FACULTY OF ENGINEERING AND ARCHITECTURE

Bias and fairness in AI: bridging the gaps between research, business, and law

Marybeth Defrance
Student number: 01701893

Supervisor: Prof. dr. Tijl De Bie
Counsellor: Maarten Buyl

Master's dissertation submitted in order to obtain the academic degree of
Master of Science in Computer Science Engineering

Academic year 2021-2022

GHENT UNIVERSITY

Permission of use

The author gives permission to make this master dissertation available for consultation and to copy parts of this master dissertation for personal use. In all cases of other use, the copyright terms have to be respected, in particular with regard to the obligation to state explicitly the source when quoting results from this master dissertation.

MaryBeth Defrance

June 2022

Acknowledgements

The process of writing this dissertation gained me a lot of knowledge. The concept of fairness in artificial intelligence combines two interests of mine, namely the way the world works and technology. I have always been intrigued by the role the technologies that we develop can play in society. This dissertation gave me the possibility to research many different domains, such as law, economics and technology, confronting me with many different ways of thinking.

I would like to thank my supervisors Tijl and Maarten for supporting me and giving me the freedom to explore what I wanted to do in this work. While this freedom was daunting at times, it gave me the opportunity to explore and make something slightly different.

I would also like to thank the people who took the time to talk to me about their experience of implementing AI systems in the work field. They provided some interesting viewpoints and aided in keeping a broad view.

From the central organisation of Ghent University I would like to thank Hanne Elsen who helped me gain an insight on how organisations handle these legal aspects with regards to technology and more specifically AI.
Special thanks to Lot Fonteyne, who collaborated with me on the SIMON-test aspects of this dissertation. I was really glad that I could discuss my ideas with an expert on the topic and your openness for discussion was a motivating factor.

Finally many thanks to my family who supported me during the process of writing my dissertation, for trying to take as many things as possible from my plate so that I could put the time and effort in researching and writing.

Bias and fairness in AI: bridging the gaps between research, business, and law

MaryBeth Defrance

Supervisor: Tijl De Bie

Counsellor: Maarten Buyl

Master's dissertation submitted in order to obtain the academic degree of Master of Science in Computer Science Engineering

Academic year 2020-2021

Abstract

Bias and fairness in AI are topics that recently gained traction in research, but also with lawmakers. Bias and fairness in AI mainly focus on the possible mistreatment of people by an AI system and how to mitigate that risk. Many factors can cause disparate treatment and require both a technical and social eye to identify. The EU proposal for an Artificial Intelligence Act will force many companies to evaluate the AI they currently use and plan to use in the future. The goal of this dissertation is to go through the process of creating a simple logistic regression classifier and exploring the notion of fairness. This requires first looking at society and the role of the AI system in it. Then, different fairness definitions and types of bias are discussed. From there, a toy example is used to test and explain different simple techniques to influence an AI system's fairness. Finally, it is shortly discussed how and why an AI system must be monitored after being taken into production.

Bias and fairness in AI: bridging the gaps between research, business, and law

MaryBeth Defrance*

Supervisor: Prof. dr. Tijl De Bie
Counsellor: Maarten Buyl

***Abstract*: Bias and fairness in AI are topics that recently gained traction in research, and also with lawmakers. Bias and fairness in AI mainly focus on the possible mistreatment of people by an AI system and how to prevent this. Many factors can cause disparate treatment and require both a technical and social eye to identify. The EU proposal for an Artificial Intelligence Act will force many companies to evaluate the AI they currently use and plan to use in the future. The goal of this dissertation is to go through the process of creating a simple logistic regression classifier and exploring the notion of fairness. This requires first looking at society and the role of the AI system in it. Then, different fairness definitions and types of bias are discussed. From there, a toy example is used to test and explain different simple techniques to influence an AI systems's fairness. Finally, it is shortly discussed how and why an AI system must be monitored after being taken into production.**

***Keywords*: Fairness and Bias in Artificial Intelligence, Machine Learning, Dataset bias, AI Fairness definitions**

I. Introduction

Bias and fairness in AI are fairly recent topics, with the first mention of it being made in 2010 [1]. However, the topic gained much traction in recent years, with even several dedicated conferences [2]. Bias and Fairness in AI research focus on creating an AI system which treats everyone fairly, meaning no discrimination against non-relevant attribute of a person.

This dissertation will focus on the different decisions and analyses that must be made when creating a binary classifier. This binary classifier will use logistic regression and pertains to the subject of succeeding or not in higher education. In the theoretical analysis, the example used will be the SIMON-test, and during the practical analysis, the OULAD data set will be used because the data from the SIMON-test is not publicly available.

The SIMON-test is a tool for prospective students that predicts whether or not they are likely to succeed in a certain degree within higher education. Any disparate behaviour in such a system could negatively impact the futures of prospective students from a certain group within society.

The importance of bias and fairness in AI can be seen through likely the most well-known real-world example. This is the example of Northpointe's COMPAS software [3]. The AI system was used for a risk assessment on whether or not someone would commit a crime again. The original analysis of the system did not include whether it showed any discrimination.

After the software was already in use, an analysis showed a difference of two percentage points between black and white men. This indicated that the system did not discriminate. However, ProPublica journalists made a more thorough analysis, also taking into account the type of mistakes the system made. This is relevant as predicting someone would commit another offence when in fact, they would not is less favourable for the individual than the other way around. The analysis from ProPublica showed that people of African American descent were twice as likely to be predicted to commit another offence when in truth, they would not in comparison to white people making the AI system clearly discriminatory against people of African American descent.

II. Related work

The inspiration for this dissertation came from the work of Boris Ruf et al. in *Towards the Right Kind of Fairness in AI* [4]. In this work a type of fairness compass was constructed in order to help people decide on the correct measure to indicate fairness. This dissertation took a different route by not giving a recommendation but rather including all relevant information to give the reader the ability to make the decision themselves. The work of Tai Le Quy et al. [5] shows an interesting way of finding biases in the data sets through the use of Bayesian networks. In *A survey on Bias and Fairness in Machine Learning*, many sources of biases and different types of fairness definitions are included, together with some context, in order to understand the contents better [6]. The final work that is interesting within the context of bias and fairness in AI is the work of Allesandro Castelnovo et al. in *A clarification of the nuances in the fairness metrics landscape*. This work gives a strong socio-technological analysis of the concept and different definitions [2].

III. Social framework

The society in which the AI system is released is vital to determine the concept of what is fair and what is bias. The first elements that indicate these aspects is the relevant legislative framework. On the other hand, it is also important to analyse the business case for an AI system and what factors are important to make it profitable.

*M. Defrance is a student of Computer science engineering at Ghent University (UGent), Gent, Belgium. E-mail: marybeth.defrance@ugent.be .

A. Current European Union legislation

The current legislation in the European Union is not equipped to handle the arrival of AI systems [7]. Currently, the plaintiff must prove that an AI system makes incorrect predictions or, in other words, discriminates. However, an AI system is a fairly black-box system, making it nearly impossible to prove that it enforces disparate behaviour.

B. European Union proposal for an Artificial Intelligence Act

In order to handle these shortcomings in the current legislation, the European Union proposed an Artificial Intelligence Act. This would introduce a new system that requires certain high-risk applications to guarantee fair behaviour. These high-risk applications are applications that, when in use, could violate an individual's human. This characteristic of high risk depends on the situation in which the AI system is used. A couple of situations are explicitly mentioned in the proposal as high-risk, but these are not limiting.

This makes the SIMON-test an interesting example of an AI system as it operates in one of the named high-risk areas, namely education. If a certain group receives worse predictions than they should on average, this directly affects their future and subsequent opportunities.

C. Business aspect

Businesses also benefit from creating fair AI systems, or rather they would suffer from implementing an unfair system. Implementing an unfair AI system would cause harm to the image of the company and could cause customers to stay away [8]. It could even further lead to more restrictive legislation being introduced if unfair AI systems became too prevalent [9]. This legislation could lead to more expensive development costs or less powerful models, both possibilities being undesirable.

IV. Types of Fairness

It is important first to define what fairness is. The Cambridge Dictionary defines fairness as the quality of treating people equally or in a way that is right or reasonable. However this definition is not interpretable by computers which depend on mathematical expressions. This means that this subjective interpretation needs to be translated into a mathematical expression and thus quantifying what fairness is. This expression is also called a fairness definition.

There are two main types of fairness definitions. The first type consists of definitions that work on an individual level. These mainly work on the premise that similar individuals should receive the same results. While this type might seem the best, it has shortcomings in that the definition is either too lax or very difficult to implement correctly without introducing a different form of bias.

The other types of definitions work with group fairness. These definitions require a form of equality between two groups in a statistical metric. This division of groups happens on a sensitive attribute. A sensitive attribute is a characteristic of an individual on which a decision should not be made. Common examples of sensitive attributes are age, gender, and ethnicity.

Often, one of these fairness definitions are not sufficient on their own and requires some combination of them. However, not all definitions can be combined, specifically those from the group fairness definitions. This is due to the equality of the statistical metrics; the relationship between these metrics is partially dictated by their base rates (the fraction of actual positives in the set). As base rates can differ between groups, this means that not all statistical measures can be equal due to this dependency on the base rate.

Deciding what definition best encapsulates the system's goal, is very dependent on the context. In the case of the SIMON-test, there is a clear value difference between false positives and false negatives as mistakes. A false positive predicts that a student would succeed at a certain degree while they would not in reality; false negatives are the inverse of this. It is thus much worse for a student to receive a false negative compared to a false positive. Therefore the fairness definition is dependent on the fraction of false negatives compared to true positives. This definition is called equal opportunity. However, equalling this statistical measure is slightly insufficient, and therefore it also seems best to require equal accuracy between both groups.

V. Dataset analysis

A. Theoretical sources of bias

An AI system would not create an unfair situation by itself. Instead, this is due to bias inside the data set on which the system was trained. This bias arises from the people who generate the data for the data set. It is beneficial to know how and where these biases can arise in order to possibly prevent these biases or at least be aware of their existence.

In order to understand where the biases are introduced in the data set, it is necessary to know how the life cycle of data gathering works. However, these different stages influence each other, creating a type of feedback loop portrayed in Figure 1.

The data collected for fair AI systems contain some form of human information. This information can be collected passively or actively but will always require some form of human input. This human input is the first possible location where bias can be introduced [10]. These biases can arise because of a different setting in which the data is collected. If this difference is inherent to one of the groups defined by the sensitive attributes, then it introduces bias. Another possibility is that the bias is introduced by inherent differences between these groups in society. Although part of society, these differences can often still be undesirable, especially with older data.

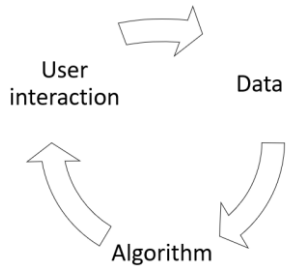


Figure 1: The feedback loop between different stages relevant to data collection

The next step in the cycle is the data itself, meaning the collection of all the user input. Bias in the data set can be correlated with how accurate the data set is compared to the real world. These inaccuracies or biases can be that the collected data does not reflect the use population of the application. This means that certain groups are under or overrepresented. Another possibility might lie in the data itself, with certain features not reflecting what they should or missing information in the data set. These biases are almost surely going to be present because making the perfect data set nears the impossible.

The final point where bias can arise is through the algorithm itself. The first possible way an algorithm introduces bias into the data set is that the algorithm dictates with what a user can or most likely will interact. If the algorithm does this in a disparate way across users, then this will result in skewed data. Another possibility is that bias is introduced during the development of the new algorithm. These are types of biases the developer needs to be wary of and prevent them from happening. The final type of bias introduced by the algorithm is related to the deployment of the algorithm. If the deployment is skewed from the original intention or certain elements shift in its environment, then bias can arise because it is not used in the setting for which it was intended. This means that the decisions made during the development process can be wrong as they were not made with this type of use of the application in mind.

B. Analysing biases on the SIMON-test

The analysis for the user interaction type biases is possible with the current knowledge of the SIMON-test because the platform is available to analyse. The most likely introduction of bias by the user interaction would be the environment in which the prospective students fill out the SIMON-test. This is an uncontrolled variable, but it could significantly impact their results for the different tests inside the system.

As the data set of the SIMON-test is not available, it is near impossible to determine if any biases are present inside it. Biases through the algorithm can be evaluated from a theoretical standpoint. Biases that arise because the algorithms show different things to different people will not arise in the SIMON-test as there is no AI behind it with which a prospective student would interact. Currently, a simple form of regression is used as the model, making it unlikely that the development process introduced bias into the system. However, this is impossible to verify without access to the algorithm and data itself. Finally, the deployment of the system is carried out by the same people that developed it. This, together with the regular upkeep, makes it unlikely for any bias to be introduced through its deployment.

C. The OULAD data set

The OULAD data set is a public data set that contains the activity information of students on the online learning platform of the Open University [11]. The data in this data set is then transformed into training data for the task of predicting whether or not a student would pass this course based on their activity on this online learning platform. One of the strongest advantages of the OULAD data set is that a lot of sensitive attributes are encoded into it. It is possible to check the fairness of whether the student has a disability, their gender, their age range, and welfare based on where they live.

For the OULAD data set, an extensive analysis was performed on the data set in order to detect biases. The distribution in the data set was compared to the demographic characteristics of society in order to assess if the data set was representative of society. However, this is a generalisation as it is not sure if the general population should be seen as the user population. In the creation of the data, the labels were also simplified in order to make the decision binary while still containing a sufficient amount of data.

VI. BIAS MITIGATION TECHNIQUES

The biases discussed in the previous section cannot always be prevented from occurring in a data set, such as biases that arise from society itself. Therefore, it is sometimes necessary to compensate for these biases in the algorithm. In order to do this properly, a fairness definition or a combination thereof needs to be established. As mentioned before, not all definitions can be used at the same time, which is something that needs to be taken into account when choosing a combination of fairness definitions. The goal of these bias mitigation techniques will be to satisfy the chosen definitions.

Within these techniques, three categories can be distinguished. These categories are based on the moment in the process of training the machine learning algorithm where they intervene. While a certain type of technique might be much more effective than others, there is also an element where it may be possible to apply the technique given certain constraints most systems are under. The three categories are pre-processing, in-processing and post-processing.

A. Pre-processing

Pre-processing techniques will make changes to the data before it is fed into the machine learning algorithm. The goal of this technique is to remove bias from the data itself. This bias removal can be performed pre-emptively or at runtime [12].

A technique that is somewhat controversially called pre-processing is fairness through unawareness. In fairness through unawareness, the sensitive attributes are removed from the data set. This means that two people who only differ in their sensitive attributes will receive the same prediction because the algorithm cannot determine the difference between them. The system will thus not have any causal discrimination, which is when two people differ in a sensitive

attribute, which causes them to receive different predictions. While not having any causal discrimination is very good from a legal standpoint, it is often insufficient. Due to the power of finding correlations between data, AI systems can still discriminate based on the latent effects of the sensitive attribute on the relevant features.

An extension of fairness through unawareness is suppression. With suppression, the sensitive attributes are used to remove the correlation in the relevant features with the sensitive attributes. This is thus stronger than fairness through unawareness as the sensitive attributes are also removed from the data before going into the algorithm. However, as the correlation is removed from the data set, it changes the relevant features. This correlation is removed based on the correlation that could be found in the training set, so this is different depending on the sensitive attributes of the person. This entire process means that it is possible to have causal discrimination in the system, making it more legally more difficult.

B. In-processing

In-processing techniques work directly on the mathematics of the algorithm that tries to find the relations in the data to predict the label [12]. This means that in-processing techniques convey the goal of fairness into the system. Because in-processing techniques work on the inner workings of the machine learning algorithm, they can be slightly harder to interpret. However, the largest problem with in-processing techniques is that they require access to the algorithm itself. In many companies, but even in many applications such as sklearn, this can be quite difficult to achieve, making them less desirable.

One very simple in-processing technique is discussed, where the weights for the samples in the loss functions is are customised. The weights for the positive samples were set to be three times the weight of the negative samples. This conveys to the algorithm that it is more important to avoid false negatives than false positives, as was determined by the problem statement. In very sensitive situations, this could decrease the model's overall accuracy. However, in most cases, adjusting the weights of some samples by a non-exorbitant amount will result in a slight loss of accuracy but increase the fairness. This increase of fairness is likely to happen but is not guaranteed and should always be checked. This checking is relevant for all bias mitigation techniques used.

C. Post-processing

Post-processing techniques make changes after the model has been trained and after the prediction has occurred [12]. A post-processing technique will most likely not benefit from receiving the label the model predicted, but rather the value of the logistic regression functions. Based on other parameters and characteristics of the sample it can then derive the label.

A very simple post-processing technique is changing the threshold for a binary label based on the sensitive attributes of the label. This means that, unlike in normal machine learning algorithms, the positive label will not be given if the value from the logistic regression is above 0.5, but rather a higher or lower threshold. These thresholds will depend on the sensitive attributes of the sample. Like the pre-processing technique suppression this means that samples which only differ in their sensitive attributes can receive different labels as their thresholds can be different. In other words this system can exhibit causal discrimination.

VII. Monitoring Bias

After creating a fair AI system, the work is not done. As this application lives inside an ever-changing society, it is possible that it does not fulfil its original purpose if it does not handle the society's change with it. This means that fair AI systems require monitoring after they have gone into production [13]. Often an AI system still collects data after it has gone into production. This data could be used to retrain the model to be more performant. However, it might be that the new data might be skewed for some reason; therefore, when retraining a model, it is necessary to see if all fairness requirements set up during the development are still valid.

It is also advised to check for other biases in the system regularly. While the hard check to see if the fairness requirements are still satisfied, having a critical analysis of all possible biases might predict future unwanted behaviour.

One of the most important biases to keep an eye on when a system has gone into production is if it is still being used as intended. The entire AI system was tuned to be used in a certain way. If there is a significant deviation from this original concept, then discriminatory practices might take place.

VIII. Conclusions

Bias and fairness in AI is an upcoming topic that deserves a significant amount of attention. As already stated by the European Union, if an AI system is brought into use that is not fair, then there is a real possibility of people's rights being harmed. On top of this, adding that AI systems work automatically and on a large scale could negatively influence many lives. This calls for ethical and fair AI systems to be developed, which rather than hurting society, will help it thrive.

These AI systems are most likely going to be deployed by businesses and governments. In order for fair AI systems to be developed the people who are going to use them are involved in their development in order to reflect correctly what their function in society will be. The goal of this dissertation is to give the insights necessary to develop and work consciously with these systems in order to get the best possible results for everyone.

References

[1] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, (New York, NY, USA), p. 59–68, Association for Computing Machinery, 2019

[2] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, "A clarification of the nuances in the fairness metrics landscape," Scientific Reports, vol. 12, no. 1, 2022.

[3] J. Larson and J. Angwin, “Machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks.,” May 2016.
[4] B. Ruf and M. Detyniecki, “Towards the right kind of fairness in AI,” CoRR, vol. abs/2102.08453, 2021.
[5] Tai Le Quy, A. Roy, V. Iosifidis, and E. Ntoutsi, “A survey on datasets for fairness-aware machine learning,” CoRR, vol. abs/2110.00530, 2021.
[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” CoRR, vol. abs/1908.09635, 2019.
[7] M. Fierens, E. Van Gool, and J. De Bruyne, “De regulering van artificiële intelligentie (deel 1) - een algemene stand van zaken en een analyse van enkele vraagstukken inzake consumentenbescherming,” Rechtskundig Weekblad, vol. 2020-2021, p. 962–980, Feb 2021.
[8] R. V. Loon, “How corporate c-levels can be the guardians of ethical ai,” Jun 2020.
[9] World Economic Forum Global Future Council on Human Rights 2016-18, “How to prevent discriminatory outcomes in machine learning,” tech. rep., World Economic Forum, March 2018.
[10] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, “Social data: Biases, methodological pitfalls, and ethical boundaries,” Frontiers in Big Data, vol. 2, 2019.
[11] J. Kuzilek, M. Hlosta, and Z. Zdrahal, “Open university learning analytics dataset,” Scientific Data, vol. 4, no. 1, 2017.
[12] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, and et al., “Bias in data-driven artificial intelligence systems—an introductory survey,” WIREs Data Mining and Knowledge Discovery, vol. 10, no. 3, 2020.
[13] World Economic Forum Global Future Council on Human Rights 2016-18, “How to prevent discriminatory outcomes in machine learning,” tech. rep., World Economic Forum, March 2018.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation focuses on bias and fairness in AI with the goal of going through the process of creating a fair AI system. The AI system used as an example in this dissertation will be a binary classification problem that uses logistic regression. This classification problem will focus on succeeding in higher education, with the SIMON-test being used as a thought experiment in the theoretical analysis. The goal is to give the reader an understanding of the different elements that are in play when developing a fair AI system and provide an insight why technical people alone cannot make all of the decisions on their own but rather require input from the experts that will make use of the application.

