about people regarding criminal justice, cannot predict people perfectly. Sometimes, the algorithm predicts

someone to be innocent when he or she is not and vice versa. The developer of the algorithm should balance the harm of mistakenly predicting an innocent person to be guilty, against the harm of predicting a guilty person to be innocent. The costs of these two mistakes could differ from each other. In some cases it is more costly to falsely accuse someone to be a criminal, but in other cases it is more costly to predict that someone is innocent when he or she is not. For example, when an algorithm has to predict possible terrorists, false negatives could have deadly consequences. Developers of these algorithms struggle with the problem that decreasing the amount of false positives indirectly leads to an increase in false negatives.

The fairness of these algorithms can be measured with the TPR and FPR. As explained earlier, the TPR is calculated by dividing all positive values that are correctly predicted by all real positive values [16]. So in this case it means that all real criminals that are identified correctly, will be divided by all real criminals. In other words, the TPR means the probability that a criminal will correctly be identified. Thus, a high TPR means here that the algorithm can predict criminals well. The FPR is calculated by dividing all negative values that were wrongfully predicted by all real negative values. So here, it means that all real innocent people that are wrongfully accused will be divided by all real innocent people. In other words, the FPR means the probability that an innocent person will be wrongly accused of being a criminal. Therefore, a high FPR means that the algorithm does not perform well, because a lot of innocent people are accused wrongfully.

N. Scurich, et al. [14] further considers a debate about these criminal justice policy preferences and how the "veil of ignorance" could affect this. To determine a acceptable and wished balance between false positives and false negatives, many people proposed a trade-off. The eighteenth-century jurist William Blackstone said that "[B]etter that ten guilty persons escape, than that one innocent suffer."(p. 2). Even though other jurists changed these numbers in the years that followed, the ratio of Blackstone remained the most populair. In the 20th century, philosopher John Rawls came up with the idea of the veil of ignorance. The veil of ignorance reduces judgments by letting people evaluate a policy without knowing how this would affect him or her personally. The study discusses an experiment where participants were first asked which mistake they think is more worse and to what extent. After that, they were asked whether they would rather be violently assaulted by someone or be in locked up in jail for one day. They were also asked how this choice would compare to the other option. Results show that a nontrivial number of participants were not consistent in their preferences. For example, a part of the participants thinks false positives are more worse than false negatives, but rather suffer the consequences of a false positive than a false negative.

M. Xiong, et al. [18] discussed several experiments across different countries on the attitudes of citizens towards judicial errors. The term 'judicial errors' has been used in their study as a collective term for wrongful convictions and erroneous acquittals. People were asked in surveys which of these two mistakes they think is worse. The many surveys that have been done were different, so it was admitted that comparing them is questionable. For example, some surveys included an option for when the participant could not choose between the two mistakes. When these answers are included in the results, the outcomes are very different from surveys were this answer was not an option. However, the results of surveys from 23 countries showed that the majority think that convicting innocents is worse than releasing criminals. The survey experiments have been done from 1985 until 2006. The outcomes of these surveys also showed that the preference for

this type of mistake has generally become less popular over the last decades. One of the reasons could be because of the increase in terrorist attacks in recent decades.

3 An Empirical Illustration: COMPAS algorithm

For the empirical research of my thesis, the case-study of COMPAS will be the main point of focus. The reason that this software tool is chosen, is because it is one of the best known examples of a discriminatory algorithm. The problem of decision-making models in justice systems can be illustrated very clearly with the analysis of this algorithm. First, an overview will be given of what COMPAS exactly is. After that, an explanation will be given on what the analysis is about and why this experiment is examined.

3.1 COMPAS overview

A well-known example of a biased AI-based prediction algorithm among computer scientists is COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) [11]. In the U.S., this tool is often used by courts. This software predicts the chance that a certain defendant will commit a crime again, also known as recidivism, and gives them a risk score for this. The tool is mostly used to help judges make decisions about jail terms, sentencing, probation, and pretrial release and parole. After a person gets arrested, a survey of 137 questions about the arrest has to be completed. A police officer answers the first 30 questions and the defendant answers the remaining. Based on the answers, the defendant receives a score from one to ten for three different categories. These categories are "Risk of Recidivism", "Risk of Violent Recidivism" and "Risk of Failure to Appear". The first category indicates the chance of committing a nonviolent crime in the future. The second one represents the likelihood of committing a violent crime in the future. The last one indicates the probability that the defendant will not show up to trial [3].

In 2014, U.S. Attorney General Eric Holder warned that the risk score might cause bias in trials [1]. He asked the U.S. Sentencing Commission to study the use of these risk scores. However, this study was never launched by the U.S. Sentencing Commission. That is why ProPublica (an independent, non-profit newsroom that produces investigative journalism in the public interest) started an investigation to this. In 2016, the researchers of ProPublica reported that the AI-based algorithm COMPAS is biased against African Americans [7]. Researchers said that African American people who did not re-offend were twice as likely to have a high risk score than white people [2]. This disadvantage against African American people is a very serious problem, because the outcomes of the COMPAS software could have a huge impact on people's lives. A lot of innocent people were accused of being a high potential risk to become a criminal again, also known as a recidivist. However, Northpointe (the developer and owner of COMPAS) rejected this accusation and said that ProPublica was focused on the wrong way of measuring the software's outcomes. This discussion has arisen because there are different definitions of algorithmic fairness. Many papers have been written about it and computer scientist still could not agree on this [16, 12].

3.2 COMPAS analysis

ProPublica also shared their reports on the research into COMPAS. After this, experts of different specializations, such as law, computer science and statistics, tried to replicate their methods and results [17]. They found out that although the precision for predicting recidivism is almost the same for African Americans and white people, the algorithm does not meet the conditions of the equalized odds definition [12]. This is because American-Americans are twice as much predicted to be a recidivist than white people. In other words, the FPR of African Americans is twice as higher than the FPR of white people. Furthermore, the FNR is relatively lower for white people than for African Americans. This means that COMPAS correctly accuses few white recidivists compared to African American recidivists.

The construction of the COMPAS algorithm was of interest to this study, so that the algorithm could be recreated and experiments could be done with it. However, Northpointe refused to reveal the source code of the software [3]. Northpointe has the legal rights to keep this secret. Therefore, nobody knows exactly how the algorithm was programmed and what data was used for this. Since there is nothing known about how the COMPAS algorithm was built, a replication of this model was not possible for the case-study. Therefore, a linear algorithm and a non-linear algorithm is programmed. For each algorithm type, two different thresholds were implemented. These different properties of the algorithm show how these implementation choices influence the outcomes of the algorithm. These algorithms will be trained with a COMPAS dataset. In the next chapter, the methodology of this experiments are explained further on.

4 Methodology

In this thesis, we will analyze how and for which demographic groups the unfairness in AI-based algorithms used in decision-making processes in justice systems may emerge. This will be done by a four-algorithmic experiment that contain two different variables. These variables are the type of model and the threshold. By making four algorithms, the influences of each programming choice can be studied and compared to each other. After this, a last experiment is implemented to zoom in on the threshold influence.

4.1 Data

The data that is used for the experiments comes from a COMPAS dataset that contains information about 6172 defendants. The information consists of sensitive characteristics, behavioral characteristics, whether the person is a recidivist and if he or she is accused of it. Each row stands for a defendant and each column stands for a characteristic. The data is mostly binary, so a 1 means a true value and a 0 means a untrue value. The features are described in table 1.

Feature	Explanation
Two_yr_Recidivism	Specifies whether the defendant has actually become a recidivist after two years.
Number_of_Priors	This is the number of previously committed crimes. This value is not binary.
Score_factor	Specifies whether the COMPAS algorithm accused the defendant of being a recidivist.
Age_Above_FourtyFive	Specifies whether the defendant is older than 45 years.
Age_Below_TwentyFive	Specifies whether the defendant is younger than 25 years.
African_American	Specifies whether the defendant is an African American person.
Asian	Specifies whether the defendant is an Asian person.
Hispanic	Specifies whether the defendant is a Hispanic person.
Native_American	Specifies whether the defendant is a Native American person.
Other	Specifies whether the race of the defendant was not one of the foregoing, making him or her a white person.
Female	Specifies the sex of the defendant. A 1 means that the person is female and a 0 means the person is male.
Misdemeanor	Specifies the seriousness of the crime that the defendant committed. A 1 means that it is a misdemeanor, such as basic assault or possession or drugs. A 0 means that it is a felony, such as murder or rape.

Table 1: The features of the dataset that is used for the experiments.

Out of these characteristics, several demographic groups can be made. These groups are used for the experiments, but this will be more explained later on. The proportions of the groups compared to the total dataset are described in table 2.

Group	People	Proportion
African Americans	3175	0.514
Native Americans	11	0.002
Whites	2446	0.396
Young people	1347	0.218
Middle aged people	3532	0.572
Old people	1293	0.209
Females	1175	0.190
Males	4997	0.810
Young African American females	145	0.023
Young African American males	664	0.108
Middle aged African American females	335	0.054
Middle aged African American males	1563	0.253
Old African American females	69	0.011
Old African Americans males	399	0.064
Young white females	87	0.014
Young white males	334	0.054
Middle aged white females	300	0.049
Middle aged white males	1022	0.166
Old white females	153	0.025
Old white males	550	0.089

Table 2: The distribution of the used data for the experiments.

4.2 Linear and non-linear method

Within algorithm models, there is a difference between linear an non-linear algorithm. A linear algorithm assumes that there is a linear relationship between x and y. The function that describes this relation would be as follows: $y(x) = ax + b$. A non-linear algorithm assumes that this relationship is not linear, but that the rate of change is different for each x [8]. The function for this could be very complex. The linear algorithm that is used for the experiments in this thesis is logistic regression. The non-linear algorithm that is used for the experiments is random forest classifier.

K. Kirasich, et al. [9] did research on the difference between logistic regression and random forest with a binary classification problem. They discovered that when the noise variables were increased, the TPR and FPR of random forest was higher than logistic regression. The FPR for random forest was found statistically different than logistic regression. The researchers also found out that the accuracy of logistic regression was higher than random forest, when increasing the variance in the explanatory and noise variables. The experiments of this thesis will not, however, use any explanatory and noise variables. Nevertheless, some people are convinced that the performance of random forest is more similar to human decision-making than the performance of logistic regression. [8].

4.2.1 Logistic regression

The first two algorithms that are programmed for the experiment contain a logistic regression method. Regression analysis is a method that can model the relationships between variables. It can predict a variable based on one or more other variables. Between regression methods, there is a difference between linear regression and logistic regression. Linear regression is used in problems where the predictive variable is a metric variable. Logistic regression is useful when the predictive variable is a binary variable, which means that it has two categories. For this experiment, the predictive variable is a binary variable. That is why we use logistic regression for this experiment. The logistic regression function is stated in 2.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

Figure 2: The logistic regression function [8].

Logistic regression models the probability that Y belongs to a certain category. The probability is a value that always stays between the range of [0,1]. That is why the function and the graph also has to stay between 0 and 1. The regression coefficients β_0 and β_1 will be calculated based the training data. To estimate the unknown coefficients, the method 'maximum likelihood' is mostly used. Here, we search for estimates for β_0 and β_1 so that the predicted probability of each individual reaches the real status as closely as possible [8].

4.2.2 Random forest classifier

For the second two algorithms that are implemented for the experiment, a random forest classifier is used. Random forest analysis is a method that combines the outputs of decision trees to determine the prediction. The multiple decision trees that are made with this method are trained by a random subset of the training data. In figure 3, an example of an ensemble of decision trees is illustrated. These decision trees start with a single point, which repeatedly splits into two ways.

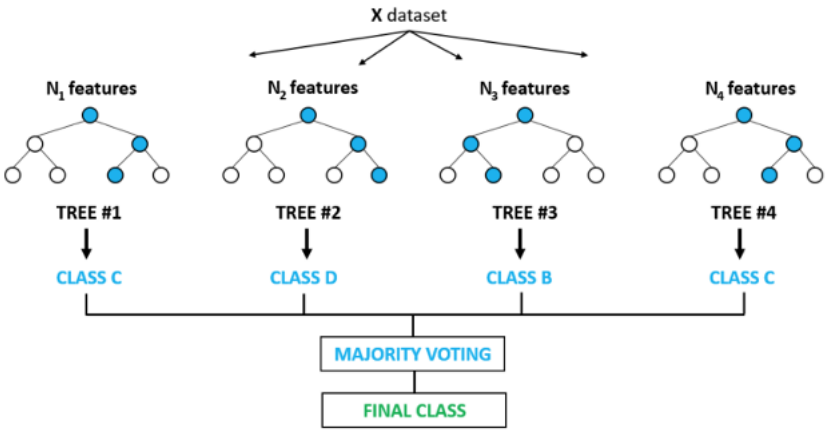


Figure 3: An example of an ensemble of decision trees [9].

Depending on whether the individual meets the condition of each point, it loops over a certain path in the tree. These conditions are based on the features that are used to train the model. When the individual ends in the last point, a certain value or class is appointed to this individual. For regression problems, the mean of all outcomes of the decision trees is used to make a prediction. For classification problems, the majority class of all decision trees is used to make a prediction. The target of our experiment consists of two classes, which is why random forest classification is used for this experiment [8].

4.3 Algorithm

For the four-algorithmic experiment, two algorithms are built with the logistic regression model and two are built with random classifier. For each algorithm method, the difference of the two algorithms is the threshold of the prediction model, but that will be explained later on. Each of the four algorithms is built up as follows. First, the prediction model is implemented for all defendants. This group is called 'All'. The target for this model is 'Two year Recidivism', because we want to predict if the defendant has become a recidivist or not. If this is not the case, this person is called an innocent person. The features that are used for this model are whether the committed crime is a misdemeanor and the number of previously committed crimes. We only use these information about the defendants, because in this way the prediction is not based on any sensitive characteristics.

After the features and target are selected, the data is split in a training set and a training set. The test size parameter is 0.1, so the test set consists of 10% of the whole dataset. The random state parameter is set on 400, so the data is shuffled 400 times before splitting it. After this, the prediction model is made. The specifications of these are explained later on. The model is trained with the training data. Then, the model made predictions on the defendants, which are compared to the real values. From this comparison, a confusion matrix is made.

After we made the prediction model for all defendants, we separated the data in groups based on the sensitive characteristics. For race, three groups are made; African Americans, Native Americans, and Whites. The column 'other' is used to represent white people. For age, three groups are made. The names of these groups are Young people, Middle aged people, and Old people. The first group is younger than 25 years old, the second group is between 25-45 years old, and the last group is older than 45 years old. Two groups are made for gender, which are Females and Males. This results in 8 'main groups'. Finally, specific demographic groups are made by combining the three characteristics gender, age and race. The concept of intersectionality is the reason for this. Intersectionality means that every specific demographic group experience a different kind of discrimination. That is why 'intersected groups' like Young African American females or Old White males are made. The main groups and intersected groups make up 20 groups. For each of these groups, a prediction model is implemented. When this groups are used in prediction model, the differences between these groups could be measured and visualised. For race, only the African Americans and Whites are used. This is because the group of Native Americans was too small in the original dataset, so that intersected groups could not be made.

4.3.1 Threshold

For each of these two model methods, two algorithms are made with a different threshold. The threshold is a value that is used to determine if the defendant will be a recidivist or not. When the model makes a prediction of each defendant, it calculates the probability that he or she will be a recidivist. The model will predict that this individual will be a recidivist, when the calculated value exceeds the threshold. Conversely, if the value falls below the threshold, the model will predict that the individual will not be a recidivist. The first two algorithms use the default threshold for the prediction models that are used, which is 0.50. For the last two algorithms, the threshold 0.70 is chosen. This choice was made so that the difference in threshold values is neither too large nor too small. A higher threshold results in less accusations of recidivism. The probability that the defendant is a recidivist has to be higher to accuse him or her of being one. For the last two algorithms, it is detrimental for the defendants with the probability between 0.50 and 0.70. They will be accused of being a recidivist here, but not by the first two algorithms.