An AI system needs to be fair if it is used in a decision process concerning people. This fairness attribute means that there will be no disparate treatment of people based on non-relevant factors such as age, race, sexual orientation, etc. It has been shown that creating an AI system without ensuring fairness guarantee that it can lead to situations where people are being discriminated against.

The most famous example of a biased AI system is Northpointe's Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS [1]. The goal of the application was to predict the risk if an offender would reoffend. By 2010, nearly all of New York State was using this software as a tool to determine whether people should be allowed probation. An evaluation of the tool published in 2012 by the state of New York also showed that it was 71% accurate in its task. Northpointe conducted a validation study in 2009 on a sample of 2 328 people. In this study they found that the tool was 67% accurate for black men and 69% accurate for white men. This difference in accuracy was deemed small enough in the eyes of the company, and they stated that the tool did not discriminate.

This would have been true if there was not a clear difference between the possible predictions of the tool. It is far worse for someone to be deemed high risk when, in fact, they are not more so than the other way around. Assessing such a tool solely on the accuracy between groups is insufficient as it does not capture this value difference of the predictions. ProPublica eventually performed the analysis, comparing which incorrect predictions were made by the system. It turned out that the AI system was nearly twice as likely to incorrectly label someone who is African American to re-offend compared to a white person when in truth, they would not reoffend. It is clear that such behaviour disfavours people of African American descent, making the tool discriminatory.

This example of the COMPAS tool shows how important it is to thoroughly evaluate a tool before it used somewhere it could impact people's lives significantly. This evaluation should be done by both the company designing the tool and the institution using the tool. The author of this dissertation believes that the elements discussed in this dissertation can help people in gaining an insight into the decisions and relevant aspects for creating such a tool and evaluate what is required for the application to be deemed fair.

The use of AI applications is expanding which require these guarantees, strengthened by companies such as DataRobot offering software solutions to create fair AI systems [2]. However, when using such applications, it remains important that the company which will implement this software has the knowledge to determine its own requirements for fairness as they are know the social landscape of the tool best. Even local governments started looking into AI systems to strengthen their operations with AI. For example, the Flemish Government has planned to experiment with AI applications such as risk analysis of child abuse or automating repetitive administrative tasks [3].

1.1 SIMON-test

The SIMON-test[1] is used in this dissertation as a thought experiment throughout the Chapters 3, 4 and 6. The SIMON-test is a tool which aids prospective students in choosing their field of study in higher education. The reader will be taken through the thought process of Fairness in AI with the hypothetical process of making the SIMON-test or a similar system and checking whether it is fair. The data from the SIMON-test itself is not available for analysis; therefore, this process is mainly hypothetical. For this reason, the SIMON-test will not be discussed in chapter 5, but a different data set related to education will be used.

[1] www.vraaghetaansimon.be/dashboard

The SIMON-test is an existing assessment tool made by the University of Ghent that predicts a student's chances of obtaining their bachelor's degree. It also provides a tool to gauge the student's interests in different topics. These tools work in a degree-specific manner. There are four categories of information which the system uses to determine these scores. Personal information is gathered about the students themselves, such as age, previous education and hours of mathematics they had in high school. However, only the hours of mathematics is used for the predictions. The other information is used to test if the systems disfavours a certain group of people. The other three aspects used in the predictions are results from tests on the platform. The second category and the first of these three types of tests gauge the student's study abilities, including test anxiety and self-control. The third category is a collection of tests about non-cognitive abilities; like academic confidence and motivation. The last group of tests estimates the student's cognitive abilities. The tests for gauging cognitive abilities on the platform are reading comprehension, vocabulary, chemistry, physics, basic mathematics skills, mathematics for sciences and language proficiency.

The novelty of the SIMON-test lies in that it combines cognitive and non-cognitive skills in order to try and predict if a student will succeed. This creates more of a complete picture of the student compared to when only a subset of elements is used. This combination of skills also makes the data highly related to the individual student. While this is evidently the goal of the system, it also bears some difficulties. It increases the chances of an AI algorithm making bias decisions based on biases present in past data. These biases would then occur on a student's personal attributed that should not influence the decision, such as gender, socioeconomic situation, race, etc. The first solution that comes to mind to mitigate these biases is not including the personal attributes in the data. This works to some extent. However, due to the power of computers and their abilities to recognise patterns in data, it is not unusual that they can infer this information from the other attributes that are relevant to the decision. The possibility of this occurring increases in systems such as the SIMON-test because their features are highly correlated with the individual.

The SIMON-test is an interesting case to discuss because of these data characteristics and its domain of application. The application domain of the SIMON-test is interesting as education is one of the "high-risk" domains specifically mentioned in the EU legislation proposal [4]. It is high-risk because a student's degree has a large influence on the rest of their lives. If a system aiding in the decision for a degree shows bias, it can lead to disadvantaging certain groups of people. This does not mean that such an application should not exist; the application has a useful goal. It is mainly important that the possibility of bias is highly monitored to prevent them from occurring.

This dissertation discusses the SIMON-test as a specific example of a high-risk AI system. The decision to use a precise example was for the tangibility of the process and the ability to work with the experts from the university who designed the application. The analysis can be generalised to other student assessment tools for determining their success before starting their education. A different example of such a tool is *Luci - Leuvens Universitair competentie instrument*[2]. This tool only contains one test. Other tests are available on the platform, but these were made independently from the platform. There is no selection process during admissions in higher education in Belgium, meaning that every student can start the degree they want (with a few exceptions). However, in other countries, there may be a selection process between students regarding who is allowed to start a degree, such as is the case in the United Kingdom. In those cases, the use of such a tool that predicts a student's success can become part of the selection process. While the SIMON-test has much lower stakes as a student can decide autonomously, implementing such a tool during a selection process can have dire consequences if the tool turns out to be biased. The general concepts are still very similar between these tools, the stakes are different. The analysis made in this dissertation could provide support when creating or assessing either type.

1.2 Related work

The inspiration for this dissertation came from the *Towards the Right Kind of Fairness in AI* paper by Boris Ruf et al. [5]. In this work a type of fairness compass was created to aid in choosing the correct fairness definition for an application. The idea of this dissertation is also to aid in making the correct decisions related to fairness, however more in an informative matter and complete fashion. Another interesting work to read is *A survey on Bias and Fairness in Machine Learning* [6]. This work contains many origins of biases and fairness definitions similar to what is contained in this dissertation. The paper of Tai Le Quy et al. *A survey on data sets for fairness-aware machine learning* [7] has an intriguing approach to analysing of data sets through the use of Bayesian networks. A similar work to this dissertation, however not as elaborate is the work of Alessandro Castelnovo et al. in *A clarification of the nuances in the fairness metrics landscape* [8].

The earliest mention of fairness as a goal for model optimisation only appeared in 2010 [9]. However, in recent years, it has gained much momentum with several dedicated conferences being organised [8]. This leads to many more papers being published, not only from a purely technical standpoint but also with a strong social aspect.

[2] www.kuleuven.be/luci/verken_jezelf

Chapter 2

Social Framework

In this chapter, the social framework in which the AI system would function is discussed. These aspects are important to keep in mind when developing an AI system that needs function fairly. As it is the social framework around the AI system that determines what would be fair. The system should first and foremost comply with the laws and regulations in effect. This is necessary not to have any legal liability, but it is also a first step in creating an application that mirrors society's values. The current legal framework in Europe is not equipped for AI systems. Therefore the European Union currently has a proposal for legislation, the European Union proposal for Artificial Intelligence Act [4]. If this proposal passes, there will be specific legislation on the use and development of AI systems.

The second aspect is business. Although some critics warn against the use of AI and its effects, there are also many experts who indicate that the adoption of AI could create large economic gains [10]. Therefore it is in the best interest of businesses to start this adoption early and do this in a correct matter as implementing a bias AI system could cause harm to the company's image. These aspects will be discussed in Section 2.3.

2.1 Current European Union legislation

The current legislation on consumer and personal protection is insufficient to cover the uses of digital products, certainly not related to artificial intelligence [11]. Currently, there are some guidelines around informing the user if an AI system used somewhere in the process. For example, users should be informed if an AI system is used to create custom pricing or the parameters used when ordering results in search engines.

Another aspect of consumer protection which in a way pertains to AI systems is the updates required for digital systems. This means that businesses are required to adjust the AI system itself if it does not satisfy the initial conditions, and this non-compliance falls within the legal guarantee period. This means that if the conditions in a contract are specified to include fairness as agreed, then that could be seen as some legal guarantee. This does however mean that some oversight needs to kept in order to monitor if these conditions remain satisfied.

Discrimination is illegal in Europe, as is noted in Article 21 of the *Charter of fundamental rights of the European Union* [12]. This states:

> Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

However, current anti-discrimination laws are insufficient to protect people from discrimination in AI systems as the proof of discrimination is not tangible. Many AI systems do not work in a rule-based fashion, making it fairly easy for a company to refute an individual's the claims that their systems are discriminatory. It is incredibly difficult to provide real proof of discrimination by an AI system as there will not be a specific rule stating to treat someone differently based on a protected attribute. This makes the black-box nature of an AI system something companies could use to protect themselves from liability. In fact this means that current legislation does not protect people from possibly discriminatory AI systems.

2.2 European Union proposal for Artificial Intelligence Act

Due to the shortcomings mentioned above in AI systems, the European Union is currently creating legislation around AI under the directive of President von der Leyen [4], to fill in the shortcomings that the current legislative framework has concerning AI. The proposal mainly focuses on defining what use of AI should be prohibited and certain high-risk areas for which extra precautions should be taken and what those extra caution measures should be. The idea is that the law strengthens the use of AI while still protecting the people of the European Union.

Certain uses of AI applications should be outright banned, as they give no benefit to the people and would cause harm. These prohibited applications include (but are not limited to) an AI-based social scoring system for general purposes by public authorities, similar to what is currently implemented in China [13] and 'real-time' remote biometric identifications (such as person identification based on cam-

era images) for the purposes of law enforcement. However certain exceptions exist for using 'real-time' biometric identifications.

Certain other uses of AI applications can be deemed "high-risk". The term high-risk is granted if an AI system's intended purpose could pose a high risk to the health and safety or the fundamental rights of persons if the system would not function in a fair manner. This could, for example, occur if the system were to contain bias against people with a certain background, resulting in less favourable treatment of certain groups within the population. The concept of fair in a legal setting can be seen as without discrimination. This aspect of the proposal for legislation is thus the most interesting for this dissertation. Certain key elements from the proposal will be discussed in the following paragraph.

Interesting requirements for these high-risk AI systems in the proposal are that the high-quality of the data set must be ensured for the training, validation and test set, the development and performance of the system must be documented throughout its lifecycle, the system must perform consistently throughout its lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity and further the system must be designed and developed in such a way that a natural person can oversee its functioning [4, (§43, §44, §46, §48 ,§49)]. This means that the AI system needs to be developed and maintained properly by law. These are factors that should increase people's confidence in such systems and ensure that their rights will be respected in the long run.

2.3 Business aspect

The main adoption of AI systems into regular life is for them to be carried out by businesses. This can be done by using AI systems internally, providing an AI system as a service to their customers, or making a product that makes uses an AI system. In order for this dynamic to function, both the business and their customers must be open to using this technology. Either one could be hesitant but for completely different reasons. This hesitation is undesirable as it generally slows down innovation. The following paragraph contains elements important to take into account when developing and implementing an AI system.

Adopting an AI system in the internal working of a business can happen in two possible fashions. One way, called automation, aims to replace certain tasks with an AI system. The second method is called augmentation, where the goal is to improve a person's performance by using an AI system [14]. Providing an AI system as a service or in a product can benefit a business of tapping into a business segment with a lot of potential.

Sometimes high costs are correlated to adopting an AI application [14]. These costs can be for infrastructure investments, the cost of the data and the cost of developing the application. Because of the possibly high costs, it is important to create these applications correctly in order to ensure a return on investment. This means that these systems need to be developed correctly, without cutting corners. From discussions with people in the field, this also returned as an important factor. If an AI system developed for one particular task would be recycled for another, then this can lead to problems. It could be that such a system is difficult to adopt in the desired business processes and often does not perform as well as on the original problem. This would be a costly investment for the business with very little return, making that the organisation might be less likely to try implement AI systems again in the future.

There are also a couple of hurdles when implementing an AI system from scratch, the first of which is the people within the organisation who will use the application once it is implemented. Through discussions with an organisation currently implementing a basic version of an AI system, certain key factors to smoothen the process were identified. First of all, the people who were going to use the AI application most were invited to brainstorming sessions. In these sessions, they had the possibility to voice their opinion on the functionalities and workflow of the system. This led to the employees being more open about the eventual arrival of the technology. Next, a close relationship should be kept between the developers and the department that would use the application. Through regular meetings the developers are able to run certain mechanics by the users in order to keep them fully aware and onboard.

The last and possibly most important hurdle is the client's perception of using an AI system. In the case that poor AI systems, such as systems that show discrimination, were to go into production it could make clients distrust these technologies. This loss of trust from the clients can directly result in lower revenue streams and hurt the image of the company [15]. However, not only could this cost the company in revenue, but if these poor systems become frequent that might lead to harsher regulations [16], which in turn often results in more difficult and costly design processes.

Chapter 3

Types of Fairness

3.1 Defining Fairness

The Cambridge Dictionary defines fairness as the quality of treating people equally or in a way that is right or reasonable. It is already not easy for humans to determine what is fair and what is not. The concept of fairness is far from universal and depends on the society in which someone lives [17]. This indicates that fairness is subjective and is based on the morals of a society. It further means that a computer in and of itself cannot to determine what is fair, as the terms subjective and computer do not go together. In order for a computer to determine whether something is fair, fairness must first be quantified in a way for a computer to be able to interpret it, meaning translating a social concept into a mathematical equation.

An important concept to discuss in fairness is sensitive attributes. These are characteristics of a person on which no judgements should be based as per the *Charter of fundamental rights of the European Union* [12]. It is however possible that it can be justified to use some of these attributes listed in the charter in a decision process. Justifiably using those attributes is called explainable discrimination [6]. The concept of sensitive attributes does not contain the features that would fall under explainable discrimination for that situation. An example of explainable discrimination is the variable of age when applying for a mortgage, which makes it not a sensitive attribute. However, there should be no discrimination on age in an employment situation, making it a sensitive attribute in that situation. Common sensitive attributes are gender, age, income, birthplace and sexual orientation. When defining a fairness definition, the concept of sensitive attributes is often used to formulate what is fair. In Section 3.3.2 about group fairness definitions, interchanges a group with a certain sensitive attribute with the term protected group.

3.2 Data properties that influence fairness

In the following section, different data properties, which have an influence on fairness definitions are explained. First of all, the real-world effects of different labels are explained. This is not inherent to AI systems but rather a consequence of the social setting in which fair AI systems work. The following paragraphs give critical insights into the features and labels of data sets. The more distortion or noise exist in the features and labels of the data set, the lower the quality of that data set. If these distortions were to start systematically disadvantaging people with a certain sensitive attribute more than others, then these distortions become biases. These biases are often the cause for unfair behaviour of AI systems and are thus influential factors in the entire process.

The preferred outcome for the user is an important factor for the fairness of an AI system. In a situation where the system would make perfect decisions, it would not be necessary to distinguish between labels. Alas the world is not perfect, and it is necessary to determine what the difference entails for the user. Certain mistakes the AI system could make entail worse consequences for the user than others. For example when applying for a loan, incorrectly not being allowed to receive a loan is much worse than getting a loan which the user will not be able to pay back. Perhaps the bank would prefer the latter, but the focus is on the user and not the owner in a context of fairness. When working with fairness in AI, the differences between incorrect predictions should always kept in mind.

The features on which the AI system bases its decisions are evidently important. The first question to ask is which features can be collected about the user, whether they are relevant to base the decision on, and how these features will be represented? Representation is an important factor as often some form of aggregation is used to simplify features and this aggregation should not remove important information [9].

The last discussion point is the ground truth used in the data. An example of an ambiguous ground truth is in a system that predicts whether a certain car needs to stopped in traffic [18]. The data shows that stopping a car with African American people had a higher chance of being justified. An investigation revealed that officers were equally likely to stop people of African American descent as someone of different ethnicity. However, it turned out that the officers would search the vehicles of African Americans more frequently than other people they have stopped. This led to a higher chance that stopping a car with African Americans would be justified. In this instance, the data set does not correctly represent the ground truth even though there is no mislabelling. Shortcomings in the ground truth of the data is something that should be taken into account in the choice of which fairness measures are optimal.

Another sensitivity is the labels itself of the data set to represent the ground truth. Often, the prediction that an AI system makes is simpler than reality. This is because certain aspects are abstracted in order to simplify the classification [9]. To return to the example of applying for a loan, the question is what would count as paying back the loan. It could be never to have missed a payment, paying the loan back by the agreed-upon time, or there might be some leniency? These different interpretations can affect different groups in different ways, making it an important factor.

3.3 Fairness grouping

Two types of fairness definitions can be distinguished; individual fairness and group fairness. As the names might suggest individual fairness defines fairness on an individual level, while group fairness ensures that different groups of people are treated equally, meaning that in group fairness, a certain individual could still be treated unfairly. It seems obvious that individual fairness would be preferred to group fairness, and from a theoretical standpoint, it is. However, the difficulty of applying individual fairness fully makes it less desirable. Some literature also suggests a definition that falls between both these categories, but as they are not common and not proven to be useful, they will not be discussed.

3.3.1 Individual Fairness definition

Individual fairness is defined as how individuals with different sensitive attributes should be treated. This section contains two definitions with their respective techniques. The first definition that prevents causal discrimination with fairness through unawareness is easier to achieve but does not impose a strong enough condition to be seen as real fairness. The middle section about suppression is an additional technique to achieve the first definition and improve further upon it. The second definition imposes a much stronger condition on the system, but is much harder to implement practically.

No causal discrimination [19]

Causal discrimination occurs when two people receive a different prediction from an AI system, even though they have identical values for their informative features and only differ in their sensitive attributes. This is a direct form of discrimination, and if such a case can be proven for any system then it could be grounds for a lawsuit even in current legislative framework. When creating an AI, system it is necessary to check whether the system shows any indication of this behaviour.

In order to have causal discrimination in an AI system, one can assume that there is something wrong

with either the data or the algorithm used. For example, if a rule-based approach is used in the AI system and shows signs of causal discrimination, then one of the rules makes an unfair decision based on one or a combination of sensitive attributes. If the AI system is something like a simple logistic regression and was to show causal discrimination, this could indicate data bias. It is possible that the current feature set in the system does not provide sufficient information, and the model is encoding some noise due to a lack of better information. Another possibility is that the ground truth is not correctly encoded in the decision process. This could either mean that task of the system is not correct or that the data used to train the system contains bias. This bias is often present because of human bias. In Section 4.1, the different ways bias can be introduced into a system are discussed.

A popular method of preventing causal discrimination is called *Fairness through unawareness*. In this method, the model is not given any sensitive attributes either as a feature or any other form. As an AI system is generally deterministic when predicting, causal discrimination becomes impossible. Two people who only differ in their sensitive attributes seem like the same exact person to the system. Thus when the same input is provided twice, same prediction will be returned. Note that this deterministic characteristic exists when making predictions with an AI system, the training is often not deterministic. This non-deterministic behaviour can be due to how the data is split for training, which is always slightly random, or because the system itself is initialised with random weights.

The problem with fairness through unawareness is that it is not a sufficient measure to guarantee fairness. While the sensitive attributes are not directly passed to the system, they can still have a hidden effect on certain features more relevant to the task. An example of this can be found in the *German Credit Data* data set [20], where indirect discrimination is present against non-single women [21]. The model could still sense the bias in the data set and disadvantage a certain group based on slight differences in their features. Therefore fairness through unawareness is insufficient to determine fairness, but preventing causal discrimination is mandatory for any AI system to be deemed fair.