4.4 Evaluation

To evaluate the four-algorithmic experiment that is done, a choice had to be made how to do this. There are a lot of ways to measure prediction algorithms and the confusion matrices that came out of these. For this research, a choice is made to measure the TPR and FPR of the confusion matrices that came out the prediction models. As explained earlier, the TPR is calculated by dividing all positive values that are correctly predicted by all real positive values. So in this case it means that all real recidivists that are identified correctly, will be divided by all real recidivists. In other words, the TPR means the probability that a recidivist will correctly be identified. When a confusion matrix is measured, the TPR is the same as the recall. Because of this, a simple function was used to get the recall. The FPR is calculated by dividing all negative values that were wrongfully predicted by all real negative values. So here, it means that all real innocent people that are wrongfully accused will be divided by all real innocent people. In other words, the FPR means the probability that an innocent person will be wrongly accused of being a recidivist. To get the FPR, the top box on the right of the confusion matrix is selected first. Then, the value in this box is divided by the same value plus the value in the top left box. Finally, graphs are made for the algorithm experiments, where the points of the groups are plotted. Here, the FPR is the x-coordinate and the TPR is the y-coordinate for each point. Two separate graphs are created for each of the four algorithms, one for the main groups and one for the intersected groups. In both groups, the point of group All is shown. As a result, the different groups could be compared to the All group easily.

4.5 Zoom in experiment

After the four-algorithmic experiment is done, a last algorithm is made to focus more on the influence of the threshold. Apart from the two different threshold experiments, many insights can be gained from this final experiment. By investigating the algorithms with a broad range of thresholds, the impact of the choice of the threshold can be shown. For this algorithm, three intersected groups are chosen to demonstrate the influence of different thresholds clearly. These three intersected groups are Young African American males, Middle aged white males, and Old white females. These groups are used again to train a random forest classifier, with the same parameters as the previous

one. However, this is done multiple times now for each group, with a different threshold each time. The thresholds that are used to do this range from 0.10 to 0.95 with a step size of 0.05. After this, the TPR and FPR are calculated for each threshold. From this, two graphs are made. One graph shows the influence of the threshold on the TPR and the other shows the influence of the threshold on the FPR.

5 Results

In this section, the results of the experiments will be shown. First, the results experiments with threshold 0.50 are shown, then the results with threshold 0.70. Per experiment, the tables are shown where the two types of methods are compared. After this, these results are illustrated in graphs. The distances from each point to the base are shown in table 7 in the Appendix. After the results of the four-algorithmic experiment are shown, the results of the the zoom-in experiment is illustrated.

5.1 Four-algorithmic experiment

In the tables, the results are shown for the main and intersected groups in separate tables. The logistic regression(l) results and the random forest(r) results are compared with each other here. The TPR and FPR of both methods are shown and after that, they are compared to each other by calculating the difference between the two values. Additionally, the region where the group's point is plotted on the graph is indicated. These regions are northeast(NE), southeast(SE), southwest(SW), and northwest(NW). The regions are based on the base point, which is the point of the 'All' group. If the region of the point of a group changed when the method changed from logistic regression to random forest, the cell of this region is marked gray.

5.1.1 Experiments with threshold 0.50

The results of the experiments with threshold 0.50 are collected and shown in tables3 and4.

Group	FPR(l)	TPR(l)	Region(l)	FPR(r)	TPR(r)	Region(r)	ΔFPR	ΔTPR
All	0.15	0.40	n/a	0.24	0.53	n/a	0.09	0.13
African Americans	0.24	0.57	NE	0.30	0.68	NE	0.06	0.11
Native Americans	1.00	0.00	SE	1.00	0.00	SE	0.00	0.00
Whites	0.07	0.27	SW	0.13	0.37	SW	0.06	0.10
Young people	0.39	0.60	NE	0.31	0.53	NE	-0.08	-0.07
Middle aged people	0.21	0.40	NE	0.25	0.51	SE	0.04	0.11
Old people	0.11	0.39	SW	0.18	0.61	NW	0.07	0.22
Females	0.08	0.42	NW	0.18	0.45	SW	0.10	0.03
Males	0.18	0.43	NE	0.37	0.59	NE	0.19	0.16

Table 3: The results of the experiments for the main groups with threshold 0.50.

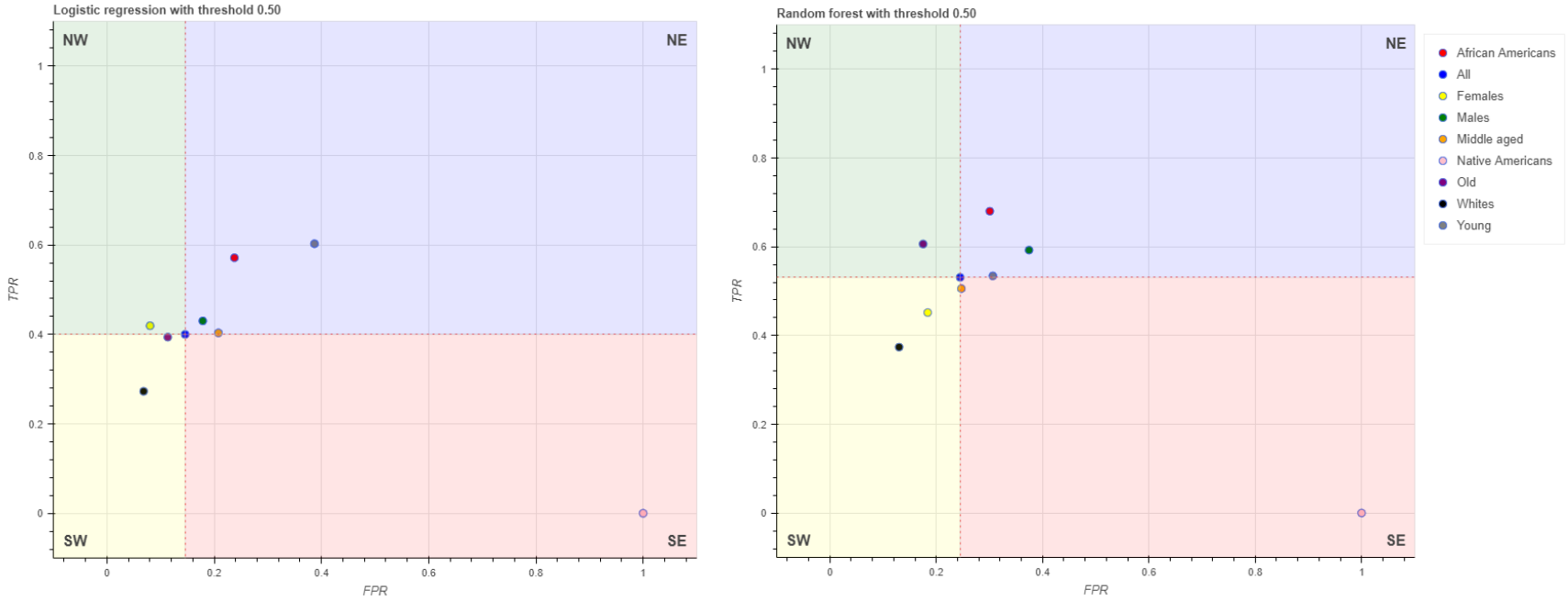
What strikes you at first when looking at table 3, is that the TPR and FPR of almost every group are higher with random forest than with logistic regression. Only the TPR and FPR of Young

people are lower with random forest, meaning Young People benefit from random forest, unlike the rest. The groups with the biggest difference in TPR and FPR are Old people and Males. This implicates that these two groups are relatively dependent on the algorithm method. When we look at the regions, we see that most groups are in SW or NE. The groups African Americans, Males, and Young people are in NE mostly. The group Whites is far to the bottom left compared to the base. This is also notable in the graphs in 4. Therefore, it can be said that the results are detrimental for African Americans, Males, and Young people, because they have a high TPR and FPR. The results are favourable for the group Whites, because it has a low TPR and FPR.

Group	FPR(l)	TPR(l)	Region(l)	FPR(r)	TPR(r)	Region(r)	ΔFPR	ΔTPR
All	0.15	0.40	n/a	0.24	0.53	n/a	0.09	0.13
Young African American females	0.26	0.64	SE	0.29	0.13	SE	0.03	-0.51
Young African American males	0.67	0.67	NE	0.67	0.67	NE	0.00	0.00
Middle aged African American females	0.08	0.33	SW	0.36	0.33	SE	0.28	0.00
Middle aged African American males	0.26	0.64	NE	0.38	0.73	NE	0.12	0.09
Old African American females	0.00	0.00	SW	0.00	0.00	SW	0.00	0.00
Old African American males	0.28	0.40	NE/SE	0.32	0.53	NE	0.04	0.13
Young white females	0.14	0.00	SW	0.29	0.00	SE	0.15	0.00
Young white males	0.55	0.60	NE	0.55	0.65	NE	0.00	0.05
Middle aged white females	0.07	0.33	SW	0.13	0.40	SW	0.06	0.07
Middle aged white males	0.10	0.40	SW	0.13	0.49	SW	0.03	0.09
Old white females	0.00	0.50	NW	0.00	1.00	NW	0.00	0.50
Old white males	0.02	0.33	SW	0.05	0.33	SW	0.03	0.00

Table 4: The results of the experiments for the intersected groups with threshold 0.50.

In table 4, we can see that the results for the intersected groups are almost the same as for the main groups. Only now, the points are more spread. This can also be seen in the graphs in 5. Groups with the characteristics African American, young, and male are mostly in the upper right corner. This means that the results are unfavourable for these groups. Groups with characteristics such as being white, old, and female are most often located on the left side or bottom of the graphs. These groups are advantaged in this experiment. When we compare the two methods in table 4, we notice a few things. First, the difference in TPR of the groups Young African American females and Old white females is very big. For Young African American females, the TPR is a lot higher with logistic regression. For Old white females, however, this value is a lot higher with random forest. Besides this, the TPR of Young African American females is the only value that decreased when the method changed to random forest. There are three groups that moved to another region



(a) Logistic regression with threshold 0.50.

(b) Random forest with threshold 0.50.

Figure 4: The graphs shows the differences between the two methods for the main groups with threshold 0.50.

when the method changed. These groups are Middle aged African American females, Old African American males, and Young white females. From this, it can be said that the method has a lot of influence on the outcomes of several groups.

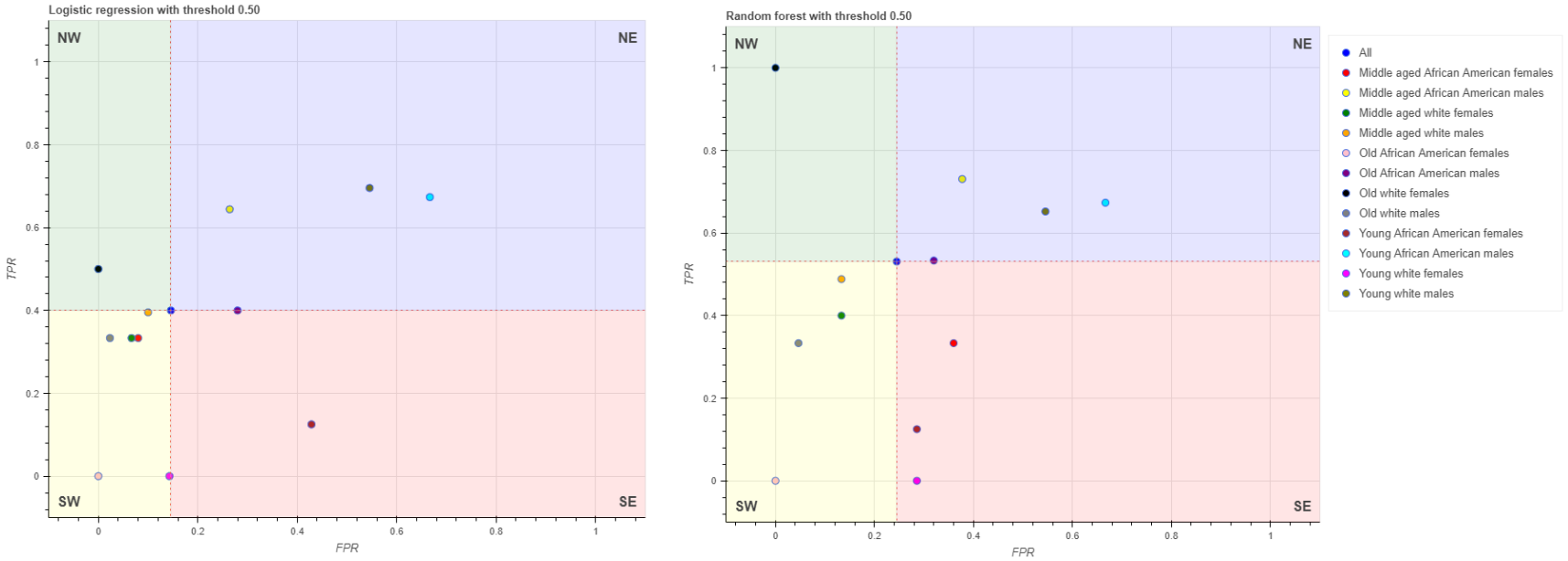
5.1.2 Experiments with threshold 0.70

Now, the threshold of the algorithms are reset. The threshold is increased to 0.70. Consequently, defendants previously identified as innocent in earlier experiments may now be identified as recidivists. The results of the experiments with threshold 0.70 have been put together and are shown in tables 5 and 6.

Group	FPR(l)	TPR(l)	Region(l)	FPR(r)	TPR(r)	Region(r)	ΔFPR	ΔTPR
All	0.03	0.13	n/a	0.05	0.17	n/a	0.02	0.04
African Americans	0.05	0.17	NE	0.08	0.23	NE	0.03	0.06
Native Americans	1.00	0.00	SE	0.00	0.00	SW	-1.00	0.00
Whites	0.04	0.12	SE	0.03	0.12	SW	-0.01	0.00
Young people	0.10	0.21	NE	0.19	0.33	NE	0.09	0.12
Middle aged people	0.06	0.19	NE	0.08	0.20	NE	0.02	0.01
Old people	0.05	0.24	NE	0.05	0.24	NE	0.00	0.00
Females	0.03	0.12	NW	0.02	0.16	SW	-0.01	0.04
Males	0.03	0.16	NW	0.04	0.20	NW	0.01	0.04

Table 5: The results of the experiments for the main groups with threshold 0.70.