Suppression [8]

A radical approach to solving the remaining problems in fairness through unawareness is suppression. Suppression aims to remove the remaining correlation between the useful features and the sensitive attributes. This can be done in a fairly drastic approach where the useful features that are highly correlated with a sensitive attribute are removed from the feature set. A decision must also be made as to what amount of correlation would count as a high amount of correlation. This has the downside of removing quite some information from the data set and thus making the task harder.

Another solution was proposed where instead of removing the highly correlated features, they are transformed in order to remove or at least lessen the correlation with the sensitive features. This can be done by projecting the feature space onto an orthogonal space. However, using this technique can become very convoluted due to feature interactions. Another possible solution would be to learn a fair representation of the data set. This fair representation should be able to reconstruct the original representation with as few errors as possible while remaining independent from the sensitive features. Removing the correlation from the useful features is done in Section 5.2.2 as an example of bias mitigation.

Fairness through awareness [22]

Fairness through awareness goes a step further than having no causal discrimination. In fairness through awareness, a distance metric is defined that can determine how different two individuals are from each other. These differences should not be determined by the sensitive attribute of the individual but rather based on useful features. If the individuals are not very different from each other according to the distance measure, then they should receive the same predictions. This is called the Lipschitz property.

The first difficulty with this definition is determining that distance metric. It would be extremely dependent on the social setting in which the application will be used. The distance metric should be based on relevant similarities and differences in the respective features. The importance of each feature must be determined and encoded into the system, meaning that an actual mathematical answer to what is fairness must be formulated. The second difficulty is using this definition itself. It requires implementing the distance metric and also using it in an AI system in which this metric type can be enforced. These difficulties can be a contributing factor as to why this type of fairness is not very popular to use [6].

3.3.2 Group Fairness definitions

Group fairness definitions are statistical in nature. An overview of basic statistical metrics useful for understanding the following definitions can be found in Appendix A and will be added to the relevant definitions. In the following sections, we will discuss different definitions and some trivial examples of classifiers that satisfy these definitions in order to aid in forming a critical view of these definitions.

These definitions are created to use in a situation where the people in the data set can be split up into two groups. The definitions work by requiring the equality of some statistical measure between both groups. The split of these two groups happens on a sensitive attribute such as gender. For example,

one group could consist of people who identify as male and the other group of people who identify as non-male. True equality will often be an impossible requirement. Therefore equality should be seen as a margin of difference between both groups. An example of a rather strict form of equality is a difference of the statistical measure of less than one percentage point. It is possible to extend these definitions to be able to handle more groups, but this will make it more difficult to satisfy the definition, perhaps requiring setting the margin for equality to three percentage points. Out of historical context, one of these groups can often be seen as disadvantaged by society. In the example of gender it would be the group that identifies as non-male. This disadvantaged group will be named the protected group in the definitions and the other group will be named the unprotected group. The definitions used in this section will require that both groups are treated equally. The naming of protected and unprotected groups is more for understanding the examples, but there should be no difference in treatment.

Because of the mathematical nature of the definitions, some symbols must be defined in order to write them.

- G : The group to which a sample belongs, in the examples used the groups are a and b.
- Y : The ground truth, what actually happened not the prediction. (Binary)
- d: What the AI system predicted (Binary)
- Classes/Labels: 0 is used for the negative label and 1 for the positive label
- L: Is only used in conditional statistical parity, stands for a relevant attribute that is not sensitive.
- TP: True positives, number of samples for which $Y = 1 \wedge d = 1$
- TN: True negatives, number of samples for which $Y = 0 \wedge d = 0$
- FP: False positives, number of samples for which $Y = 0 \wedge d = 1$
- TP: False negatives, number of samples for which $Y = 1 \wedge d = 0$

Important to note is that the following definitions on their own are most likely not capable of defining what is fair as defined by humans [23]. Section 3.3.3 talks about combining the different definitions in order to approach what can be seen as fair.

Statistical parity [19] [18]

Definition 3.3.1 (Statistical parity) *If a classifier satisfies statistical parity then members of both the protected and unprotected group should have the same chances of receiving the positive outcome. This means that both groups have the same Positive Rate.*

$$P(d = 1|G = a) = P(d = 1|G = b) \qquad \textit{Positive Rate (PR)} = \frac{TP + FP}{TP + FP + TN + FN}$$

The idea behind statistical parity is that everyone has the same chance of receiving the prediction of the more favourable outcome. If statistical parity would be used in a situation that determines the candidates to interview for a job, then the group of candidates at the interviews should reflect the group of applicants. In other words if 60% of all applicants were from the unprotected group and 40% were from the protected group then if 10 people are invited for an interview, 6 of them will be from the unprotected group and 4 will be from the protected group.

The first thing to notice is that statistical parity does not depend on the ground truth, it only depends on the outcome of the classifier. A random classifier would, for example, satisfy this definition. A problem with the definition is that it does not relate to the ground truth. To relate back to the interview example, it is possible that the classifier will invite people of the advantaged group who are qualified for the job (the ground truth is the positive class), but from the disadvantaged group it invites people who are not qualified for the job (the ground truth is the negative class). In that case there might have been qualified people in the disadvantaged group, who simply were not invited for an interview. If such a situation occurred, the company would hire someone from the advantaged class as they are qualified for the job. This means that the disadvantaged class was disadvantaged due to the incorrect people receiving the positive prediction.

Statistical parity does not re-enforce biases in society. It can be an advantage for people from the disadvantaged group as the positive rate could be higher than the base rate (the proportion of samples with a positive ground truth). However this lower base rate could be due to their circumstances of being in the protected class and not their capabilities. The law sees this concept of statistical parity as fair. On the other hand, it can lower the accuracy of the model as the higher positive rate means that certain samples are definitely misclassified compared to the current ground truth. If a strong case can be made why a model is fair through the use of other fairness definitions than that should be preferred to statistical parity because of the burden statistical parity puts on the accuracy.

Conditional statistical parity [19]

Definition 3.3.2 (Conditional statistical parity) *If a classifier satisfies conditional statistical parity then people in subgroups of the protected and unprotected group created on some relevant attribute, have the same chances of receiving the positive outcome. This means that smaller groups are created and the split of people from the protected and unprotected group have the same Positive Rate.*

$$P(d = 1|L = l, G = a) = P(d = 1|L = l, G = b) \qquad \textit{Positive Rate (PR)} = \frac{TP + FP}{TP + FP + TN + FN}$$

Conditional statistical parity is an extension of statistical parity. It adds the possibility of refining the groups further based on a condition. This helps obtain higher accuracy. However, it does not mean that if the conditional statistical parity is fulfilled that the statistical parity is fulfilled and vice versa. Continuing with the job interview example, subgroups could be made based on a person's degree. For example, the chances of obtaining a positive outcome when you have a master's degree can be equal across both groups, but higher than the average, whereas the subgroup of people that have no higher education, it could have a lower positive rate.

Conditional statistical parity suffers from the same problem as statistical parity in that it does not depend on the ground truth. The effect can however be lessened when correctly choosing the relevant attribute on which the condition is based. Two considerations should be made when using conditional statistical parity and choosing the relevant attribute. First of all whether the given attribute introduces some form of discrimination itself. For example for a job interview a relevant attribute might be having worked abroad for some time. It might be possible that people in the protected class did not have the opportunity to work abroad because of personal reasons. If having worked abroad is not a strong advantage for that job itself and people in certain groups did not have the same opportunities this could introduce some bias in the fairness definition, the opposite of what is desired.

The second consideration is about how representative the subgroups remain. If the data set is extremely large this might not pose a problem, but when looking at the relative sizes of fairness data sets they are often fairly small [7]. This results in a trade-off between the number of subgroups and how representative the subgroup remains. Representative in this context means how well it can reflect society and the accuracy of the statistical measures. When returning to the interview situation more coarse splits could be made on the education. The subgroups could consist of people with a master's, a bachelor's or no higher education instead of splitting into specific degrees.

Predictive parity [19] [18]

Definition 3.3.3 (Predictive parity) *If a classifier satisfies predictive parity then members of both the protected and unprotected group have equal chances of truly belonging in the positive class, when they were predicted to be in the positive class. This means that both groups have equal Positive Predictive Value.*

$$P(Y = 1|d = 1, G = a) = P(Y = 1|d = 1, G = b)$$

$$P(Y = 0|d = 1, G = a) = P(Y = 0|d = 1, G = b)$$

$$\textit{Positive Predictive Value (PPV)} = \frac{TP}{TP + FP} \qquad \textit{False Discovery Rate (FDR)} = \frac{FP}{TP + FP}$$

The definition of predictive parity depends on the Positive Predictive Value. However due to the nature of the statistical metrics, the False Discovery Rate could also be used as the condition for predictive parity. This is because $FDR = 1 - PPV$. Thus, equal chances of not belonging in the positive class when predicted to be in the positive class would have the same result as the original formulation of the definition.

Predictive parity, first of all, gives information about the correctness of the positive predictions. This means that it gives certainty to each user about the correctness of their prediction. On its own, this definition does not provide much information for the user, as it only determines how much they can trust a positive result. If someone were to get a negative prediction then not much is known about the probability of that prediction being correct. It can be assumed that both parties must have at least one sample predicted as positive; otherwise, the positive predictive value is undefined and therefore unusable.

In an interview setting, predictive parity would mean that the people invited for the interview are equally likely to be capable of the job. However, it does not give an indication about how many people were invited from each class. It is possible that significantly fewer candidates are selected from one group for the interview than the other. This characteristic of the fairness definition makes it insufficient in a situation like this. It is seen as discrimination in Belgium if fewer people from a certain group are invited for an interview even though both groups are equally capable [24]. However, combining definition with another group fairness definition, could strengthen it. For example, if predictive parity were combined with statistical parity then this problem of one group being invited less is mitigated. Also, satisfying predictive equality would mean the people who were invited to the interview equally likely to be capable, resulting in the eventual hire having equal chances of being from either group. Therefore, in an interview setting, the combination of predictive parity with statistical parity can yield strong results, with the slight misfortune of a decreased accuracy for the system most likely.

In the general list of fairness definitions the inverse of the definition that would be $P(Y = 1|d = 0, G = a) = P(Y = 1|d = 0, G = b)$ is not included. However, it will play a part in the definition of conditional use accuracy. The fact that it is not a named fairness definition does not mean it should never be used. If this constraint seems useful in some use cases, it could still be used.

Predictive equality [19]

Definition 3.3.4 (Predictive equality) *If a classifier satisfies predictive equality then members of both the protected and unprotected group have equal chances of being incorrectly predicted to be in the positive group while the ground truth is that they should be in the negative group. This means that both groups have an equal False Positive Rate.*

$$P(d = 1|Y = 0, G = a) = P(d = 1|Y = 0, G = b)$$

$$P(d = 0|Y = 0, G = a) = P(d = 0|Y = 0, G = b)$$

$$\textit{False Positive Rate (FPR)} = \frac{FP}{TN + FP} \qquad \textit{True Negative Rate (TNR)} = \frac{TN}{TN + FP}$$

The definition used here for predictive equality uses the False Positive Rate as the statistical measure that should be equal. However, if the FPR is equal then the True Negative Rate will also be equal, because $FPR = 1 - TNR$. So the definition can be used with either one of the statistical properties.

In its definition, predictive equality uses false positives (incorrectly predicting someone in the positive group, while they should be in the negative group) in its definition. In most scenarios, false positives are more desirable and less impactful for the user's future than false negatives. If that is the case, then this definition does not provide much information about the system on its own as it does not include the more crucial decision about the people who belong in the positive category. In the interview situation this means that the chance of getting invited to an interview if you are not qualified is equal. This is not very advantageous for either party, but the knowledge of these odds also does not tell a lot about the fairness of the predictor. However, additional information van by very useful when the predictor also satisfies other fairness definitions.

In the case a true positive is the worst possible incorrect prediction like, for example, in a security situation where you do not want someone to be falsely cleared. In that case, this definition shows that there is no discrimination. If, for example, the false positive rate was higher for white people versus

people of African American descent then there is an argument to be made that the AI is racist towards the African Americans.

Equal opportunity [19]

Definition 3.3.5 (Equal opportunity) *If a classifier satisfies equal opportunity then members of both the protected and unprotected group have equal chances of being incorrectly predicted in the negative group while the ground truth is that they belong in the positive group. This means that both groups have an equal False Negative Rate.*

$$P(d = 0|Y = 1, G = a) = P(d = 0|Y = 1, G = b)$$

$$P(d = 1|Y = 1, G = a) = P(d = 1|Y = 1, G = b)$$

$$\textit{False Negative Rate (FNR)} = \frac{FN}{TP + FN} \qquad \textit{True Positive Rate (TPR)} = \frac{TP}{TP + FN}$$

The definition used in this case for equal opportunity uses the False Negative Rate to express the statistical properties. Another possibility would have been to use the True Positive Rate as $FNR = 1 - TPR$, making them equivalent for determining whether this definition is satisfied.

The definition of equal opportunity is similar to the definition of predictive equality. Both concern the fraction of correct/incorrect predictions given the actual class. The main difference is that equal opportunity is about false negatives and predictive equality is about false positives. So the most relevant definition of fairness between the two of them depends on what scenario is worse for the implementer and the eventual user. The same elements discussed in the section on predictive equality also count here. Namely, it depends on how detrimental false negatives are for the system in order to evaluate how useful this fairness metric is and if it would be more valuable if used together with other definitions.

Equalised odds [19] [18]

Definition 3.3.6 (Equalised odds) *If a classifier satisfies equalised odds, then it both satisfies predictive equality and equal opportunity (the two definitions above). This means that both groups have equal True Positive Rates and True Negative Rates.*

$$P(d = 0|Y = i, G = a) = P(d = 0|Y = i, G = b), i \in \{0, 1\}$$

$$P(d = 1|Y = i, G = a) = P(d = 1|Y = i, G = b), i \in \{0, 1\}$$

Equalised odds is a fairly strong fairness definition as it is about the incorrect classification rates within each class. This means that the chances of someone being incorrectly classified are known, and those chances are equal for both groups. A trivial example of a classifier that would satisfy equalised odds is one who would classify every input as the positive class (or the negative class). This would satisfy the condition as $P(d = 1|Y = i, G = x) = 1$, but would of course not be a very performant classifier. It is also important to note that equalised odds does not mean equal accuracy; in order to have equal accuracy the base rates between both groups should also be equal.

One way to look at equalised odds is as a sort of statistical parity but then defined on subgroups instead of the entire group. With these subgroups split based on the true group the people should belong in, this would take away the biggest problems with statistical parity, namely that it hurts the accuracy because it is so strict. Equalised odds differ from conditional statistical parity because the subgroups are not split on a relevant feature like in statistical parity. It splits on the ground truth which is not a feature known to the system when it is being used.

Conditional use accuracy equality [19]

Definition 3.3.7 (Conditional use accuracy equality) *If a classifier satisfies conditional use accuracy equality then it satisfies both predictive parity and the inverse of predictive parity (chances of truly belonging in the negative class, when being predicted in the positive class). This means that both groups will have equal Positive Predictive Value and Negative Predictive Value.*

$$P(Y = 1|d = 1, G = a) = P(Y = 1|d = 1, G = b)$$

$$P(Y = 0|d = 0, G = a) = P(Y = 0|d = 0, G = b)$$

$$\textit{Positive Predictive Value (PPV)} = \frac{TP}{TP + FP} \qquad \textit{Negative Predictive Value (NPV)} = \frac{TN}{TN + FN}$$

$$\textit{False Discovery Rate (FDR)} = \frac{FP}{TP + FP} \qquad \textit{False Omission Rate (FOR)} = \frac{FN}{TN + FN}$$

The definition uses the statistical measures of positive predictive value and negative predictive value. False discovery rate (FDR) and false omission rate (FOR) could be used in their stead as FDR = 1 - PPV and FOR = 1 - NPV, meaning that if one side of the equation were to be equal across the groups, then the other side would also be equal. The choice is up to the developer regarding which values they use to determine whether the definition is satisfied.

The definition of conditional use accuracy equality might seem very similar to equalised odds, but the difference lies in some of the details. While equalised odds keeps the same balance with regard to the

ground truth, conditional use accuracy equality keeps the balance given a certain prediction. If the classifier satisfies conditional use accuracy equality then the user knows what their chances are that this prediction is correct regardless of the group to which they belong.

Using a classifier that would always return the positive class 1 as the prediction will not work for conditional use accuracy because the positive predictive value will be dependant on the base rate and that might not be equal across groups. Therefore, it is possible to satisfy equalised odds but not conditional use accuracy equality.

Conditional use accuracy equality is simpler for the user to understand and to draw conclusions about their own situation. When they get their results they can be sure about the likelihood that the result is correct. The interpretability for an individual is not present in equalised odds, as it depends on the ground truth, which is not known to a user otherwise the AI system would serve no purpose.

Overall accuracy equality [19] [18]

Definition 3.3.8 (Overall accuracy equality) *If a classifier satisfies overall accuracy equality then both the protected and unprotected group have equal chances of receiving the correct prediction regardless of the class they actually belong to. This means that the accuracy for both groups is equal.*

$$P(d = Y|G = a) = P(d = Y|G = b) \qquad accuracy \ = \frac{TP + TN}{TP + FP + TN + FN}$$

This is definition tends to be currently used to determine if an AI system is fair. An example of where this was used is in the COMPAS case explained in the introduction [1]. However using this definition as the only fairness definition to satisfy is only possible if there is no preference between possible classifications. For example, if the AI should simply determine if you look best in round or square glasses, then it would be equally bad if it would mistake round glasses as the better option while it should be square glasses or vice versa. However, if there is a preferable group to be predicted in, like for a job interview, then it is a lot worse for the user to not be invited when they are qualified than the other way around. This does not mean that the "lesser evil" option would not affect the user; rather, the extent of the effect is significantly less.

The goal in classification is often to maximise the accuracy. If it were possible to reach 100% accuracy then the model would be fair; it would make no mistakes, given that the ground truth is the reality. This is, however, realistically impossible to achieve in nearly all applications of AI systems. One could

argue that maximising the accuracy increases the fairness of the system, but this is both legally dubious and often incorrect when there are underrepresented groups within the population in which the AI would be used. The incorrect predictions for these people in the underrepresented groups would only contribute to a small fraction to the eventual global accuracy, making it possible for an AI system to maximise its accuracy by simply focusing on optimising its performance for the samples from the majority group.

The definitions of equalised odds and conditional use accuracy equality might seem similar to overall accuracy equality; however, satisfying one of these conditions does not mean satisfying the other. This can be deduced using the Bayes rule. In order for equalised odds and overall accuracy equality to be satisfied, then both groups need the same base rates. Meaning that if two groups have the same base rate satisfying equalised odds is the same as satisfying overall accuracy equality. Satisfying both conditional use accuracy equality and overall accuracy equality requires that the model has equal positive rates for both groups, which, unlike equal base rates, is something that could be tuned. However, equalling the positive rates often comes at the cost of accuracy as was discussed in the section on statistical parity (Definition 3.3.2).

Treatment equality [19]

Definition 3.3.9 (Treatment equality) *If a classifier satisfies treatment equality then the ratio of false negatives to false positives is equal between the protected and unprotected group.*

$$\frac{FN_a}{FP_a} = \frac{FN_b}{FP_b}$$

Treatment equality is the only fairness definition in this dissertation which cannot be expressed as chances. Therefore this characteristic is more vulnerable to different base rates between both groups and should be something for which to watch out. If the fraction used in the definition would be higher for the unprotected group than for the protected group, then there are relatively more false negatives than false positives in the unprotected group than there are false negatives relative to false positives in the protected group. Some information can be gained from this definition but should be handled with care and especially with a good knowledge of the data distributions before using. The discussion of it was done in Chapter 5.

3.3.3 Combining Fairness definitions

While it would be possible to combine the fairness conditions from sections 3.3.1 and 3.3.2, certain combination are better to make than others. For example, the similarity-based metric from fairness through

awareness already encompasses the entire system, partly because it is so strict. Therefore, it would not be necessary to combine it with other definitions. However, most of the other definitions are not very strict in nature and thus it is beneficial to try and use them together.