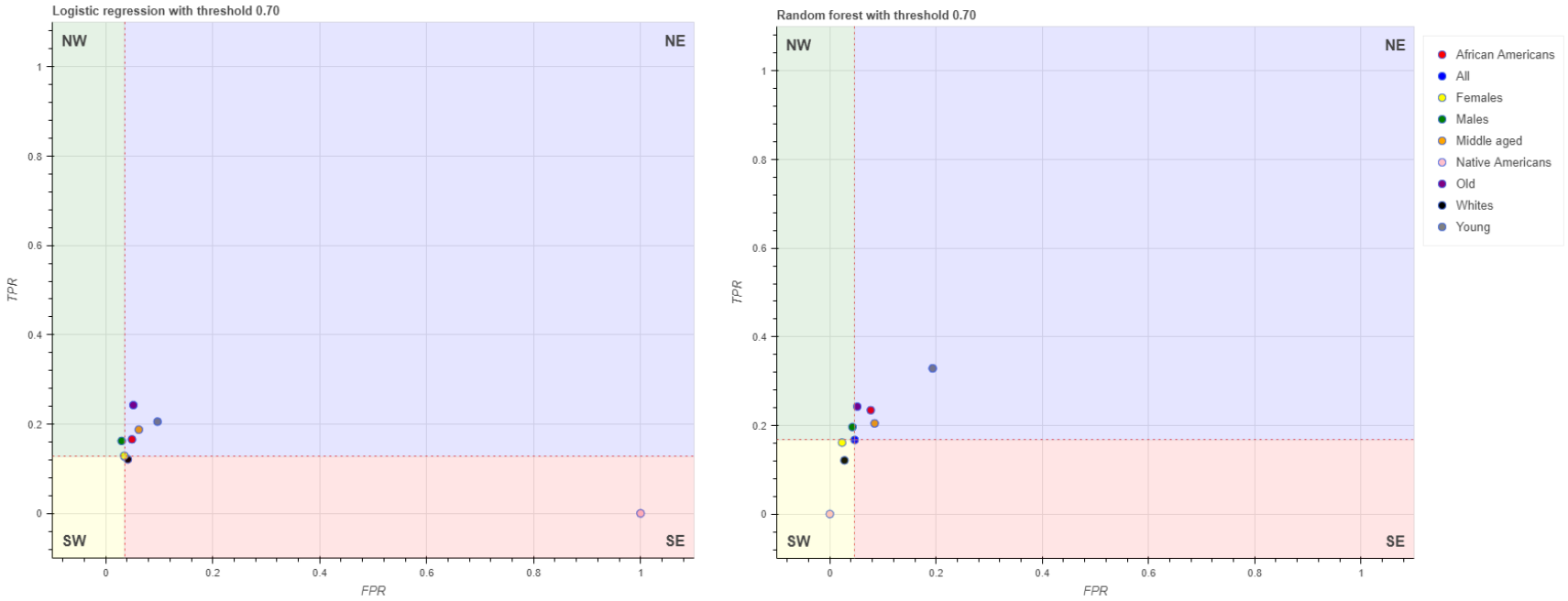
When we look at table 5, we see that the points of the two methods do not differ a lot from each other. Besides the point of Native Americans, which is a outlier every time, the point of Young people has also changed relatively much. This can also be seen in the graphs in 6. The point of Young people is in 6b a lot more to the upper corner right than in 6a. Because of this, we can say that the method does not have much influence on the outcomes when the threshold is high, except for the group Young people. However, there are three main groups that changed region when the random forest was used instead of logistic regression. These groups are Native Americans, White, and Females. In addition, we see that the points of the groups are positioned relatively close to each other, compared to the experiments with threshold 0.50. This implicates that the groups are treated more equally than with threshold 0.50. With the last of the four algorithms, we observed this as well.



(a) Logistic regression with threshold 0.50.

(b) Random forest with threshold 0.50.

Figure 5: The graphs shows the differences between the two methods for the intersected groups with threshold 0.50.



(a) Logistic regression with threshold 0.70.

(b) Random forest with threshold 0.70.

Figure 6: The graphs shows the differences between the two methods for the main groups with threshold 0.70.

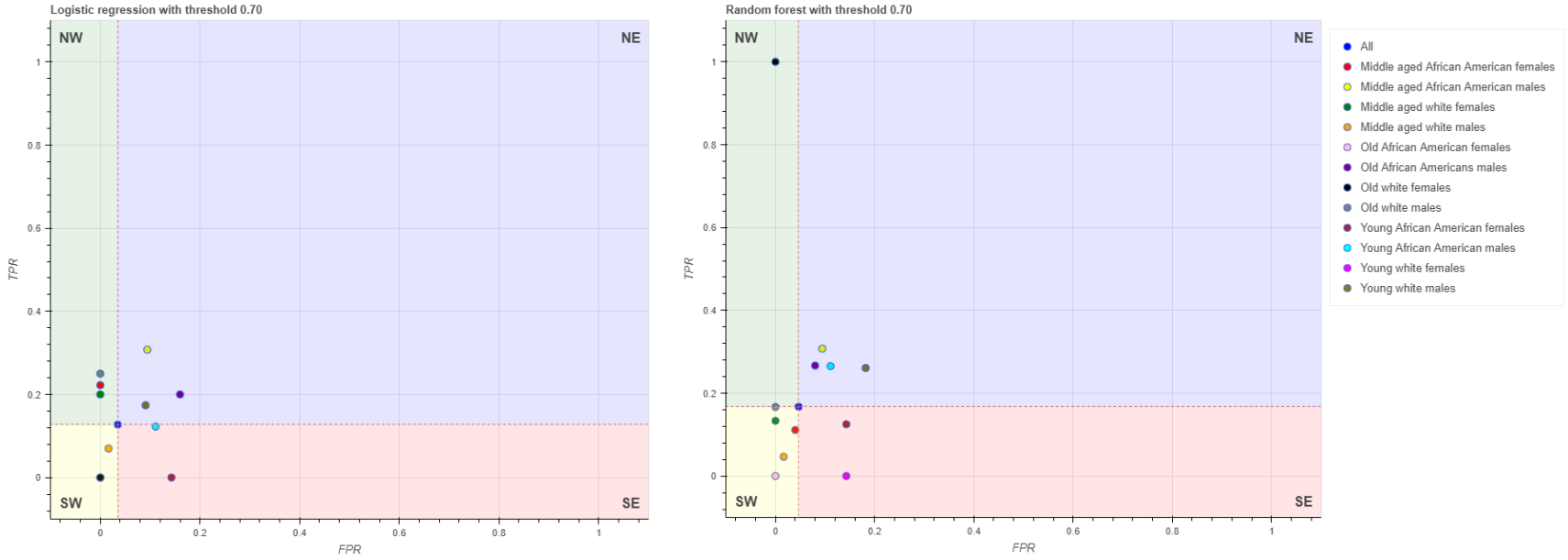
Group	FPR(l)	TPR(l)	Region(l)	FPR(r)	TPR(r)	Region(r)	ΔFPR	ΔTPR
All	0.03	0.13	n/a	0.05	0.17	n/a	0.02	0.04
Young African American females	0.14	0.00	SE	0.14	0.13	SE	0.00	0.13
Young African American males	0.11	0.12	SE	0.11	0.27	NE	0.00	0.15
Middle aged African American females	0.00	0.22	NW	0.04	0.11	SW	0.04	-0.11
Middle aged African American males	0.09	0.31	NE	0.09	0.31	NE	0.00	0.00
Old African American females	0.00	0.00	SW	0.00	0.00	SW	0.00	0.00
Old African American males	0.16	0.20	NE	0.08	0.27	NE	-0.08	0.07
Young white females	0.00	0.00	SW	0.14	0.00	SE	0.14	0.00
Young white males	0.09	0.17	NE	0.18	0.26	NE	0.09	0.09
Middle aged white females	0.00	0.20	NW	0.00	0.13	SW	0.00	-0.07
Middle aged white males	0.02	0.07	SW	0.02	0.05	SW	0.00	-0.02
Old white females	0.00	0.00	SW	0.00	1.00	NW	0.00	1.00
Old white males	0.00	0.25	NW	0.00	0.17	SW	0.00	-0.08

Table 6: The results of the experiments for the intersected groups with threshold 0.70.

In table 6, the last results of this experiment are shown. Here, we also see that the points of the intersected groups are more spread than the points of the main groups. Additionally, the groups with characteristics as African American, young, and male are primarily located in the NE region, similar to the experiment with a threshold of 0.50. Again, these groups are disadvantaged in the experiment. On the other hand, groups with characteristics as white, old, and female are mostly on the left side or bottom. These groups have favourable results in this experiment. We can see this in the graphs of 7 too. When we compare the methods in 6, we see that six groups have moved to another region when the method changed. These groups are Young African American males, Middle aged African American females, Young white females, Middle aged white females, Old white females, and Old white males. Therefore, it can be said that the choice of the method has a lot of impact on the outcomes of this algorithm.