Important to note is that not all group fairness definitions can always be combined [25]. In the Section on Overall Accuracy Equality (Definition 3.3.2), this was already brought up when trying to satisfy overall accuracy together with equalised odds or together with conditional use accuracy. In order to satisfy the combination, certain properties are required of the data or the model. While the model is somewhat controllable, the base rates are characteristics from the data set and should not be altered, unless this inequality does not reflect the real world. One special case is where the base rate would not matter if the system would achieve perfect accuracy. In this case, the equality of $P(Y = x|d = x) = P(d = x|Y = x) = 1$ holds making that all fairness definitions will be satisfied. This also means that there are no false positives or false negatives.

When trying to combine fairness definitions, it is important to check if the probabilities will work out and what constraints might be put on the system as a result of it. As many combinations of fairness definitions are possible, it would not be possible to give a complete list. The user can try and calculate it themselves, mainly through the use of Bayes' rule $P(A|B) = \frac{P(A \cap B)}{P(B)}$ if $P(B) \neq 0$ and with the knowledge that $P(Y = 1|G = a)$ equals the base rate of group a and that $P(d = 1|G = a)$ equals the positive rate of the model for group a.

3.4 Fairness on the SIMON-test

In the case of the SIMON-test the prediction whether a student succeeds equals whether they receive their bachelor's degree within four years after starting the degree. However for the notion of succeeding a myriad of different definitions could be used. The first element would be determining what a student should achieve to succeed, is this completing the first year, the bachelor's degree or the master's degree. The next choice would be in what time frame the result should be achieved. For example in the first year does it require passing all the exams at the first attempt or is it having passed all courses by the end of the academic year. For the bachelor's and master's degrees the question can be raised if this should be through getting the degree eventually, completing the standard study path (three years for a bachelors) or could this be flexible to a certain extent. The definition of what success looks like will affect fairness itself. For example, students with a disability often take slightly longer to achieve their bachelor's degree [26].

It is also important to analyse the effects of the different results that the AI system for the SIMON-test might return. This will be observed from the student's standpoint as they are the most affected by the AI system. For a student, the worst possible result would be a false negative, receiving that they would not succeed while, in reality, they would have. A false positive, predicting a student would succeed while in reality they will not, is less detrimental to them. Currently, it is not unheard of for a student to reorient towards a different degree, and that is what most likely would happen when a student is given a false positive. Because a false negative would result in reducing the options of a student for their future it is the most crucial case on which the fairness definition should focus.

3.4.1 Individual fairness

Fairness through unawareness Fairness through unawareness is often seen as the baseline of fairness. It prevents causal discrimination, which is legally the clearest form of discrimination. However, this will not guarantee that the system does not discriminate. Due to the nature of the data in the SIMON-test it seems likely that a certain sensitive attribute influences relevant features for making the prediction. While there would be no causal discrimination in the system, another form of discrimination is not impossible to occur. It seems insufficient to use fairness through unawareness in the SIMON-test and deem it fair without any further investigation.

Suppression Suppression was briefly discussed as an extension of fairness through unawareness. It either removes the highly correlated features or makes changes to them. Removing the highly correlated features from the SIMON-test does not seem like a good decision as the features are all closely related to the individual. This could mean that removing the highly correlated features would either mean removing too many features from the data set or setting a very high correlation threshold. This very high correlation threshold might remove some correlation but the intention of removing all significant correlation will not occur. The possibility of transforming the data is possible, but will be highly complicated. This combined with the limited data set it will be very difficult to generate a proper result. If a strong notion of fairness could be achieved through the other definitions then that option should be preferred.

Fairness through awareness As was discussed in subsection 3.3.1, it is difficult to achieve fairness through awareness. Because the SIMON-test is backed by people with a strong knowledge about education and students it is not unfeasible that a correct distance metric could be devised. Or at least that no use of an incorrect distance would occur. The greatest difficulty in achieving fairness through awareness in this situation would be the implementation of an algorithm. While this could be very much possible it is

not widely tested and is still raising several questions. Therefore fairness through awareness might be an interesting aspect to research in the future but does not seem like the best solution at the current time.

3.4.2 Group fairness

Statistical parity Statistical parity on its own is a sufficient characteristic to determine that there is no discrimination in the system, in the sense that no legal liability is present if statistical parity is satisfied. However in general statistical parity is not very desirable to achieve. It often comes paired with compromising on the accuracy of the system, the opposite of what is desired in an AI system. If there were only small differences between the base rates of the sensitive and non-sensitive groups then this loss in accuracy could be minimal. However for the sensitive group of students who come from a lower socio-economic background the numbers often reflect a lower chance of succeeding than student from a higher socio-economic background [27, p. 38]. This difference in base rates indicates that using statistical parity might require a fairly high compromise with overall accuracy.

Conditional statistical parity For conditional statistical parity it might be difficult to find a feature on which the different student groups can be divided. The best option seems to be on the hours of mathematics a student had in high school. This seems like a logical attribute to split the students as it ascertains their prior mathematical knowledge and interest. However the question must be raised if splitting on this feature does not introduce some bias of itself. Not all schools provide the possibility of following for example eight hours of mathematics. This means that the school a student attended can influence the group they will be split into. This presents a hardship as often the school a student attended is linked to a student's personal situation. This decision can probably be defended on legal basis, but morally it could raise some more questions.

Predictive parity Predictive parity in this context means that when a student receives a positive prediction of the system, irrespective of their sensitive attribute, they have the same chance of this prediction to be correct. In the context of choosing a profession this translates that the same ratio of the students across groups who received a positive prediction will be able to go into the profession. It is clear that this is not a definition which encapsulates the goal of this system. The main importance is assisting students with their choice in higher education, this definition gives very little reassurance to the student that they can trust the system. They get reassurance if they receive a positive prediction, but they know very little if they receive a negative prediction. Therefore while this definition can create some benefit for the system, but it does not encapsulate its goal and the tied fairness concept of it.

Predictive equality Satisfying predictive equality provide some benefit to the system. In the setting of the SIMON-test it makes sure that one group will not be overestimated compared to another. This introduces some sense of fairness into the system, but false positives are not the worst thing to happen for a student. False negatives affect students a lot more, and these are not included in the definition of predictive equality. If the goal of the application would be used in an admission's setting as discussed in the introduction, then the university might be more interested in false positives as they are not beneficial to them. Alternatively maximising the accuracy would reflect their interests more. This makes predictive equality a reasonable fairness definition to use, more so in an application used in an admissions settings than for a supportive tool such as the SIMON-test. The definition does however not encapsulate certain situations which are more crucial to control.

Equal opportunity Equal opportunity is one of the more crucial fairness definitions in the context of the SIMON-test. If the AI system satisfies equal opportunity then students from different groups are equally likely to get a false negative, which is the worst possible situation. It is best to try and avoid false negatives as much as possible, but as most AI systems are not perfect it is not practically feasible to prevent them completely, aside from a trivial system which only returns positive predictions. Equalling the chances of a false negative across the groups makes that they have equal opportunity (like the name of the definition) to start their education. A higher false negative rate for one group would result in less opportunities given to them, which is the concept of discrimination. Thus equal opportunity is an important fairness definition to satisfy in the case for the SIMON-test.

Equalised odds As equalised odds combine the two definitions above this section and equal opportunity is on its own a good fairness definition, that makes equalised odds definitely a good fairness definition for the SIMON-test. It must be noted that the stronger/restrictive a fairness definition is, the harder it could be to achieve. So if an AI system would satisfy equalised odds it would be a great feat, but the decision can also be made that equal opportunity is sufficient.

Conditional use accuracy equality Conditional use accuracy equality mainly equalises the chances that a given prediction is correct. In a situation like the SIMON-test, this useful information. In a situation where the base rates are balanced across groups, then the value of the definition increases. However in this situation the only added value that can be seen is that a student knows they are just as likely as anyone else in their year to succeed. Therefore it does not seem crucial for a system to satisfy conditional use accuracy equality.

Overall accuracy equality While overall accuracy equality seems like a good fairness definition to use, it has many pitfalls when used as the sole fairness definition. The main reason overall accuracy equality is not a desirable definition on its own, is the large difference between the effects of a false positive on a student's life versus a false negative. Accuracy could be equal across groups but where one group could have significantly more false negatives than the other group, leading them to be disadvantaged. In the situation of the SIMON-test it could also be difficult to achieve equal accuracy across groups. The fairness guarantee provided with the overall accuracy equality is less compared to the equal opportunity fairness definition. Large discrepancies in accuracy across groups will evidently not be beneficial, as the incorrectly predicted students would not obtain the degree. They either would not start (false negative) or not succeed (false positive). Overall accuracy equality is important to a lesser extent than for example equal opportunity, but is still beneficial to satisfy (with perhaps a larger error margin between groups) for the SIMON-test.

Treatment equality It is difficult to determine if treatment equality would work in this situation. In order to evaluate the usability of treatment equality some notion of the proportions within the data is important. If the data set is large enough and the base rates between the groups are similar enough then treatment equality could be a good fairness measures. It would achieve the similar effect of equal opportunity, where one group would not necessarily be more disadvantaged. When using treatment equality it is beneficial to include overall accuracy equality as an additional definition to satisfies in order to prevent that fraction could otherwise get out of hand.

3.4.3 Recommendation

The simplest definition to hold an AI system to is causal discrimination. This can be seen as one of the benchmarks for fairness. However a combination of different fairness metrics could achieve a more fair system that might break causal discrimination, but not necessarily would. An added condition would be that causal discrimination is only broken when the differing predictions would be correct. This is however difficult to ensure in every situation and should be done with caution when chosen instead of normal causal discrimination.

One fairness definition that seems vital for the SIMON-test is equal opportunity. If the SIMON-test satisfies this definition and has fairly similar accuracy across the groups than it has a justifiable case of stating that it is a fair application. Its fairness could even be improved upon if it were to satisfy equalised odds. Of course it is still for the developer to determine which margins in relation to the data are acceptable between metrics to classify the statistical measures between two groups as equal.

Chapter 4

Biases in the data set

Data sets are the backbone of every Machine Learning application, a phrase that has become common in the setting of Computer Science is *Garbage in, Garbage out*. This means that a computer program could use a really great algorithm, but if it were to get bogus inputs, it will not be able to generate the desired outputs. In the context of Machine Learning this translates to the data set not containing the appropriate information in order to achieve the goal set out to fulfil. A simple example to imagine is if a car manufacturer wants to make self-driving cars, who of course need to be able to identify other cars on the road. If there would only be sport model cars in the data set, then there is a large chance that the car will only be able to recognise sport model cars, instead of all cars as it was intended.

The following section will introduce the different types of biases that can arise in a data set and cause undesirable behaviour in the form of discrimination. The types will be ordered firstly by category and then by likelihood. Each type will be discussed with its definition, an example and if the bias would be present in the SIMON-test. After that section the analyses of these different biases are made for the OULAD data set.

4.1 Types of Biases

The bias types discussed in this section comes from the survey conducted by Mehrabi et al. [6]. Before this discussion can be done properly the collection process of data itself needs to be outlined. The discussion is supposed to help the reader to possibly draw a parallel to their own data sets and be mindful of what biases may be present in them. As will be discussed, the type of bias present could be dependent on the data but also on the factors around it.

Important to note for all types of biases is that the differences need to be systematic. Meaning that the difference is between two groups, who could be identified in that they share a certain sensitive attribute. If the different circumstances are random between the people in the groups then it will be noise that is added to the system rather than bias.

In order to understand where the bias might arise from, some general concepts about the collection of the data needs to be understood. This dissertation focuses on *Fairness in AI* with regard to fair treatment of people, so the data collection will in some degree always have a human element to it.

In order to encompass all of the possible biases that might arise, the usage of the data is also incorporated in the discussion. This creates three elements in the data collection process, the user interaction, the data as a whole and the algorithm used. As can be seen on figure 4.1 there is a connection between the different stages creating a feedback loop throughout.

The first stage of data collection in the context of this dissertation is some form of user interaction. User interaction can be seen as the source of the data. It can be a user consciously providing data to the system, by some sort of questionnaire, or unconsciously by simply surfing on a website. It is immediately clear that there is a large human factor in this stage that should be taken into account.

The second part is the data itself and is also the main subject of data analysis, meaning learning the characteristics of the data. This can go from general characteristics such as the time when the data set was collected, such as 50 years ago or one year ago, or it could be about the distribution of the data collected, such as the gender distribution of the people in the data set. A full data set analysis will be done further in this chapter.

The final element is the algorithm. This algorithm can be the algorithm that will be designed, but it is also the algorithm in play when the data is being collected. That algorithm can be seen as the factors that influences the data a user creates. This could be because it already takes into account people's preferences or simply the format which will be more or less accessible by different groups of people.

The data of the SIMON-test itself is not available therefore following analysis added to the bias types is hypothetical. However most types of biases require a somewhat theoretical analysis.

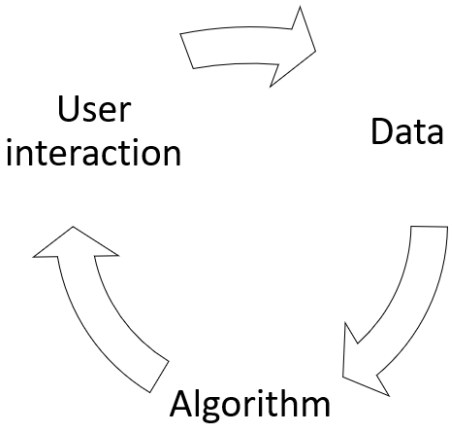


Figure 4.1: The feedback loop between different stages relevant to data collection

4.1.1 Data bias through user interaction

The SIMON-test works mostly with user input in the form of tests rather than just actions taken on the platform. This increases the chances of bias through user interaction as the data is more dependant on the user.

Definition 4.1.1 (Temporal bias) *Systematic distortions across groups or behaviours over time [28]*

Distortions defined as temporal bias can have many causes. One possibility is that the platform on which the data is being collected changed somewhat during the collection of the data. This can influence the data and create distortions between the newer and the older data. For example Twitter has changed quite a bit over the years and so have the people that use the platform [29]. This means that the data of people who used the platform more in Twitter's earlier days, could vary significantly from people who are active on it more recently, simply because of temporal bias and not because of fundamental differences between the users. Temporal bias can also occur because of changes in the mentality of the population itself, this can be seen for the profession of OB-GYN. Recently the expectation shifted from them being male physicians to becoming a predominant female field [30]. Also the work ethics within the profession itself shifted in recent years, indicating another form of a difference in mentality across time.

Temporal bias could also be related to something seasonal. The question here is if there is a possibility that the data gathered would have been different if it was done 6 months later in the year. In shopping behaviour for example, this could be an interesting factor on what someone buys and how much they tend to spend. The last possibility of temporal bias discussed is the relative time frame for the user. In the example of gathering data on how a student perceive their performance on a test, then the moment respective to when they took/will take the test has an effect on their answer [31]. So in order to

avoid this origin of bias in the data it is important that the relative time should be equal across groups. This could still be at different moments in time but those different moments should be relatively equal for all groups compared to when the test was/is.

If temporal bias is clearly present in a data set then that is a real basis for liability of the designer. Using an older data set is something the designer should be wary of. If an older data set is used then there should be some thorough analysis in order to determine which biases might be present.

SIMON-test Temporal bias can only occur if some large differences in behaviour or groups exist over time. The SIMON-test was taken into use in 2017 [27]. Due to its limited age and that the characteristics of a (prospective) student do not change that quickly, the chances for temporal bias seems very low at the time this dissertation is written.

Definition 4.1.2 (Historical bias) *Bias introduced by the world as it is or as it was [32]*

Historical bias arises through the biases which are present in society, but aren't desirable. So it is very probable that these biases are present in the data and it can be up to the AI to compensate for these biases. Natural Language Processing is a field in which this appears easily. One example is the association of a gender to a profession, for example nurse is associated with female and engineer with male. There is no valid reason to assume a nurse would be a female, but the current way our society works it is probable. This creates a bias that nurses are assumed to be female.

Historical bias is something that cannot be avoided as it is just a part of society. Depending on the situation these biases can be troublesome. Take for example an AI system that suggests possible careers to people. In this situation historical bias could be acceptable in the sense that male student tend to get the recommendation of engineer more than female students and female students get the recommendation of nurse more often. The key element in this situation is that it is not impossible for a female person to get the recommendation of engineer. If the algorithm merely mimics society then it could be accepted in a legal concept, if it however were to exacerbate these biases then that could become a problem. If the system was however to determine how likely someone is to succeed for example as an engineer then there should be no differences based on gender.

SIMON-test The SIMON-test contains certain non-cognitive personal tests. This opens the possibility for some historical bias to be introduced in the data. This warrants to look at the score distributions between different groups in order to determine if some systematic distortions have an effect on the workings of the AI system.

Definition 4.1.3 (Behavioural bias) *Systematic distortions in user behaviour across platforms or contexts, or across the people represented in different data sets. [28]*

The context in which the data is collected affects the behaviour of the user. This means that if data is collected through multiple sources, some randomness also called noise could be added to the data. In the case that one group of users is collected through one medium/platform, and the information of a different group is collected through another than these can create significant discrepancies in their data. These discrepancies are also called bias as they are correlated to a group of people.

An example of how behavioural bias might occur can be found in the research paper of Miller et al. [33]. They found that people's interpretation of emoji's across different platforms could also vary significantly. However collecting data across different platforms will not always lead to behavioural bias. Research in human behaviour has looked into behavioural bias for different survey styles and could often not find any significant differences between the methods [34].

SIMON-test The data for the SIMON-test is solely collected through its own platform. Consequently there can be no differences across platforms as there is only one platform. It is possible that prospective students fill them in different contexts, like in a classroom or at home alone which could affect their behaviour. In order to determine if there could be some behavioural bias, a survey could be devised which poses the question in what context the prospective students filled out the SIMON-test. This is one also a type of bias that needs to be checked regularly, as society and the application undergo changes it is possible for behavioural bias to suddenly occur.

Definition 4.1.4 (Content Production Bias) *Lexical syntactic, lexical, semantic, and structural differences in the content generated by people. [28]*

Content production bias encompasses the differences in language behaviour of the population. This could propose a problem when using unstructured text inside of a data set. Interesting research is being done in the field of NLP regarding lexical differences. Currently most systems within the NLP community are trained on Standard American English, however that variant of English is not is spoken by everyone else [35]. For example nearly 80% of all African Americans in the United States speak African American Vernagular English. When testing Natural Language Understanding models against these two forms of dialect it turned out that they systematically performed slightly worse for the African American Vernagular English.

Creating an NLP application such as a chatbot that is sensitive to certain dialects will also become

sensitive to certain groups of people as dialect is often correlated with demographics. For example creating a chatbot that does not work as well for people with a certain dialect will result in discriminating against those people as the model won't be able to provide them the same level of service.

SIMON-test The content generated by the student does not contain user generated text. The content produced is a combination of test results, so no content production bias can occur.

Definition 4.1.5 (Self-selection bias) *A subtype of selection or sampling bias in which users select themselves. [6]*

Self-selection bias occurs when the user is actively generating data for the data set, rather than passively. In self-selection bias the bias is introduced because the user creates some selection bias on their own. This is often due to unclear requirements for the users. In computer systems this is often a rather rare type of bias. An example could be if a study would ask hard workers to fill in a survey. What defines a hard worker is difficult to formulate and no definition would be homogeneous across the population. So it is possible that people who are not considered hard workers by the researcher's criteria to still fill in the survey and increase the noise in the data.

SIMON-test Participation in the SIMON-test is on a voluntary basis. There is however no filter criteria to participate, so there is most likely no self-selection bias. A slight form of self-selection bias can occur in that only people who see it as a possibility to follow higher education will fill in the SIMON-test. This most likely will not have a huge effect on the data, but it is good to keep this in mind.

4.1.2 Data bias in the data set

Bias in the data set is harder to analyse for the SIMON-test as the data set is not available. The following section cannot ascertain the following biases, but mentions certain aspects which should be investigated within the data set.

Definition 4.1.6 (Representation/Population Bias) *Systematic distortions in demographics or other user characteristics between the population represented in the data set and the use population. [28] [32]*

Population bias is one of the more obvious data bias. When thoroughly exploring a data set it becomes clear if there are imbalances present which can cause bias. It is important to acknowledge that there must be a definition of what the target population is and what its characteristics are. The first step is thus knowing who the use population of the AI system will be, and not work on an unfounded expectation. After that it is important to identify underrepresented groups in your data set. An example of often

underrepresented people are illiterate people in Western countries. When certain groups of people are underrepresented in the data, it can lead to the AI and other data processing techniques to not take them into account fairly.

A common cause of representation bias is also called sampling bias in statistics. Sampling bias occurs when the data set does not represent the full truth of the objective you are trying to fulfil. The Corona crisis was a good example of this, at a certain point only seriously ill people could get tested. This caused the statistics to skew as they didn't actually represent the people who had the disease, but only the people who had the disease and were seriously ill. So sampling bias arises due to how the data was collected and can be found directly in the distribution of the data set. An example of sampling bias which has real-life effects is in the accuracy of pulse oximeters. Pulse oximeters measure the amount of oxygen in the blood, this is important if someone is gravely ill or under sedation. Recent studies have shown that the error rate of these devices are much higher for people with a darker skin tone [36]. This is likely due to the calibration being mainly done on white people. A rarer case of sampling bias could be that a bad split is made in the data for training versus testing, but that would mainly be an error to the technique of creating an AI and less a socio-technical one.

Representation bias although very common is also something serious and is directly the responsibility of the designer to try and correct it. If certain groups are not well represented then it can be very difficult for the model to perform well on them. If a model underperforms for a certain group of people in society then that can be seen as discrimination. The importance of using a representative data set is also included in the EU proposal [4, §44], where it says that *They (the training, validation and test set) should also have the appropriate statistical properties, including as regards the persons or groups of persons on which the high-risk AI system is intended to be used.*

SIMON-test Representation bias can occur more easily in the data set because the training data contains only entries from people who actually started the degree. In other words the population represented are the students who start the degree. This can be different from the use population which are the students who just finished high school and are still deciding what degree to follow. An example of this can be found in engineering degrees. Only a small percentage of engineering students are female, leading to a predominantly male population represented in the data set. This is very different from the demographic of students who finished high school where around 50% are female. This is an example of representation bias in the data, but other characteristics or degrees could have other distortions. Knowing where these distortions could be present requires a thorough analysis of the data set.

Definition 4.1.7 (Measurement bias) *Systematic distortions when choosing, collecting, or computing features and labels. [32]*

The first place where measurement bias can occur is when constructing the data itself. It occurs when the features are too much an oversimplification of the concept that the system is trying to capture. It is possible that there is not sufficient information remaining for the AI to learn the ground truth. An example for this is in college success chances, only looking at the student's final grade in high school does not capture the entire picture. Other relevant characteristics can be extra curricular activities, the main subjects they studied, the school they attended and their home situation.

Another source of measurement bias is when the method of measurement varies across groups. This means that checking for a certain property is done in different ways. It could be more frequent, more thorough or simply through a different method. An example of this type of measurement bias was already presented in section 3.1, where it was discussed how the ground truth whether a car should be pulled over was distorted. This was because they were more likely to search the car when the driver was African American, making it more likely to find something. The last form of measurement bias is when the accuracy of measurement varies across groups. The COMPAS example from the Introduction (Chapter 1) is one place where this likely occurred. Due to the bias in previous decisions of judgements, African American offenders were more likely to be labelled as high-risk than white offenders [1].

SIMON-test Not enough information is present to predict if measurement bias would be present. One possible point where measurement bias could arise is in the scoring of the tests. This requires analysing the aggregation made in these tests and if they could introduce some bias into the data.

Definition 4.1.8 (Linking bias) *Differences in the attributes of networks obtained from people's connections, interactions or activity. [28]*

This is mainly relevant in data which can be structured as a graph. Linking bias could be present if there is a difference in the way the links are constructed. This could be due to user characteristics, like certain groups of the population create a different close-by network than others [37]. It depends on the use of the data if this would introduce a bias or not, it is thus necessary to be aware of these differences and take them into account in the design process. Another possibility of bias appearing among the links can be due to linkage errors, these are either missing links or extra links which should not be present. While this can often be just added noise in the data set, if these linkage errors occur because of a specific reason it might become bias in the data set [38].

The possibility of liability arises in this situation if these linkage errors were caused by human error. These things can be easily missed and checking for linkage errors is an important step as is trying to mitigate them afterwards. If it is not possible to mitigate these problems then extra steps need to be taken when checking the models behaviour for the groups of people where the linking did not occur correctly.

SIMON-test There is likely no linking bias in the data. Some links between a student's result and their SIMON-test score could be missing, making that student missing from the training data. However there is no reason to believe that these missing links would be systematic or at least not to a degree that would significantly influence the data.

4.1.3 Data bias due to the algorithm

The SIMON-test does not fall within a very complex environment, therefore it is not likely for the algorithm to introduce a significant amount of bias when the data is collected. Theoretically it could be possible that the algorithm designed to predict the scores of the students could introduce bias. This algorithm is not available and therefore it is not possible to form any conclusions.

Definition 4.1.9 (Learning/Algorithmic bias) *Amplifying performance disparities across different samples in the data. [32]*

Learning bias is the algorithm itself adding bias to the data set. This could be amplifying the bias already present or introducing it. The algorithm can be unaware to the objective of creating a fair system, especially if no measures are taken to achieve that it tries to act fairly. If this is the case then the algorithm simply tries to optimise the metric given to it and in some cases this can lead to increasing disparities. For example certain edge cases could be identified by humans and handled appropriately, while the algorithm does not. This could mean that the algorithm could make more biased decisions than a human, which is the opposite goal of a fair AI system.

Learning or algorithmic bias should be avoided at all cost. If the AI system itself were to increase the disparate treatment of people then it should never be in use. Using such a system is directly discriminating against people, which is of course illegal.

SIMON-test Algorithmic bias must be investigated when creating the AI application. This cannot be done on a theoretical basis.

Definition 4.1.10 (Evaluation bias) *Introducing bias into the system by benchmarking the model against a non-representative data set. [32]*

This bias is introduced in the development phase. Even if models are trained on perfectly representative data, they are often compared to one another using test sets. A test set can be a general benchmark data set. These benchmark data sets are used to ensure proper comparisons between models and often also to spare data, in order that all data collected can be used for training and validation. A problem arises when the test set does not capture the same use population as the AI will encounter. Based on the test scores the best model will be chosen. But if the test score is not capturing the full picture then the model which has the highest test scores might not have the best result in the specific use case.

An example of this problem was seen with AI systems that uses a person's face as input. There are quite a lot of data sets with images of people's faces, but are often racially biased. Most of these data sets contain a lot more white males than others groups from the population [39]. Based on the test scores of the system on those data sets it is more likely to choose a model that performs strongly on white males, but not necessarily as well on other people. Leading to possibly the same result as learning bias, creating systems that should not be in use and were it might not even be known that it discriminates as the test set would not indicate it.

As already mentioned in representation bias (Definition 4.1.6) the EU proposal regarding high risk AI applications specifically mentions that the test set must be representative in order to ensure bias monitoring, detection and correction in relation to high-risk AI systems [4, §44].

SIMON-test It is not likely that evaluation bias will be present in the SIMON-test. The AI system needs a specific set of features which are highly unlikely to be found in an unrelated data set. It is important that from their own data a representative test set is constructed.

Definition 4.1.11 (Aggregation bias) *The incorrect assumption that mapping from inputs to labels is consistent across subsets of the data. [32]*

Aggregation bias occurs when one model is used for everyone, but there are underlying groups where using that general model results in aberrant behaviour. This is a very complex type of bias. The best way to avoid this type of bias is including an expert about the domain in which the AI would function in the design. An example of aggregation bias is when GPT-3 [40] would be used in a legal documents framework. GPT-3 is a language model with very strong capabilities. However a general language model can underperform in a more specialised setting, which can be obscure to a layman in the field.

The main problem with the aggregation bias is the lacking performance of the model in the other situation. From a business standpoint cutting these corners will most likely result in a lost investment.

SIMON-test Aggregation bias must be investigated when creating the AI application. This cannot be done on a theoretical basis.

Definition 4.1.12 (Deployment bias) *Bias due to a mismatch between the problem intended to solve and the way it is actually used in practise. [32]*

Often models are created intended to be used in a certain way, but once released into the world start leading a life of their own. Currently quite a lot of AI tools are developed in order to support the human decision making process, providing support rather than making decisions. However a real danger exists that the person who receives the recommendation is more likely to agree with the recommendation instead of critically taking it into their decision process. This is due to some social phenomena such as automation or conformation bias.

Deployment bias is also something that is included in the EU proposal [4, §58], where it states *Users should in particular use high-risk AI systems in accordance with the instructions of use and certain other obligations should be provided for with regard to monitoring of the functioning of the AI systems and with regard to record-keeping, as appropriate.*. This also means that the developer of the system needs to determine a set of rules in order to prevent deployment bias.

SIMON-test The SIMON-test is currently in use and no clear deployment bias can be identified. A possibility for deployment bias could be that the SIMON-test is meant to stimulate a student to use additional resources, such as summer school or study support in order to improve their chances of success. But if a student were to take their result as a deterrent then the application is used differently then intended. This would mainly create problems as the fairness definition is dependent on the spirit of the AI, so the use should match between the designer and the user.

Definition 4.1.13 (Simpson's paradox) *The effect that occurs when the marginal association between the same two categorical variables is qualitatively different from the partial association between the same two variables after controlling for one or more variables. [41] [6]*

The Simpson's paradox is the effect that a certain statistical metric can vary heavily within the same data set simply by splitting the data set in different categories. In section 3.3.2, the definition was given of conditional statistical parity. The idea behind conditional statistical parity is closely related to

the Simpson's paradox as they both distinguish different groups on relevant features. There is a rather popular real life example of Simpson's paradox, a law suit at UC Berkeley over the admission numbers of female versus male applicants [41]. The law suit was about the overall lower admission rates for female applicants compared to male applicants. However when looking at the admission rates at the different departments at the university, it seemed that female applicants actually had some slight positive bias in regards to being admitted. The overall lower admission rates for female applicants were due to the fact that female students were more likely to apply to the departments which had lower admission rates. In order to avoid Simpson's paradox some thorough analysis is necessary. In that analysis it is necessary to look at these disparities between different subcategories and depending on those results making some design choices for the AI.

SIMON-test The SIMON-test works on a degree-specific level, making Simpson's paradox highly unlikely. Working degree-specific means that there won't be any more logical subgroups to define, making controlling for one or more variable not valid anymore.

Definition 4.1.14 (Emergent bias) *Emerges some time after a design is completed as a result of changing societal knowledge, population or cultural values. [42]*

Emergent bias arises after a system has in use for some time. For example the use population of a system may change from what was expected in the design. These changes can be due to different groups starting to use the system or because of larger societal changes. This is a very tricky form of bias to identify and requires a strong knowledge of the use population and their evolution. Interfaces with which users interact are also prone to emergent bias. An example of emergent bias could be a platform on which personal medical data can be consulted. The original design of the platform might have been that only doctors could access it. However some new government regulation dictates that individuals can now also consult the platform and see their medical files. The platform however was not designed to be used by people with little or no medical knowledge. This could lead to people with little medical knowledge to avoid the site, creating the appearance that they aren't interested in their medical knowledge.

In the EU proposal preventing this type of bias in a way included in paragraph 49 [4], where it states that *High-risk AI systems should perform consistently throughout their lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity in accordance with the generally acknowledged state of the art.* This can be seen that the model should be able to cope with possible emergent bias. This is also reiterated in paragraph 54 [4] where it states *The provider should (...) and establish a robust post-market monitoring system.* again indicating that the system remains monitored after going into production.

SIMON-test The author of this dissertation does not feel they can comment on emergent bias for a lack of thorough knowledge of the interaction of students with the SIMON-test and its evolution.

Definition 4.1.15 (Presentation bias) *A user can only interact with the elements that were presented to them. [43]*

Presentation bias becomes present when there are discrepancies between what different users get to see. These different views are often the result of recommender systems, those are AI systems that based on previous interests of the user presents them with items they are most likely to also be interested in. Therefore the data can become fairly heavily biased as the users do not get the possibility to generate the same information, such as watching a certain television show. One solution to lessen the effects of presentation bias is a method "explore and exploit" by Agarwal et al. [44]. The idea is that only part of the items presented come from the recommender system and the other part are items randomly chosen. This gives the user the possibility to explore items independently from their previous interests.

Presentation bias can also be extended to include the bias introduced by ranking systems. Users will be recommended items with higher scores leaving items with no scoring or with a lower score less likely to be seen, and thus also less likely for their scores to change. Representation biases are generally biases that are introduced by customising user experiences. Meaning that any system which makes the items seen by the user not random introduce some form of bias. However it is often desirable for the users to have customised view, so instead of preventing it a smart way of handling this bias is preferred.

In general presentation bias is not common as it requires quite a sophisticated system before the data is collected. None the less it is interesting how a more complex system can increase the bias in the data.

SIMON-test To a certain degree some presentation bias is present as the courses are also sorted on the interest level of the student, this could create some minimal presentation bias. However the effects of this will be minimal as other factors outside of the SIMON-test will influence a student's choice. This is different compared to ordering a recommended product online, as the decision for a degree requires more consideration.

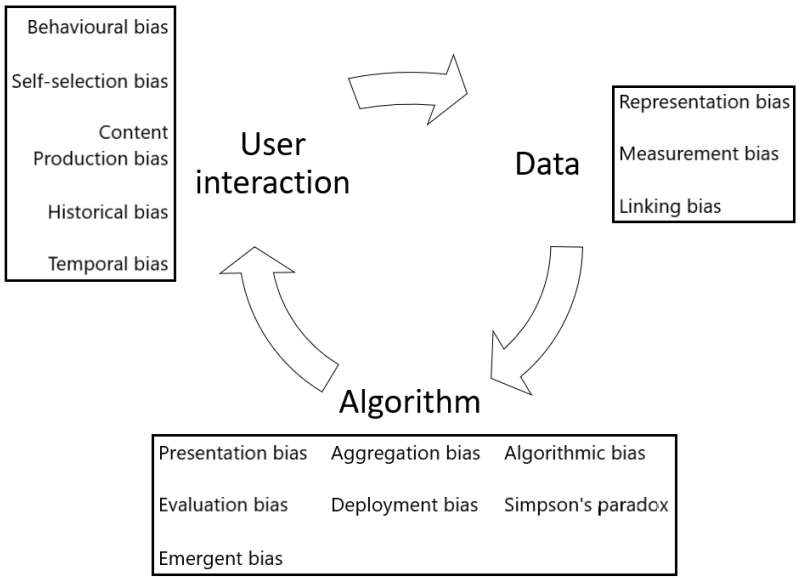


Figure 4.2: The feedback loop in data collection with bias types

4.2 Analysing the OULAD data set

The first step in analysing a data set is understanding the data set that will be used. This dissertation uses the OULAD public data set [45] as an example of data analysis and to test different techniques to enhance fairness. OULAD stands for Open University Learning Analytics Data. The OULAD data set contains information about courses, students and their interactions with the Virtual Learning Environments. The goal in this dissertation with the data set is to predict if a student would pass a course or not. For this reason certain adjustments are made to the data set. These adjustments must be analysed in order to check whether they introduce bias.

The OULAD data itself was collected over two years. It contains seven different courses, three of which are in the social sciences research area and four are STEM courses. The number of students who took a course over the two years varies from 748 to 7,909 per course. In fairness literature this is already a decently sized data set in the field of education and especially recent [7]. Another strength of the OULAD data set is the broad variety of sensitive attributes it contains, which is difficult for a public data set that needs to be anonymous. These sensitive attributes include, gender, age range, index of multiple deprivation of where they live and whether the student has a disability.

4.2.1 Data set creation

The eventual set of features in the data set can be found in table 4.1. This set of features needs to be constructed for all students that took the course in order to train the AI. The types of the features are also included as the features differ in type. The data set creation for the classification is constructed out of three tables from the OULAD data set, the student's activities on the platform, the student's test score and the general information of the student. The data set constructed for this dissertation uses the data available after twelve weeks from the start of the course. The interactions with the online environment are programmed dynamically so these can be added easily when changing the number of weeks after which the prediction is made. However the test scores of the students will require some manual changes in order to include them into the data.

The created data set only contains data for the course BBB. This was done as the AI system predicts whether a student would pass a course or not, which makes working on a course per course basis a logical choice. This also helps to avoid Simpson's paradox and to make the data aggregation simpler. All data gathered from the OULAD data set is filtered for the course BBB. The student's activities on the platform are aggregated in order to create uniform features. These features are the number of clicks a student makes on a type of content during a week, some personal information of the student and the scores they had on the tests that fell in the first twelve weeks. Including the clicks a student made on the platform can give an indication of how actively engaged they are with the course material. This does also mean that a student who has no activity logged on the platform won't be included in the data set.

For the course BBB two test moments were organised in the first twelve weeks of the course, both these test moments were graded by a tutor. In certain sessions a third computer graded assignment was organised. Because this was not present in all the sessions, this test session was not included as a feature. The first three sessions had the same score weights for the two included tests, however the fourth session did not. Because it is not known why this change was made, it is ignored for the data set. If a student had no results for either of the tests then they are not included in the data set. This means that students who dropped out before the first test (day 12 or 19 of the course) will not be included. Some students also had missing values for one of the evaluations. Because logistic regression will be used for the AI system it does not handle NaN values, so it was necessary to change this value. The decision was made against removing those students from the data set, in order to save data. A default value of 0 was chosen to replace the NaN values. A zero is also the same score a student normally gets for not taking part in a test.

Forumng_clicks_X	The number of forum interactions by the student during week X	Integer
Homepagepage_X	The number of homepage interactions by the student during week X	Integer
Oucontent_X	The number of content interactions by the students with Open University content during week X	Integer
Subpage_X	The number of clicks to a subpage by the student during week X	Integer
Url_X	The number of clicks on an URL by the student during week X	Integer
Resource_X	The number of times the student interacts with resources during week X	Integer
Glossary_X	The number of times the student opens a glossary during week X	Integer
Ouelluminate_X	The number of time the students uses Blackboard Collaborate[1] on the Open University platform during week X	Integer
Oucollaborate_X	The number of times the student uses collaborate during week X	Integer
Quiz_X	The number of times the student interacts with the Quiz tool during week X	Integer
Sharedsubpage_X	The number of clicks to a sharedsubpage by the student during week X	Integer
X_test	The score of the student in the Xth test	Integer
Passed	Whether or not the student passed for the course in the end. If true this means the student passed or passed with honours. If false this means that the student failed or dropped out.	Boolean
Male	If the student identifies as male or not	Boolean
Age_band	The age range (in years) in which the student falls	0: 0-35 1: 35-55 2: 55 $\leq$
Disability	Whether the student has a disability or not	Boolean
Imd_band	The multiple deprivation index [46] in which the student falls based on their home address	0: 0-10% 1:10-20% ... 9: 90-100%
Highest_education_X	Whether the students highest education is of type X	Boolean

1 https://www.blackboard.com/

Table 4.1: Adjusted feature set based on the OULAD data set

Lastly the student's personal information is added to the data set. Part of the features generated from the personal information are sensitive attributes, the other part are informative features. In the student's personal information their result for the course is also included. This is the label that the AI system will try to predict. The student's result is however simplified, passing with distinction will be mapped to pass and withdrawing from the course is seen as the same as failing the course. As mentioned in the previous paragraph, people who withdrew fairly early from the course will not be included in the data set. Simplifying these results removes some information from the system, but also simplifies the task.

The highest degree a student has obtained is included as an informational feature. A hierarchy could be made amongst degrees, however it was chosen not to do this as this can lead to some difficult decisions, which would introduce bias. Therefore this feature is one-hot encoded, leading to some sparse features. Another informative feature that was included is the number of credits a student is following. This can give an indication of the student's motivation and their work load. The rest of the features gathered from the student personal information are sensitive attributes. Two binary features are included, the gender of the student and whether they have a disability. Originally these were binary features encoded with strings, these were changed to booleans.

The last two features added to the system are the sensitive attributes of age range and the index of multiple deprivation (MDI) of the student's home. The index of multiple deprivation indicates the chances a person is deprived of based on where they live. These are expressed in percentage ranges with lower percentages indicating that they were more deprived of chances. Some of the MDI values are missing from the data set. Because this only occurs in 54 instances, the decision was made to remove these rows from the data. Just like the age range, the MDI is encoded from strings to a numerical hierarchical feature.

This results into a data set with 5989 rows and 156 columns. For an AI application this is a fairly small amount of data. This further supports the decision to work with a simpler model to try and achieve the classification. Important to note that with the way the data is constructed, it is possible for the same student to appear twice in the data set. This could happen because they followed the course in more than one organised session.