5.2 Zoom in experiment

For the zoom in experiment, three groups are chosen to use for the algorithm. These groups are chosen, because they ended in different areas in the previous experiments. In three of the four experiments, the group Young African American males ended in region NE, Middle aged white males in SW, and Old white females in NW. This means that the groups are treated differently by the algorithm. With this in mind, interesting insights can be obtained by observing



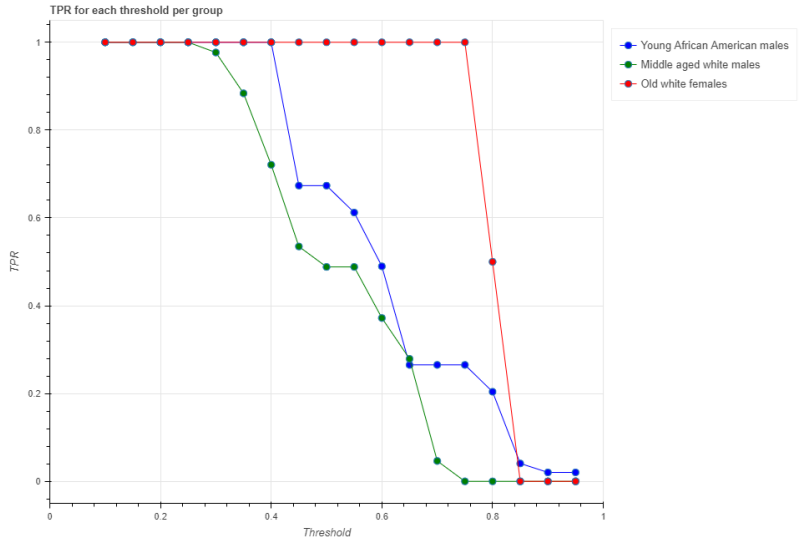
(a) Logistic regression with threshold 0.70.

(b) Random forest with threshold 0.70.

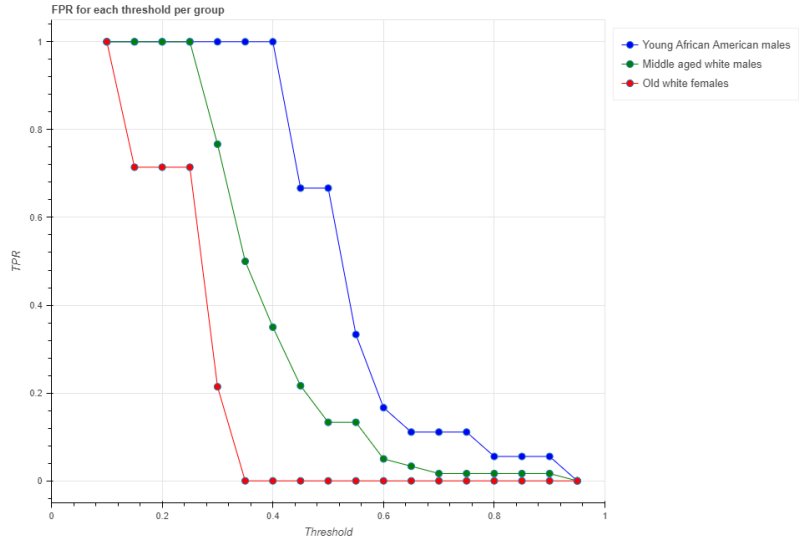
Figure 7: The graphs shows the differences between the two methods for the intersected groups with threshold 0.70.

these three different groups in this experiment. Graph 8a shows the TPR of the three groups for each different threshold. As you can see, the TPR stays on 1.00 until threshold 0.25. Then, the TPR of the group Middle aged white males goes down. The TPR of the group Young African American males goes down after threshold 0.40. The TPR of the group Old white females goes down quite late, compared to the other groups. There is a deep downward trend for all three groups. The results show that recidivists in the group Old white females can be identified relatively well with high thresholds. This holds until the threshold reaches 0.80. The groups Young African American males and Middle aged white males have an almost equivalent course in the graph.

In graph 8b, the FPR is shown for each group with the different thresholds. In this graph, the group Old white females is the first one with a descending line. This already happens after threshold 0.10. This red line also reaches the zero point very quick, compared to the other groups. After threshold 0.25, the FPR of the group Middle aged white males goes down. Finally, the FPR of the group Young African American goes down. This line is descending after threshold 0.40. Again, there is a deep downward trend for all three groups. However, at the end of the blue and green lines, the line becomes less and less steep. What is also notable in this graph, is that the group Young African American males have the highest FPR (or equal to the others) for each threshold. The group Old white females have the lowest FPR for each threshold. These results are the most beneficial for the group Old white females, because a low FPR means that few people have been wrongly accused by the algorithm. A high FPR means that a lot of defendants are wrongly accused of being a recidivist, which is why the results are detrimental for Young African American males here.



(a) The graph shows the TPR for the three groups with each threshold.



(b) The graph shows the FPR for the three groups with each threshold.

Figure 8

6 Analysis

In this section, the results of the experiments are explained and further discussed. First, the meaning of the regions of the graphs will be explained. After this, every group is discussed separately. Then, the discussion of the choices when building such algorithms follows. The nature of the unfairness in algorithms within criminal justice is discussed in the last part.

6.1 Discussion per region

As mentioned earlier, we divided the area of graphs 4, 5, 6, and 7 in four regions by drawing a horizontal and vertical line through the base point. Now, we will interpret the four regions and determine what the outcome of this analysis means for the groups. We will examine this from two perspectives: on the one hand, from the viewpoint of society, and on the other hand, from the perspective of an individual within that group.

- **NE**. In the region above the base right, the FPR and TPR are higher than the base. This means that for the groups in here, chances are higher that innocent people are wrongly accused and recidivists are correctly accused. From a social perspective, the outcome of these groups is better for a safe world without criminals. This is because criminals will be correctly accused more often, so they can be arrested faster or get a higher punishment. However, it will bring some social costs with it, because innocent people are wrongly accused of being a high risk more often. From an individual perspective from the group, it is very disadvantageous to be judged in this way by the algorithm. It is unfavourable for a recidivist because the chances are higher that he or she will be accused of being such. Moreover, this outcome is unfavourable for innocent people, because the chances are higher that they will be accused of being a recidivist.

- **SE**. In the region under the base right, the FPR is higher than the base and TPR is lower than the base. This means that for the groups in here, chances are higher that innocent people are wrongly accused. The chance that a recidivist will be accused of being one is lower. From a social perspective, this outcome is unfavourable for society. Criminals will be accused correctly less often, so there will be more criminality in the world. Besides that, innocent people are being wrongly accused more often, which is something that you also do not want for a community. From an individual perspective from the group, these results have two implications. Now, chances are higher for recidivists to be accused as innocent people, which is good for a recidivist. For innocent people, it is unfavourable because they will be wrongly accused more often.

- **SW**. In the region under the base left, the FPR and TPR are lower than the base. This means that for the groups in here, the chance is lower that innocent people are wrongly accused. In addition, the chance that a recidivist will be accused of being one is lower. From a social perspective, this result is on the one hand unfavourable, because recidivists are accused of being a recidivist less often. On the other hand, innocent people are correctly judged by the algorithm more often, so less innocent people will be seen as a high risk. From an individual perspective from the group, these outcomes are very good. It is favourable for recidivists, because they will be accused as recidivists less often. Moreover, innocent people are being judged as innocent more often, which is also beneficial for individuals.

- **NW**. In the region above the base left, the FPR is lower than the base and the TPR is higher than the base. This means that for the groups in here, chances are lower that innocent people are wrongly accused. Besides this, the chance that a recidivist will be accused of being one is higher. From a social perspective, these outcomes are very favourable. Recidivists will be accused of being one more often, so they get caught faster or get a higher penalty from the judge. Moreover, innocent people are being correctly judged more often, so they will not be

seen as a high risk. From an individual group perspective, this outcome is detrimental to some and beneficial to others. It is detrimental for recidivists, because the chance is higher that they will be accused as recidivists. However, it is beneficial for innocent people, because they will be accused of being recidivists less often.

Out of this analysis, the following conclusions can be drawn. For society, the best outcome is to end in the region above the base left (NW) and the worst outcome is to be in the region under the base right (SE). The region under the base left (SW) is the best outcome for individuals in the groups. The worst outcome for individuals in the group is to end in the region above the base right (NE).