4.2.2 Bias analysis

In the following section the data set is analysed based on the biases discussed in section 4.1. This is an example of how a bias analysis can be done for a data set, where the data itself is available. This is however not representative for all data sets as characteristics, such as how the data is collected and

the use population have a large influence on the bias analysis. Certain types of biases will require more analysis than others.

User interaction bias

The user interaction bias is difficult to assess when there is no access to the method used to collect the data. In this case it is the online platform of the Open University. Another difficulty is the limited social knowledge of the author about online universities.

Behavioural bias The data was collected through the online platform of the Open University. The user has three possible methods of interacting with the platform, either they use the website on a desktop, or they use the website on a mobile device, or they make use of the mobile application. There is no information available through what platform the data points are gathered and what differences there are between the platforms. Therefore no assumption of user interaction bias is present in the data set. If it was possible to change the collection process then the advice would be to include the method to the interaction itself. This would make it possible to analyse sensitive attributes against the collection method. If the result shows that these are not spread equally then a deeper evaluation for bias is necessary.

Content Production bias Content Production bias cannot arise in this data set. The content is generated by clicks on the forum and have no link with producing text.

Historical bias The analysis of historical bias needs to be done by someone with a background in social sciences and knowledge of online universities. The situation is more complex to analyse for historical bias as the data is collected passively, so correlations in actions with a sensitive attribute are more difficult to ascertain than in a questionnaire. Because the author of this dissertation does not contain these capabilities no statements about the prevalence of historical bias will be made.

Temporal bias The data itself was collected over the period of two years, 2013 and 2014. There is no reason to believe that any systematic distortion would have occurred during those two years. Therefore it is assumed that no temporal bias is present in the data set. It is however possible that temporal bias would arise when the application is implemented for the students to use. As there is a time gap between training and the implementation of the data. Nothing can be done in the design process to ascertain if this would be a problem as no current data is available to compare with.

Self-selection bias The users of this data set have not selected themselves as this data was collected through the normal use of the Open University platform. The decision of the user to be present in the

data set also did not depend on some sort of criterion that they needed to interpret. For this reason no self-selection bias should be present in the data set.

Bias in the data set

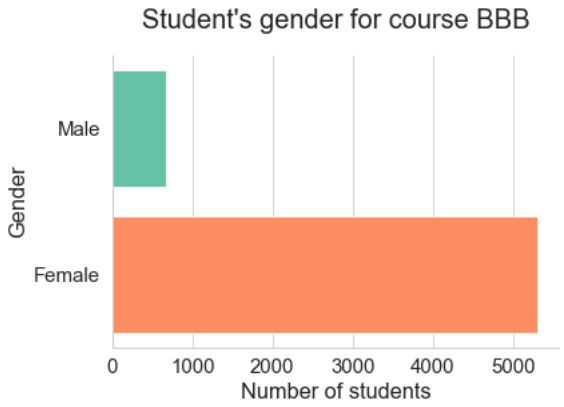
In the following section the possible biases introduced by the data set will be discussed. This discussion will be more elaborate because it involves the statics in the data set itself. These kinds of biases can either be introduced in the data set of OULAD itself or through the adaptations made for the classification task.

Representation bias Representation bias requires a fairly large analysis because the data set needs to be compared to the use population. This requires different characteristics to be investigated. The general characteristics of the population in the entire OULAD data set could be seen as the use population. This however would make some big assumptions as can be seen at the first characteristic. Therefore a societal viewpoint is necessary to determine if the data set is representative of the use population.

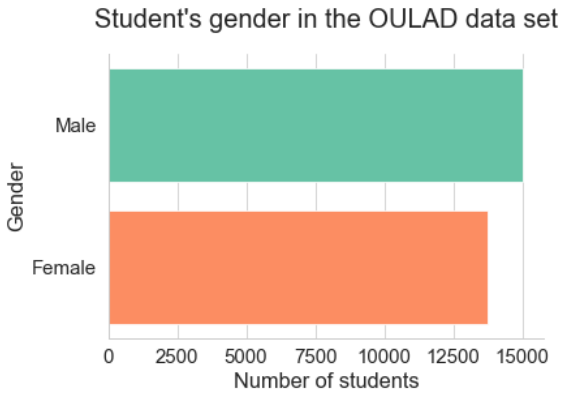
The first characteristic that will be checked is the gender balance. The OULAD data set only contains male and female as gender identities, so only a reflection can be made about that distribution. In figure 4.3 the gender distribution for both the BBB course specifically and over the entire OULAD data set is depicted. It is clear that the distribution in the course BBB is very different from the general OULAD data set. The course BBB is a course in the social sciences domain [45]. This give an explanation of why there are clearly proportionally a lot more women taking this course versus the general population of the Open University courses present in the data set. However the balance that 11.35% of the students identify as male is still very low even for a course in the domain of social sciences. For this reason we suspect some bias will be presented in the gender balance as males are less represented in the data set than in the possible use population.

The second element is the proportion of people with a disability taking the course. In figure 4.4 the disability distributions for both course BBB and the entire OULAD data set are given. It is clear that these distributions are very similar between the general use population of all the included courses and the course BBB at around 9.62%. This differs from the general population numbers in the United Kingdom[2], where the percentage of all people with a disability lies at 22%. This number can be refined more assuming that the use population of the open university will not be children and State Pension adults, leaving

[2] https://www.gov.uk/government/statistics/family-resources-survey-financial-year-2019-to-2020/family-resources-survey-financial-year-2019-to-2020

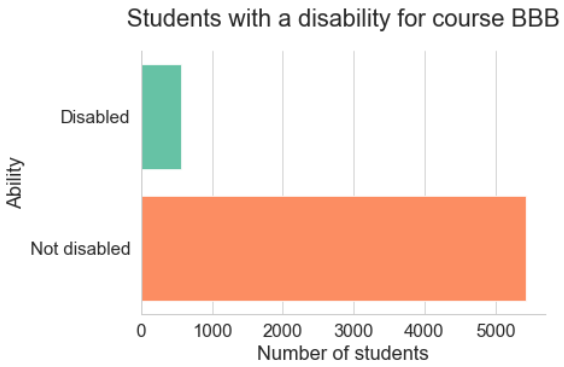


(a) Student gender distribution for the course BBB

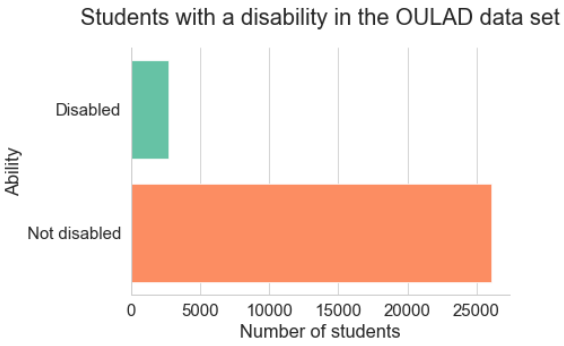


(b) Student gender distr. in the OULAD data set

Figure 4.3: Gender distributions of the students in the OULAD data set



(a) Student disability distribution for the course BBB

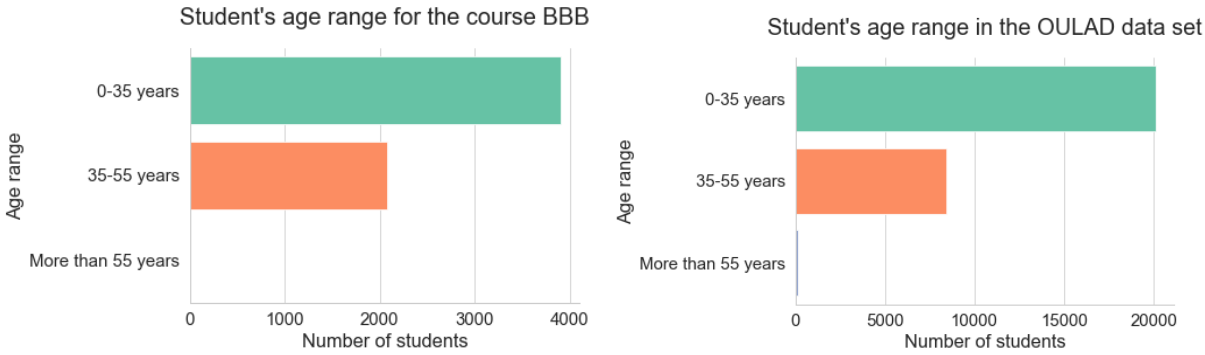


(b) Student disab. distr. in the OULAD data set

Figure 4.4: Disability distributions of the students in the OULAD data set

people at work-age where around 19% have a disability. In order to conclude if there is representation bias for people with a disability the proportions should be looked at on a larger scale. Because online universities are easier to attend for people with a disability than other universities it is difficult to decide this without further research.

Age ranges should also reflect the use population of the Open University. In the public data set available the age ranges are fairly broad in part to ensure the anonymity of the students. Because of this the representation bias analysis will also be fairly broad. In figure 4.5 the age range distribution for both course BBB and the entire data set are given. The first very noticeable difference is that for the course BBB there are nearly no people above the age of 55, while for the entire data set their share was at least visible on the graph. When looking in the data set for the course BBB itself, it was found that only 6



(a) Student age ranges for the course BBB

(b) Student age ranges in the OULAD data set

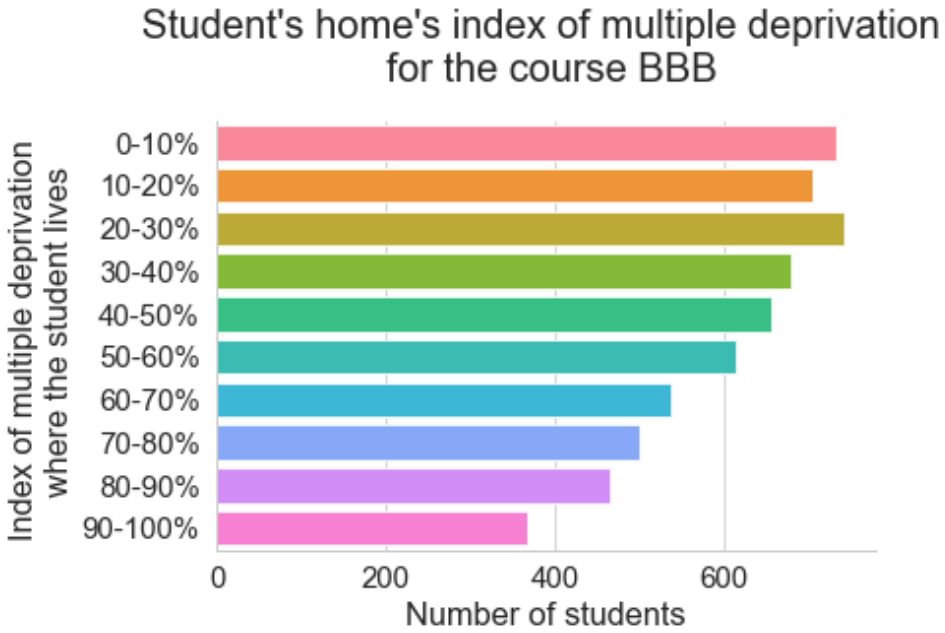
Figure 4.5: Age range distribution of the students

people over the age of 55 took the course. Proportionally slightly more people in the age range of 35 to 55 took the course BBB compared to the other courses. In the age range category there is probably some slight representation bias for the people above the age of 55. Even in the general OULAD data set the question of representation bias could be posed. As this proportion of people above the age of 55 is very small compared to the general public. The same recommendation is done as for the disability characteristic to do a more general research in the numbers and not only for these courses during the period of 2 years.

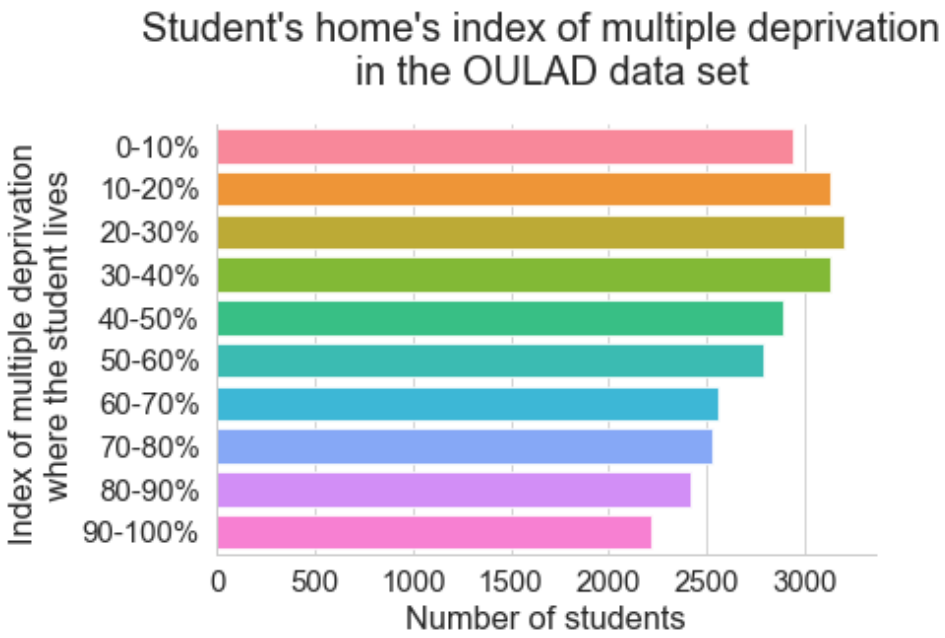
The final attribute that needs to be researched is the multiple deprivation index of where the student lives. This indicates the possibilities the student has gotten in their life. Representation bias with regard to the MDI could introduce some severe biases possibly against people with less resources than the average person. In figure 4.6 two plots are presented, representing distribution of the multiple deprivation index on a course-specific level and on the entire OULAD data set. The only real differences between the data for the BBB course and over all the courses in the data set, are in the higher percentage ranges of the MDI. This shows some slight representation bias. If assumed that the OULAD data set can be seen as representative for the use population of the Open University.

Measurement bias The first possibility of measurement bias is in features concerning the test scores. If the difficulty of the tests differ between the sessions then some noise will be introduced into the system. This however is not correlated to a certain sensitive attribute but rather the moment they took the course. This means that their is no bias introduced in the data, but only noise.

In the original data set there were four possible outcomes that a student could get for a course: Distinction, Pass, Fail, Withdrawn. Combining distinction and pass to simply passing the course does not



(a) Student's home's multiple deprivation index for the course BBB



(b) Student's home's multiple deprivation index in the OULAD data set

Figure 4.6: The multiple deprivation index distributions of the students' homes

change anything significantly, only a small amount of information is lost in that process. This information can be deemed not necessary for the task. However fail and withdrawn were also mapped onto the same value as fail. From the standpoint of the university it can be seen as the same, but for a student there is a reasonable difference. However the reasons for a student to withdraw from the course are not known. It could be that they withdrew from the class because it was too difficult for them rather than personal reasons. The decision was made to map these values as the same, largely because a large portion of the enrolled students dropped out of the class. In appendix B.2.1 the distribution of the dropped out students can be found. It turns out that both people who identify as male and people who have a disability are slightly more likely to drop out. More significant is that people who come from a region with a higher multiple deprivation index, thus who are less deprived, are less likely to drop out of the course compared to people who live in other regions. As there are only small differences across the groups of people with different sensitive attributes, a small amount of measurement bias can be suspected due to aggregation of withdrawn and failed as the same label.

The last form of measurement bias that can arise in the adjusted form of the data set is due to missing values. In general most AI systems don't function properly if there are NaN values in the data. In most cases one of two approaches is used. The first possibility is to remove the data lines which have missing values. If this method is chosen it could introduce representation bias, but no measurement bias. The second method is assigning an arbitrary value for the missing feature. This could introduce some measurement bias, but for this the missing values would need to be more prevalent for one group of people compared to the rest. The analysis was done for the distribution of people who were left out of the data set because not enough data was generated. This analysis showed that most distributions were similar to the data distributions themselves when split on the sensitive attribute. The distribution when split on the IMD property was not very similar to the distribution in the data set. However the IMD distribution in the data set is more closely related to the distribution of the general population, so there should be no representation biased created by removing these features.

Linking bias The data set cannot really be constructed as a graph. However because the data set is constructed from multiple tables in a data set it could be that some information can get lost when combining these tables. Because of how the data set was constructed and the quality of the OULAD data set, there should be no linkage error or at least no systemic errors disadvantaging one group of students based on their sensitive attributes. It is assumed that the OULAD data set contains the correct student interactions with the platform, thus not having any linkage errors through there. The construction of the classification data set was done as to remove the person where data was missing for them in one of the

tables. This should also prevent missing links and duplicate links did not seem to arise in the data set.

Bias through the algorithm

The bias introduced by the algorithm in this section will be about the Open University platform and the possible deployment method of the classifier. It is assumed that the classifier will simply be a module in the course which calculates if the student would succeed or not, based on their current efforts.

Presentation bias It is assumed that the Open University functions like all learning platforms. This means that no personalisation is used for the platform but rather everything is presented in a chronological order. All students are getting the same elements presented to them, so there would be no presentation bias in the system.

Aggregation bias The proposed system would be a simple logistic regression model and it would only be trained using the data from the specific course in which it will be used. There should be no aggregation bias as the training data is from the specific task trying to be achieved.

Algorithmic bias Algorithmic bias is what will be investigated in section 5. The goal in that section is to even reduce the overall bias through the use of the model.

Evaluation bias The evaluation of the AI system is done against a subset of the data. So the training and test data come from the same data source, which is specifically for this task. The only form of bias that can be introduced is when the split between the training and test data is done in a manner that they are not representative anymore for the use population. In section 5.1.2 the method is discussed for making this correct split.

Deployment bias Deployment bias is difficult to predict in an application which has not gone into use yet. A possible deployment bias is that it could demotivate the students to continue the course, leading to more drop-outs. It is supposed to be used as a motivation metric and not a deterrent.

Simpson's paradox Simpson's paradox does not apply on this problem. If however an application would be made for all the courses offered by the Open University than Simpson's paradox comes into play as that problem could be way too general and break the information from the statistical metrics.

Emergent bias For an application still in development it is not possible to yet have emergent bias. This should however be checked regularly, but this is discussed in broad lines in chapter 6.

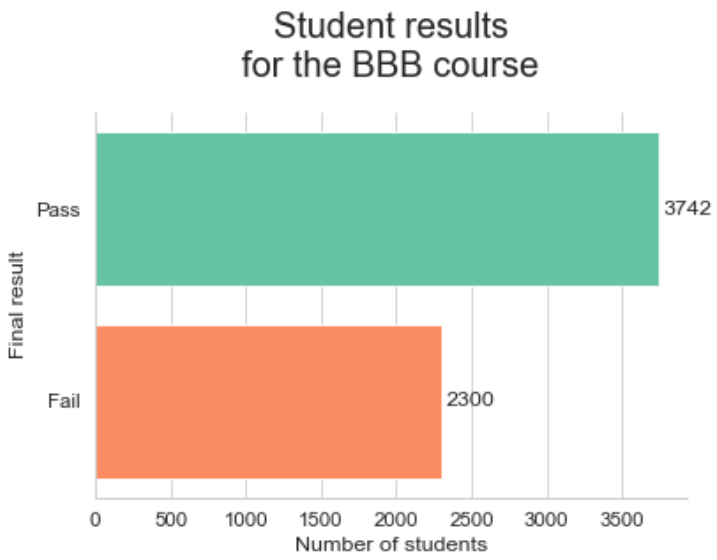
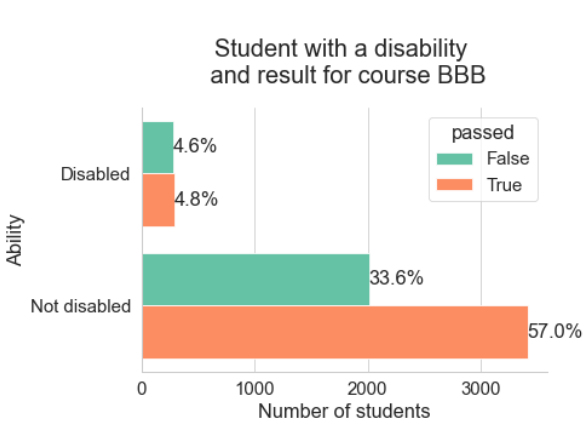


Figure 4.7: Base rates for the BBB course

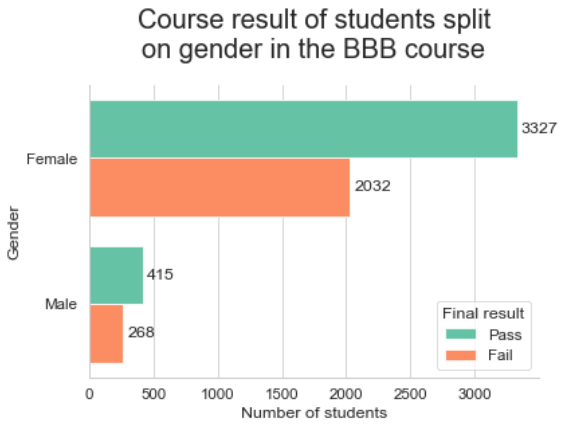
4.2.3 Analysis of the base rates

The structure and content of the data was already discussed in detail in subsection 4.2.1. It is however important to gain an insight on the balance of the student's results based on their sensitive attributes. This balance will play an important role in how the different fairness definitions could be satisfied. This balance is called the base rate in statistics and will also be included when evaluating different fairness techniques in chapter 5. The general balance of people passing and failing can be seen in figure 4.7. The figures depicting these balance based on the sensitive attributes can be seen in figures 4.8a, 4.8b, 4.9a, and 4.10. When analysing these figures it is clear that only the split on gender results in the same balance of passed and failing in each group. The other sensitive attributes show that people with a disability, or a lower age or that live somewhere that has a lower index of multiple deprivation on average have a lower base rate than the other groups based on that sensitive attribute.