6.2 Discussion per group

Here, the results of the four algorithm experiments that are illustrated in graphs 4, 5, 6, and 7 are discussed for each group separately.

- **All**. The group that contains all the defendants, also known as the base, had the average score for each experiment. However, the base changed for each experiment. The TPR and FPR of the base were a lot higher for the experiments with threshold 0.70 than with 0.50. The random forest method also provided a small increase in these values over logistic regression.
- **African Americans**. This group ended in the NE region with every experiment. The point was often relatively far away from the base. It could be said that this group is very disadvantaged compared to the other groups.
- **Native Americans**. This group ended in the SE region three times and in the SW region one time. This point was a outlier every time, which could be because this group did not contain a lot of defendants. Because of this, the models were tested with a small amount of defendants.
- **Whites**. This group ended in the SW region three times with the experiments. The group ended in the SE region one time, but it was positioned very close to the base and the SE region then. When the threshold was 0.50, this group was one of the groups with the greatest distance from the base. From individual perspective, it could be said that this group is advantaged in the algorithm.
- **Young**. This group ended in the NE region every time, often with a big distance to the base. When the method was random forest and the threshold 0.50, the group almost ended in the SW region. It could be said that this group is disadvantaged in the outcomes of this experiment.
- **Middle aged**. This group is positioned in the NE region two times and in the SE region one time. As a result, middle aged people are slightly disadvantaged. However, the point stayed close to the base mostly.
- **Old**. The group Old ended in the NE region when the threshold was 0.70. With threshold 0.50, the group ended in NE and NW with logistic regression and random forest respectively. Therefore, this group is very dependent on the choices that are made when the algorithm is built.

- **Females**. This group is positioned in the region SW two times and in the region NW two times. When the method random forest was used, the group ended in SW. When the method logistic regression was used, the group in NW. Because of this, the group is very dependent on which method is used in the algorithm. It can also be said that Females is a little advantaged with these outcomes.

- **Males**. The group Males is positioned in the region NE two times and in the region NW two times. When the threshold was 0.50, the group ended in NE. When the threshold was 0.70, the group ended in NW. Therefore, the group Males is very dependent on which threshold is used in the algorithm.

- **Young African American females**. This group ended in the region SE four times. Often, the distance between this point and the base was very long. This means that the outcomes of this group are very detrimental from social perspective.

- **Young African American males**. This group is positioned in the region NE three times. When the threshold was 0.70 and the method was logistic regression, this group only just finished in region SE. When the group ended in NE, the distance to the base was very big. Therefore, it could be said that this group is disadvantaged in the models.

- **Middle aged African American females**. This group ended in SE one time, in SW two times, and in NW one time. By this, it is hard say something about these outcomes. However, it could be said that this group is very dependent on the choices that are made when building the algorithm.

- **Middle aged African American males**. The group Middle aged African American males is positioned in the region NE every time. The points of this group often have a big distance to the base. As a result, it could be said that this group is very disadvantaged.

- **Old African American females**. This group ended in the region SW four times. The points of this group were (0.00, 0.00) each time. It could be said that this is very advantageous from individual perspective. However, the results of this outlier could be questioned because of the small size of this data group.

- **Old African American males**. The group Old African American males ended in the region NE three times. With method logistic regression and threshold 0.50, the point was exactly on the line between NE and SE. Over all four experiments, the distance between this group and the base was average. Because of this, it could be said that this group is a little bit disadvantaged in the models.

- **Young white females**. This group is positioned in SE with logistic regression and in SW with random forest. Therefore, this group is very dependent on which method is used in the algorithm. What also strikes us when we look at the outcomes of this group, is that the TPR of this group is exactly zero for every experiment.

- **Young white males**. The group Young white males ended in the region NE with every experiment. Often, the distance between this group and the base was relatively big. By this, it could be said that this group is very disadvantaged in the outcomes of the models.

- **Middle aged white females** . This group is positioned in the region SW three times. With method logistic regression and threshold 0.70, the group ended in region NW. The distance from these points to the base often were average. Therefore, it could be said that this could is a little bit advantaged.

- **Middle aged white males**. The group Middle aged white males ended in the region SW with every experiment. The distance between the points of these groups are average most of time. As a result, it could be said that this group is advantaged in the models from individual perspective.

- **Old white females**. This group is positioned in the region NW three times. When the method was logistic regression and the threshold was 0.70, the group ended in SW. What stands out here, is that the FPR of this group is exactly zero with every experiment. Therefore, it could be said that is group is very advantaged. However, a reason for these outcomes could be because this group contained a small amount of data.

- **Old white males**. The group Old white males is positioned in the region SW three times. When the threshold was 0.70 and the method was logistic regression, the group ended in region NW. The distance from these points to the base often were average. Because of this, it could be said that this group is a little bit advantaged in the outcomes of the models.

6.3 Discussion of variables

As shown in the previous section, the groups had all different sorts of outcomes with the experiments. This can be explained by the different variables within the algorithms. For my experiments, two variables are used to make four different algorithms; The method and the threshold. The influence of the threshold is studied further with the zoom in experiment. Here, we see that the TPR and FPR react very differently on the different thresholds for each group. Because of this, the choice of which threshold is used in these algorithms matters a lot. Some groups are advantaged with a certain threshold, but can be disadvantaged when this threshold changes. Some people will be accused of being a recidivist with threshold 0.50, but are declared innocent with threshold 0.70. That is why it as such a big impact on people's lives. However, there are many more variables when it comes developing a prediction model. All these variables have a big impact on how the prediction model performs, just like the two variables in this experiment. As a result, the impact of the choices that are made when building these algorithms are huge. Especially when the algorithms are used in criminal justice, like COMPAS. Therefore, it could be questioned that software engineers should program these prediction models all by themselves. Software engineers often do not have enough knowledge of ethics and criminal law like philosophers and law scientists do. It is important that these algorithms are built with the knowledge of both ethics and criminal law.

6.4 Discussion of the situation

The implications that are discussed in the previous section, are the reason for this worrying situation. When COMPAS is used by judges in court, the predictions of this algorithm could have a lot of influence on the court decision. The risk scores of the defendant will help the judges in making decisions about the freedom of the defendant. However, these scores can differ a lot when different variables

are used, as we saw in the experiments. For some groups, it mattered a lot which method was used in the experiment. Some were advantaged with logistic regression and disadvantaged with random forest, and vice versa. In addition, the threshold has a lot of influence on the outcomes too. Some groups were advantaged with a high threshold and disadvantaged with a low threshold, and vice versa. Therefore, it is important to think well about which methods and parameters are used in this algorithm. Besides this, it is important to generate the risk scores without considering sensitive characteristics. Defendants should not be advantaged or disadvantaged based on their race, gender of age.

The situation of unfairness in algorithms within criminal justice is sensitive. This is, because people can not agree on how to measure fairness in these algorithms. For this case-study on COMPAS, the measurement of the TPR and FPR is used. After calculating these, these rates are compared to each other to compare the algorithms. However, there were also other ways to do this, as described in figure 1. The choice could have be made to measure the PPV and NPV of each model for example. Then, the results of the experiments would probably have been very differently. Besides this problem, people also have a different preference on which mistake is worse; A false positive or a false negative. Together, a society determines what is ethical when it comes to this issue. However, this common preference also changes over the years. These developments have to be taken into account when algorithms like COMPAS are created.

When looking at the broad situation of the use of decision-making algorithms, we can suggest that the problems are the same. These algorithms are all built to gain more efficiency and provide smarter and better solutions. However, there is a risk on biased outcomes for all of these algorithms and it is hard to prevent this. Besides this, it is difficult to measure if these algorithms are biased. Every situation and used algorithm is different, which means that every measurement is different. Every biased outcome has some degree of impact on people. However, when the algorithms make predictions about important things in people's daily lives, the seriousness of these impacts becomes huge. Algorithms that predict whether someone can pay back a loan, is a good job applicant or is a criminal have a lot of impact on our society. That it is important that we make sure that these algorithms make fair predictions.