The last personal attribute of which its correlation with succeeding is investigated is the highest degree the person has achieved when taking the course. The results of this can be seen in figure 4.9b. It is clear that there is a correlation between the degree someone has achieved and them succeeding. This supports the decision of making this a feature in the data set rather than a sensitive attribute.

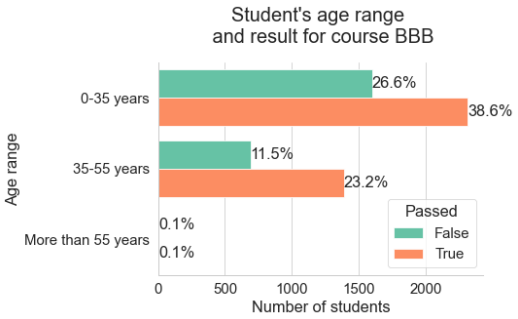


(a) Base rates of students split on disability

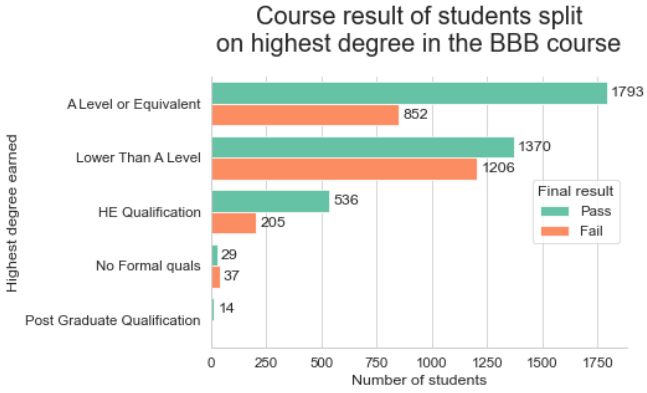


(b) Base rates of students split on gender

Figure 4.8: Base rates of students split on disability or gender for the BBB course



(a) Base rates of students split on their age range for the BBB course



(b) Base rates of students split on the highest degree earned for the BBB course

Figure 4.9: Base rates of students split on age range or highest degree for the BBB course

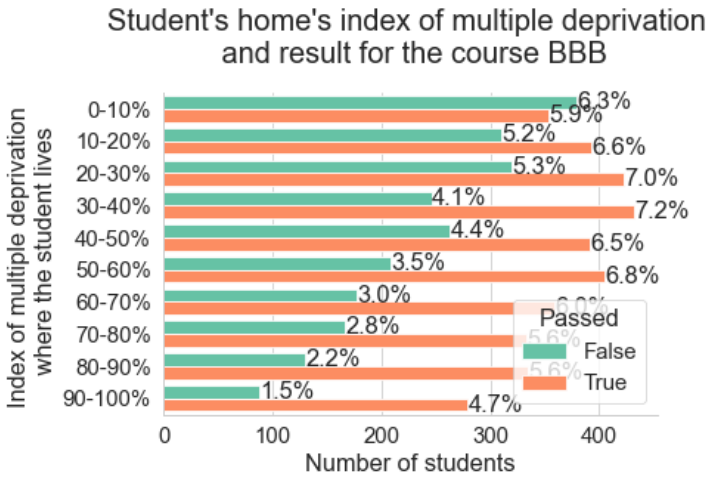


Figure 4.10: Base rates of students split on the multiple deprivation index of their home for the BBB course

Chapter 5

Bias mitigation techniques

5.1 Aspects of bias mitigation

Bias mitigation techniques have the goal of removing or lessening the bias from the data. In essence it is the opposite of Algorithmic bias (Definition 4.1.9), where the algorithm creates or amplifies bias. In order to reduce bias in the algorithm the technique intervenes at a certain stage, this is discussed in Section 5.1.1.

It is important to create good train, validation and test sets in order to avoid bias from arising in the system. The concept of creating good sets is also included in the European Union proposal for AI [4, §44]. The train and test sets are static and their construction is discussed in Section 5.1.2. The validation set used for tuning the hyperparameter is created using k-folds, this means that multiple sets are split off from the training set in order to achieve the validation score, this is discussed in Section 5.1.3 for the base model. Finally the evaluation process is discussed for the base model in Section 5.1.4. The base model serves as a type of benchmark to compare the different techniques to. This is valuable to determine the changes in behaviour and accuracy that the technique bring forth.

5.1.1 Types of bias mitigation

Three types of bias mitigation techniques types can be distinguished. These types are based on where in the process they take place, hence the names pre-processing, in-processing and post-processing. These are separated this way because not all AI systems are accessible in every way.

Pre-processing

Pre-processing is making changes to the data which the AI system receives [47]. The idea behind it is to reduce or remove the bias already present in the data. In order to use pre-processing a good knowledge of the data itself is necessary. One problem perceived with the pre-processing technique is that it is based on human insight and not dependant on statistics. This can make it difficult to tune this technique to achieve the best possible result.

In-processing

In-processing techniques can be more difficult to comprehend what is happening. They affect the formulation of the classification problem in order for it to be aware of the possible discrimination behaviour [47]. It is also possible to have in-processing techniques for non-classification tasks, but those are out of scope for this dissertation. In order to use in-processing techniques, it is necessary that the model itself is accessible, and it is possible to change it. This means that the system should be owned by the business that uses it; otherwise, these adjustments can prove to be very difficult. Even if the business itself own the system, it could sometimes still be difficult to implement these changes. In general in-processing techniques are the most powerful. This is because they make the system aware of the discrimination problem instead of trying to find fixes.

Post-processing

Post-processing techniques take effect after the model has already learned from the data [47]. Post-processing can happen in two different manners. If the system itself is a white-box, or in other words, it is accessible, then post-learning changes can be made in the model's inner workings. The other manner is when the system is treated as a black-box, data is fed into the system, and a prediction comes out of it, but nearly no other information can be gained. Then the post-processing technique purely works on the predictions it gets out of the system.

AI systems are often black-box or at least treated as a black-box system. This makes post-processing techniques that work with the system as a black-box very important. The white-box techniques are also interesting but are often not favoured compared to the in-processing technique, which are often more powerful.

When tuning a post-processing technique, it is important to watch out for bias introduced by the post-processing technique. This can consist of adjusting the decision boundary or the samples close to it.

Not making informed decisions can introduce bias into the system which is the exact opposite of what is desired.

5.1.2 Creating train and test set

From the data a train and a test set needs to be constructed. The train set is used for training the model. The test set serves for comparing different models against each other. The trained model is used to predict the labels of the test set. A measure can be calculated from the relation between the predicted labels and the actual labels, and different models can be compared based on that measure. In the case of fairness the accuracy and statistical properties mentioned in section 3 are best used for comparing different models.

The simplest way to create a train-test split is by using a random sampler. This will randomly assign a sample to either the train set or the test set. When a data set is not large this method possibly violates the assumption that the data is distributed equally across both sets. If this were to occur as mentioned in subsection 4.1, bias could be introduced in the form of representation bias and evaluation bias. This can be evaluated by looking at the data distributions in the train and test sets.

A possibility to compensate for this effect is creating multiple test sets (with corresponding different train sets). Averaging these sets will result in similar statistical properties as when the data set was large enough not to violate these properties. Moreover, it includes the added bonus of being able to calculate the standard deviation, which is not possible when only using one data set. The downside of this technique is that it requires the model to be retrained multiple times. When using a small model, this is not a big problem, but when starting to use larger models, one should take into account the energy and time consumption this retraining requires.

A second possibility is not making the split randomly but splitting based on the demographic properties. In this case, the split is done very consciously but requires more work to accomplish. Certain demographic groups will become very small when splitting so a strategy needs to determine in what order a split must be made and how small demographic groups need to be handled.

The latter was applied for the OULAD adjusted data set. In order to ensure the best possible split, the order of splitting was done, starting with the categories in which the smallest groups were present. It is important to note that these categories do not only contain the sensitive attributes but also other personal attributes. This is because making this split is not based on the fairness definition but more on the proper distribution of the data, a property about which machine learning algorithms are very sensitive.

For the OULAD data set, the eventual split was made in the following order: The prior education, the index of multiple deprivation, age, gender, and then disability. This leads to 4479 samples in the data set and 1506 samples in the test set.

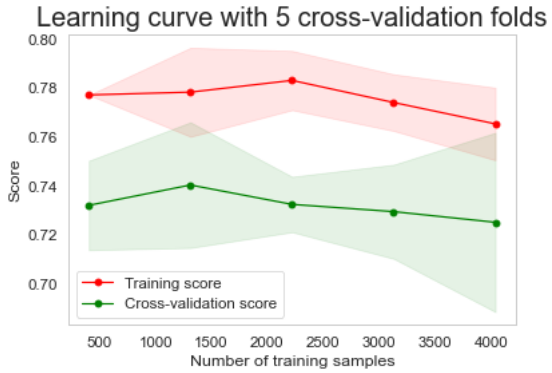
The OULAD data set is fairly small, so a train-test split of 25% was used in order for enough training power to remain and having the test results for the different models to be representative enough. The test set is used to ascertain the different group fairness definitions discussed in chapter 3.

5.1.3 Number of folds for the base model

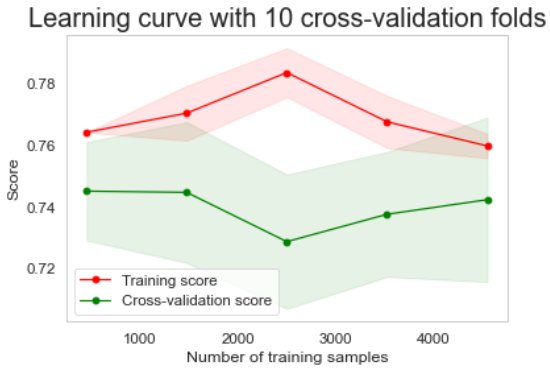
The base model is the LogisticRegressionCV model from sklearn. This is an implementation of basic logistic regression but with a built-in function that optimises the regulation strength through the use of cross-validation. In cross-validation, a train set and a validation set are created with the use of folds. Another data test split means that the same problems start to occur as with the train and test set split. However, it is more cumbersome to create a function that will make relatively fair splits as was chosen for the solution in creating the test set. Therefore it seems better to use the other proposed solution for that problem in these situations, namely working with multiple sets. This can be easily done by increasing the number of folds used.

The basic value for the number of folds used in sklearn is 5. Another common value for the number of folds is 10. Determining what number of folds are necessary for proper validation figures can be done through the use of the learning curves. The learning curve with 5 folds can be seen in figure 5.1a and with 10 folds in figure 5.1b. Both these learning curves do not look very proper. The false conclusion could be made that the model is not behaving properly. Howevern no behaviour could be ascribed to the volatile behaviour. With the prior knowledge of the data, this volatile behaviour in the learning curves can be ascribed to higher prediction errors due to the distribution of the data.

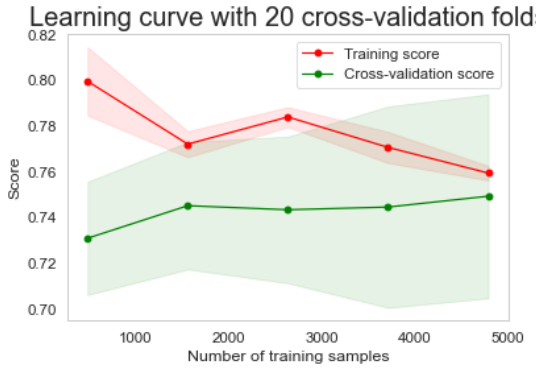
Increasing the number of folds lowers these prediction errors [48], this however comes with increasing the computation time. Looking at the learning curve with 20 folds in figure 5.1c the behaviour becomes more of what can be expected in a learning curve. However, the stagnation in the middle range for the number of training samples is still slightly strange. The final training curve portrayed in figure 5.1d uses 30 cross-validation folds. The cross-validation score increases steadily with increasing the number training samples and even starts to converge with the training score. Therefore the choice is made to use 30 cross-validation folds as a base number of folds when tuning the models. Based on the model itself, the best number of folds used can vary.



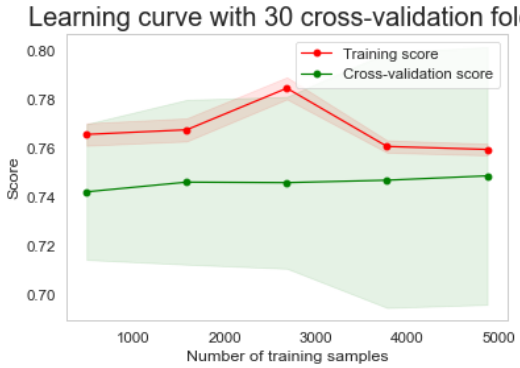
(a) Learning curve for the base model with 5 cross-validation folds



(b) Learning curve for the base model with 10 cross-validation folds



(c) Learning curve for the base model with 20 cross-validation folds



(d) Learning curve for the base model with 30 cross-validation folds

Figure 5.1: Learning curves for the base model with different numbers of cross-validation folds

5.1.4 Model evaluation of the base model

The overall accuracy of the base model is 75.1%. In the base model the sensitive attributes of a group are still given to the model as a feature. They thus have a direct influence on the decision the model makes for an individual. The weights for these sensitive attributes can be found in table 5.1. Interpreting the weights of a model can be a difficult feat where often wrong conclusions are made. Therefore the main analysis of the model will be done through the test set results, which is generally the custom. The one noteworthy thing about these weights is that the model weights are not zero for the sensitive attributes. This means that the model finds these characteristics to influence the eventual result.

As was discussed in section 3.3.1, when the sensitive attributes are given to the model as a feature, that model can show causal discrimination. A first test to indicate whether this model shows any causal discrimination is changing the values of the sensitive attributes in the test set and comparing the results. The binary attributes were simply inverted; the values for the categorical attributes such as the index of multiple deprivation and age range were shifted in order to have the biggest value change on average. The predictions between the original test set and the test set with the altered sensitive attributes were compared. No difference could be found between the predictions. This test result shows no evidence of causal discrimination; however, it is insufficient to determine that the system would never show causal discrimination. Especially with the non-zero feature weights, the user cannot be certain that there cannot be any causal discrimination.

Sensitive attribute	Model with smart split
Gender	-0.003367
Age	0.036027
Disabled	-0.048696
Index of multiple deprivation	0.102192

Table 5.1: Weights of the sensitive attributes in the base model with the smart and simple data split.

The most important part of the analysis is the statistical metrics used to determine the fairness definitions in chapter 3. The base rate (5.1), positive rate (5.2), true positive rate (5.3), true negative rate (5.4), false discovery rate (5.1) and false omission rate (5.5) are shown on a graphs such as in the figures 5.2, 5.3, 5.4 and 5.5. The accuracy will be mentioned separately. Accuracy is more subtle than the statistical measures on the graph and is, therefore, less suited to be included. To aid the user in understanding the

following questions, the formulas for the statistical measures are added below.

$$BR = \frac{nr.\ of\ positives\ in\ the\ test\ set}{nr.\ of\ elements\ in\ the\ test\ set} \tag{5.1}$$

$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{5.2}$$

$$TNR = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{5.3}$$

$$FDR = \frac{False\ Positive}{True\ Positive + False\ Positive} \tag{5.4}$$

$$FOR = \frac{False\ Negative}{True\ Negative + False\ Negative} \tag{5.5}$$

For each of the sensitive attributes the investigation must be made if the desired statistical measures are similar across the different groups. As was mentioned in other parts of this dissertation, it is possible that splitting on a sensitive attribute alone is not detailed enough and some groups can still get treated unfairly without this being visible through splitting solely on one sensitive attribute. In this instance, the splits are only done based on one sensitive attribute. Further splitting was not possible as the test set was not large enough to have further splits which were statistically large enough to be relevant. In the split on age groups, it was not possible to have a representative split for the last age groups ≥ 55. In the entire data set, only 6 samples for someone over the age of 55 were present. The decision was to have these 6 samples in the training set as no valuable information could be gotten from such a small subset in the test set.

The fairness definitions checked in the analysis are statistical parity, predictive parity, predictive equality, equal opportunity, equalised odds, conditional use accuracy equality, overall accuracy equality and treatment equality. The explanation and characteristics of these definitions can be found in section 3.3.2. Conditional statistical parity is omitted from the list because no condition could be found after which the groups were still large enough for statistical relevance.

Sensitive attribute - disability In figure 5.2 the statistical metrics can be seen for the test set split on the sensitive attribute of disability. At first glance it can be seen that the metrics can vary strongly between the group of people with a disability and without one. From these results the compliance with the different fairness metrics can be inferred as shown in table 5.2. In the test set there are 134 samples where the sensitive attribute of disability is true, and 1372 where it is false.

None of the fairness definitions except for overall accuracy are satisfied. This is most likely due to the large difference in base rate between both groups. Some statistical measures differ by around 18 percentage points between each other, while others are slightly closer related with 9 percentage points. One group fairness definition that can be seen as satisfied after consulting with the person or entity that will eventually be responsible for the application is conditional use accuracy equality. The differences are in a dubious range where they can be seen as equal enough especially because the current groups in the test are rather small. The accuracy of the group of people with a disability and without a disability differ strongly with 72.4% and 75.4% respectively. The false negative to false positive ratio in the base model is 0.321 for the people with a disability and 0.146 for people without a disability, this difference is fairly large, and it shows that people with a disability get far more false negatives than false positives compared to people without a disability. This is poor behaviour as false negatives are the worst possible behaviour that the model can portray.

Sensitive attribute - gender The situation in the split based on a person's gender is largely different from the one split on whether they have a disability, as can be seen in figure 5.3. The same table has been created while evaluating the different fairness definitions based on the gender split. The results can be seen in table 5.3. Nearly all statistical measures differ around 1% to 5% between the two groups split on gender, this is most likely because the base rate is much more similar between the groups. There are 170 samples of people who identified as male and 1336 who did not. Again there is one smaller group in the test set, in this case, for the people who identify as male. Therefore it is not unreasonable to see a difference of 5% as statistically insignificant given that the user agrees. The accuracy differs less than in the disability split with 74.1% and 75.2% for the group identifying as male and the group that does not respectively. The false negative false positive ratios are also similar with 0.222 for people who identify as male and 0.153 for people who identify as not male. The choice was made to deem any difference in false negative to false positive lower than 0.1 as satisfying equality, due to the limited size of certain groups.