Unfortunately, the situation in the U.S. where COMPAS is used will probably not change soon. The status quo of criminal justice in the U.S. is lamentable. The criminal justice system is very underfunded and the people that work here are resistant to change [5]. As said earlier, this thesis has not the goal to provide possible solutions for this problem. The problem is not solved now and it will not be for a while in my opinion. Yet, the problem does get more attention from society, scientists, and other specialists. This will probably help to find a solution for this problem.

7 Conclusions and Further Research

In this thesis, a study on the nature of the unfairness in algorithms within criminal justice is done. The investigation consists of a literature review and a demonstration of the challenges that come up when building these algorithms. Experiments have been done to study the influence of algorithm variables on the prediction models that predict recidivism. Four algorithms are implemented with logistic regression and random forest classifier and two different threshold. After this, another

algorithm was programmed to zoom in on three different demographic groups, to study the influence of the threshold on these groups. Both experiments showed the huge influence that these variables have on the outcomes of the prediction model. Some groups had three different sorts of outcome, which means that the method and threshold have had a big influence on this. On the other hand, some group stayed on the same place in the graph. Some of these cases were favourable for the respective group, but others were detrimental. This implies that some demographic groups are advantaged or disadvantaged within these algorithms. Therefore, we can say that the results of this study show that COMPAS is discriminatory. Groups that were advantaged were mostly young, male, and African American. Groups that were disadvantaged were mostly old, female, and white.

For the experiments, not all data of the dataset is used. We only selected three kinds of races; African Americans, White people, and Native Americans. However, for the groups that were based on gender and age, all defendants were selected. This could cause selection bias for the prediction models. Also, some groups were very small. This could result in invalid predictions, because the test set of the prediction models were not big enough. As a result, there were a few outliers.

For further research, the nature of the unfairness within algorithms in criminal justice can be studied more. More similar research on COMPAS can be done in the future. These studies can use other sorts of experiments with other measurement methods. When there is more known about the algorithms of COMPAS, more research can be done about this. Moreover, more studies can be done about the definitions of fairness. When more experiments are hold to determine which notions of fairness should be used, there will be more public attention to it. Finally, more research can be done on other situations where populations are disadvantaged by AI-based algorithms used in decision-making processes. When these other situations are discovered, perhaps a solution can be found for them as well.

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. *Ethics of Data and Analytics*. Auerbach Publications, 2022.

[2] L. Belenguer. Ai bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics*, 2:771–787, 2022.

[3] A. Brackey. Analysis of racial bias in northpointe's compas algorithm. *ProQuest Dissertations and Theses*, page 46, 2019.

[4] J. Chakraborty, T. Xia, F.M. Fahid, and T. Menzies. Software engineering for fairness: A case study with hyperparameter optimization. *CoRR*, abs/1905.05786, 2019.

[5] V. Chiao. Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice. *International Journal of Law in Context*, 15(2):126–139, 2019.

[6] K.A. Clarke. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22:341–352, 2005.

[7] D. Hellman. Measuring algorithmic fairness. *Virginia Law Review*, 106:811–866, 2020.

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[9] K. Kirasich, T. Smith, and B. Sadler. Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, 1(3):25, 2018.

[10] R. N. Landers and T. S. Behrend. Auditing the ai auditors: A framework for evaluating fairness and bias in high stakes ai predictive models. *American Psychologist*, 78(1):36–49, 2023.

[11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54:1–35, 2021.

[12] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys*, 55:1–44, 2022.

[13] N.A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D.C. Parkes, and Y. Liu. How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence*, 283:103238, 2020.

[14] N. Scurich. Criminal justice policy preferences: Blackstone ratios and the veil of ignorance. *Stanford Law Policy Review*, 26:23–35, 2015.

[15] H. Suresh and J.V. Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8):73, 2019.

[16] S. Verma and J. Rubin. Fairness definitions explained. *FairWare'18: IEEE/ACM International Workshop on Software Fairness*, pages 1–7, 2018.

[17] A.L. Washington. How to argue with an algorithm: Lessons from the compas-propublica debate. *Colorado Technology Law Journal*, 17(1):131–160, 2018.

[18] M. Xiong, R.G. Greenleaf, and J. Goldschmidt. Citizen attitudes toward errors in criminal justice: Implications of the declining acceptance of blackstone's ratio. *International Journal of Law, Crime and Justice*, 48:14–26, 2017.

8 Appendix

The link to the github repository is: https://github.com/jortv/Thesis_Jort_Visser.git

8.1 Truth table

These are the other values that can be calculated in the truth table in 1.

- **Positive predictive value (PPV)**. The part of positive values that is correctly predicted out of all predicted positive values. This is also known as the precision.

- **False discovery rate (FDR)**. The part of negative values that is wrongly predicted out of all predicted positive values.
- **False omission rate (FOR)**. The part of true values that is wrongly predicted out of all predicted negative values.
- **False negative rate (FNR)**. The part of true values that is wrongly predicted out of all real positive values.
- **Negative predicted value (NPV)**. The part of negative values that is correctly predicted out of all predicted negative values.
- **True negative rate (TNR)**. The part of negative values that is correctly predicted out of all real negative values.

8.2 Logistic regression

The parameters that are used for the logistic regression algorithms are as follows. The other parameters that are not mentioned are set on the default value.

- **solver**. This defines the kind of algorithm that is used for the problem. Here, it is set on 'liblinear'. This kind of algorithm results in a short runtime for small datasets. Because the dataset that is used here is small, this method seemed appropriate to me.
- **random_state**. This parameter defines how many times the data is shuffled before fitting it into the model. Here, it is set on 400. Randomness in data is important when algorithms are trained by it, otherwise it causes bias. Models could learn patterns when the training is in a certain order.

8.3 Random forest classifier

The parameters that are used for random forest classifier are as follows. The other parameters that are not mentioned are set on the default value.

- **n_estimators**. This parameter defines how many decision trees there are used for the algorithm. Here, the parameter is set on 500. This choice is made because a higher value of this often results in a better accuracy of the model.
- **max_depth**. This defines the amount of splits that each decision tree is allowed to make. Here, it is set on 4. If this value is too low, it could lead to an underfitting model. It could lead to an overfitting model when this value is too high.
- **max_features**. This parameter defines the maximum features that should be used when splitting the node in the decision tree. A higher value of this could lead to overfitting, but it also leads to a more accurate model. Here, the parameter is set on 3.
- **bootstrap**. This means that random samples of the training data are used to make simulated samples for the decision trees. Here, bootstrap is set on 'True'.

- **random_state**. This parameter defines how many times the data is shuffled before fitting it into the model. Here, it is set on 400. Randomness in data is important when algorithms are trained by it, otherwise it causes bias. Models could learn patterns when the training is in a certain order.

8.4 Distance to base

Group	LR_0.5	LR_0.7	RF_0.5	RF_0.7
African Americans	0.19	0.04	0.16	0.07
Native Americans	0.94	0.97	0.92	0.17
Whites	0.15	0.01	0.19	0.05
Young people	0.32	0.10	0.06	0.22
Middle aged people	0.06	0.07	0.03	0.05
Old people	0.03	0.12	0.10	0.08
Females	0.07	0.00	0.10	0.02
Males	0.04	0.04	0.14	0.03
Young African American females	0.27	0.17	0.41	0.11
Young African American males	0.59	0.08	0.45	0.12
Middle aged African American females	0.09	0.10	0.23	0.06
Middle aged African American males	0.27	0.19	0.24	0.15
Old African American females	0.43	0.13	0.58	0.17
Old African Americans males	0.13	0.14	0.08	0.10
Young white females	0.40	0.13	0.53	0.19
Young white males	0.50	0.07	0.32	0.16
Middle aged white females	0.10	0.08	0.17	0.06
Middle aged white males	0.05	0.06	0.12	0.12
Old white females	0.18	0.13	0.53	0.83
Old white males	0.14	0.13	0.28	0.05

Table 7: The distance from the group point to the base for each experiment.