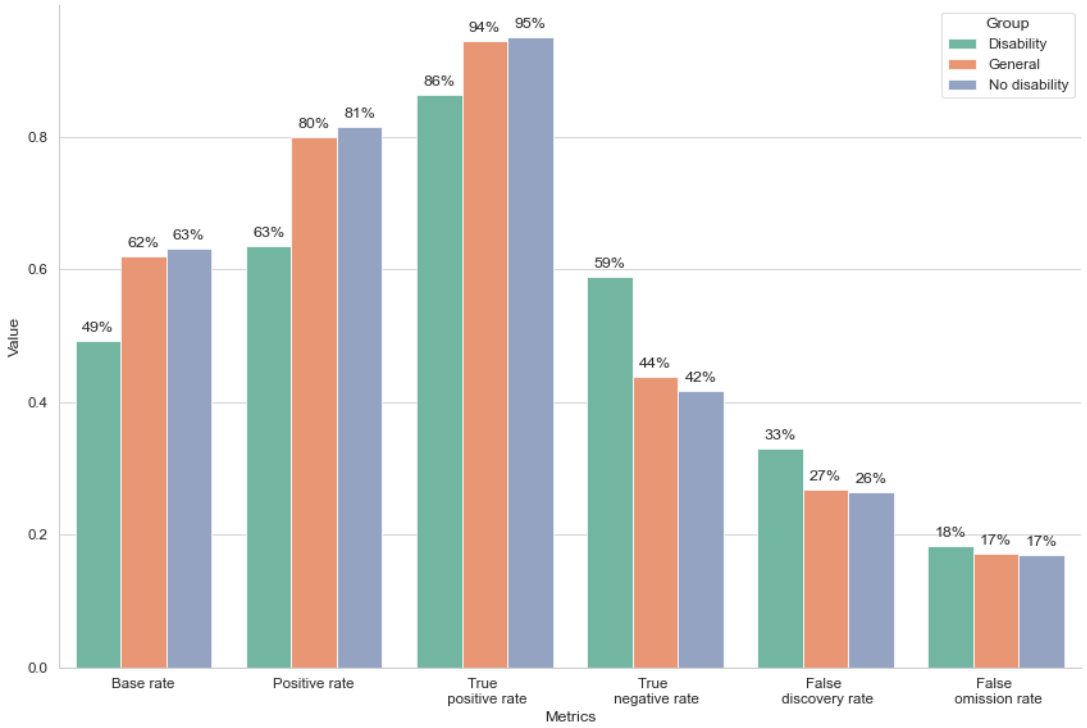


Figure 5.2: Statistical metrics of the base model with regard to the sensitive attribute of disability

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	✗	✗	✗
Difference	18pp	7pp	17pp	9pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	?	✗	✗
Difference	9pp - 17pp	7pp - 1pp	2.97pp	0.175

Table 5.2: Fairness definition compliance of the base model with regard to the sensitive attribute of disability

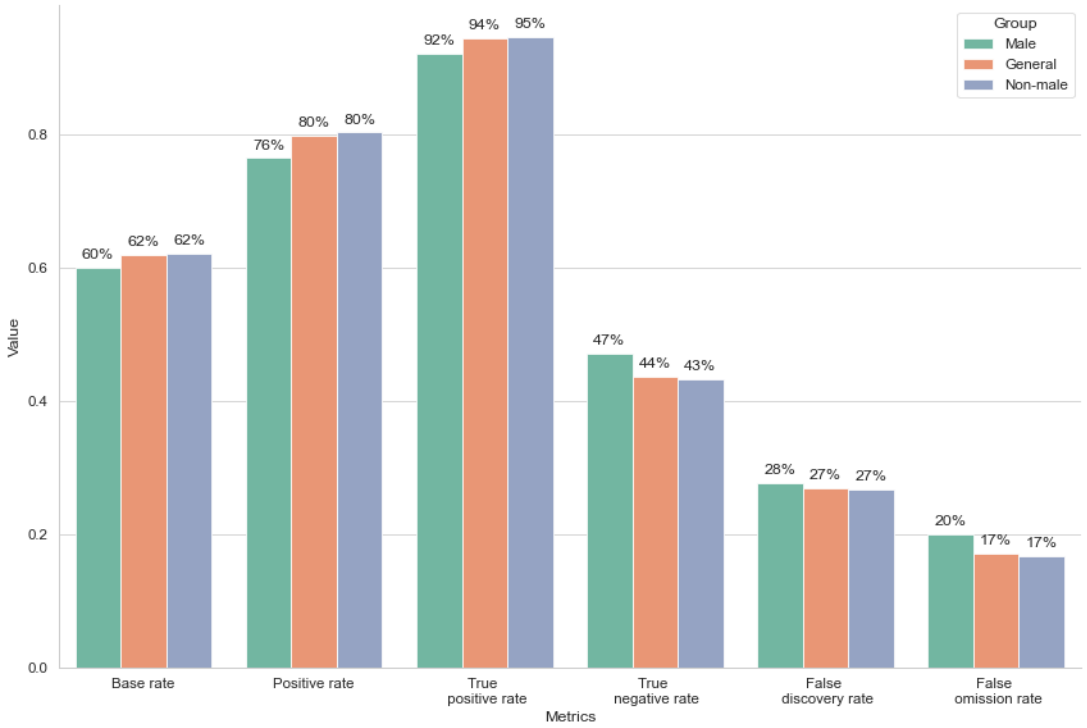


Figure 5.3: Statistical metrics of the base model with regard to the sensitive attribute of gender

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	**?**	✓	**?**	**?**
Difference	4pp	1pp	4pp	3pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	**?**	✓	✓	✓
Difference	3pp - 4pp	1pp - 3pp	1.1pp	0.069

Table 5.3: Fairness definition compliance of the base model with regard to the sensitive attribute of gender

Sensitive attribute - age range The second to last sensitive attribute to discuss is the age range, where the statistical measures are displayed in figure 5.4. The distribution of samples between these groups is a lot more balanced, with 986 samples of people under 35 and 520 samples for people between the ages of 35 and 55. As mentioned before, the age range above 55 is not included as there were not enough samples in the training data. Unlike the previous sensitive attributes, the equality between the statistical measures is clear as noted in table 5.4. By clear, it means that statistical differences of 1 or 2 percentage points are close enough in a data set of this size to be deemed equal. The accuracy between the age group under 35 and between 35 and 55 are similar with 75.3% and 74.8%, respectively. The ratio of false negatives to false positives is incredibly close compared to the previous sensitive attributes with 0.162 and 0.159 for the age group under 35 and between 35 and 55, respectively.

Sensitive attribute - index of multiple deprivation The index of multiple deprivation is the last sensitive attribute to discuss. It differs from the other sensitive attributes in that it splits the data set into ten different groups. Because there are more and thus also smaller groups, the tolerance of the difference between the different statistical measures must also be greater. Failing to be equal is seen more as a systemic distortion across the different groups, like for the True negative rate in figure 5.5. The general group is also left out of figure 5.5 as it is already very crowded and the evolution across the groups is the most valuable information that can be deduced. The accuracy varies strongly across the different groups; however, these variations are no systemic. Unlike for the false negatives to false positives ratios where there is a clear downward trend with a higher IMD. Therefore, it can be said that the overall accuracy equality is, but treatment equality is not satisfied. The specific values for the test set for the index of multiple deprivation can be found in table 5.5. This was done in order to ensure the readability of this text.

IMD range	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
Sample size	187	179	183	171	167	152	128	126	119	94
Accuracy	75.4%	75.4%	73.8%	78.9%	73.7%	71.7%	75.8%	71.4%	74.8%	81.9%
TN/TP	0.210	0.222	0.263	0.200	0.158	0.162	0.107	0.091	0.071	0.000

Table 5.5: Test set characteristics when split on the index of multiple deprivation in the base model

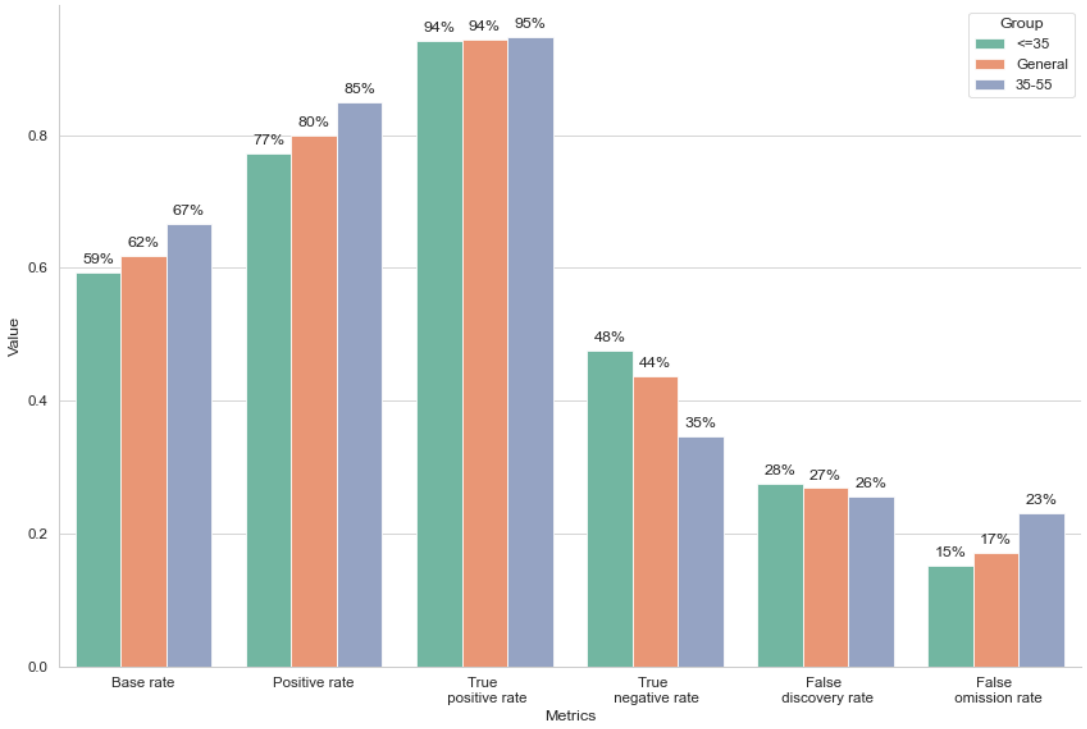


Figure 5.4: Statistical metrics of the base model with regard to the sensitive attribute of age range

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	✓	✗	✓
Difference	8pp	2pp	13pp	1pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	✗	✓	✓
Difference	1pp - 13pp	2pp - 8pp	0.5pp	0.003

Table 5.4: Fairness definition compliance of the base model with regard to the sensitive attribute of age

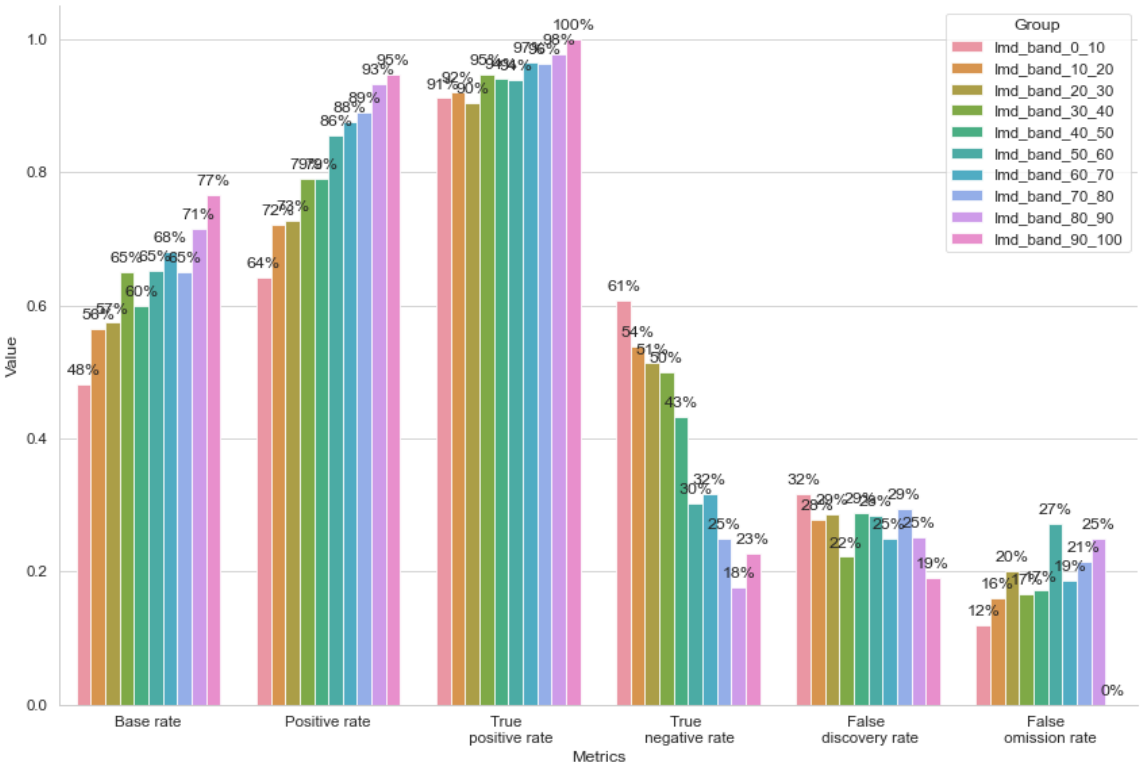


Figure 5.5: Statistical metrics of the base model with regard to the sensitive attribute of the index of multiple deprivation

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	?	✗	?
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	✗	✓	✗

Table 5.6: Fairness definition compliance of the base model with regard to the sensitive attribute of the index of multiple deprivation

5.2 Different bias mitigation techniques

In the following section different bias mitigation techniques will be discussed. The first two techniques discussed were already introduction in Section 3.3.1 about individual fairness definitions, mainly fairness through unawareness (Section 5.2.1) and suppression (Section 5.2.2). Both these techniques can be classified as pre-processing techniques. The following technique Threshold optimiser (Section 5.2.3) is the only technique discussed which specifically tries to satisfy certain group fairness definition. It uses post-processing to achieve this. The final technique discussed in Section 5.2.4 is an in-processing technique that works on the loss function of the algorithm. It is not per se a fairness technique, but can be used as one.

5.2.1 Fairness through unawareness

Fairness through unawareness differs from the base in that it does not use the sensitive attributes as features. In fact, the model itself does not use the sensitive features at all. The model's designer uses them uses them to gauge how the model acts for the different groups based on the sensitive attributes. Because the data is altered by leaving the sensitive attributes out of it, the fairness through unawareness technique can be seen as a form of pre-processing. It is not a pure form of pre-processing as it is possible that when the model is being used the data about the sensitive attributes is never gathered about someone.

As discussed in section 3.3.1, fairness through unawareness ensures that causal discrimination is impossible. So two people who have generated precisely the same data but only differ in their sensitive attributes are sure that they will receive the same prediction. The model, in this instance, cannot see the differences between the two persons as it is unaware of their sensitive attributes. Again, it is important to emphasise that this is not sufficient to rule out any form of discrimination. In the data generated the influences of one's sensitive attributes can still be present. A developer should still check if the different groups are treated equally based on their sensitive attributes, as will be done in the following section.

However, there is a cost for removing information from the data set; 0.8 percentage points are lost for the global accuracy. This cost, however, is minimal in this case and can be seen as valid for ensuring that no causal discrimination occurs.

Model evaluation

Sensitive attribute - disability Only one fairness definition is satisfied when using fairness through unawareness and splitting on disability as a sensitive attribute, which is the false negative to false positive

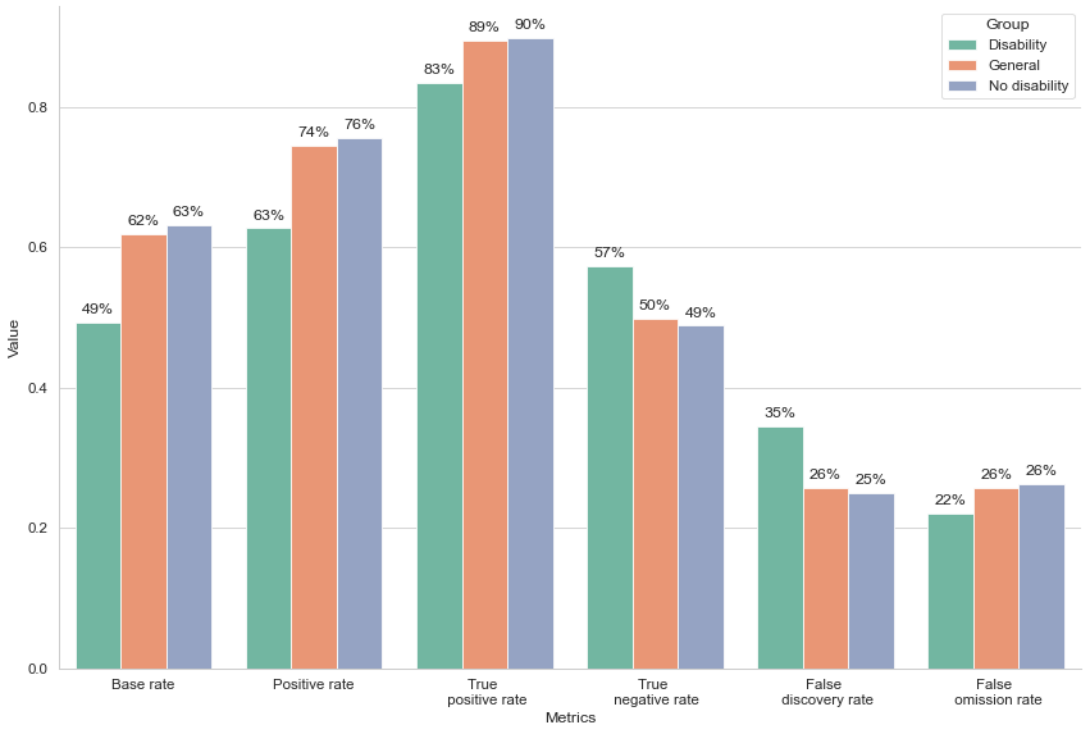


Figure 5.6: Statistical metrics of the fairness through unawareness model with regard to the sensitive attribute of disability

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	✗	✗	✗
Difference	13pp	10pp	8pp	7pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	✗	✗	✓
Difference	7pp - 8pp	17pp - 4pp	4.6pp	0.039

Table 5.7: Fairness definition compliance of the fairness through unawareness model with regard to the sensitive attribute of disability

ratio. This is noted in table 5.7 and can be deduced from figure 5.6. The other statistical measures also vary significantly more than in the base model. So while causal discrimination is removed from the system it does not mean well for most statistical measures used in group fairness definitions. The fairness through unawareness model accuracy is 70.1% for the group of people with a disability and 74.7% for the group of people without a disability. This is a larger difference than before, with a negative effect for people with a disability compared to the base model. This shows that removing simply removing the sensitive attributes will not make a system necessarily fair. However, the false negative to false positive ratio has become more fair as the ratio increased much more for the group of people without a disability, leading to 0.379 for people with a disability and 0.340 for people without a disability. However, this ratio is still to the disadvantage of people with a disability as they are more likely to receive a false negative than a false positive compared to the group of people without a disability. However one important note to make on this is the base rate of nearly 50% for people with a disability. This makes that there are the same amount of samples that can become a false positive or a false negative. While with the higher base rate of the other group proportionally there are fewer samples that can become a false positive, making a higher false negative to a false positive ratio more logical for them, which is not the case in this instance.

Sensitive attribute - gender All statistical measures included here to compare different groups are equal when the split is made on gender as is shown in figure 5.7. This leads to the model with fairness through unawareness satisfying all fairness definitions discussed and noted in table 5.8 for the grouping based on gender. This is the opposite situation from when the split is made on whether they have a disability. If the business that plans to implement this tool does not care about fairness across the other sensitive attributes, fairness through unawareness brings the optimal solution. However, it would be incorrect to simply ignore the characteristics across different sensitive attributes unless a very strong argument can be made. The reason that all the measures can be so equal is probably because of the very similar base rates. However, this should not be used as an argument for not having satisfied the chosen fairness definition for all sensitive attributes. If these base rates differed more radically, it would become impossible to satisfy all fairness definitions with the same model. The difference in accuracy is smaller than one percentage point with an accuracy of 73.5% for the group that identifies as male and 74.4% for the group that identifies as not male. After removing the sensitive attributes as features, the difference decreased slightly for false negative to false positive ratio , but the values increased. In this case, the ratio becomes 0.324 and 0.346 for the group that identifies as male and the group that identifies as not male, respectively.

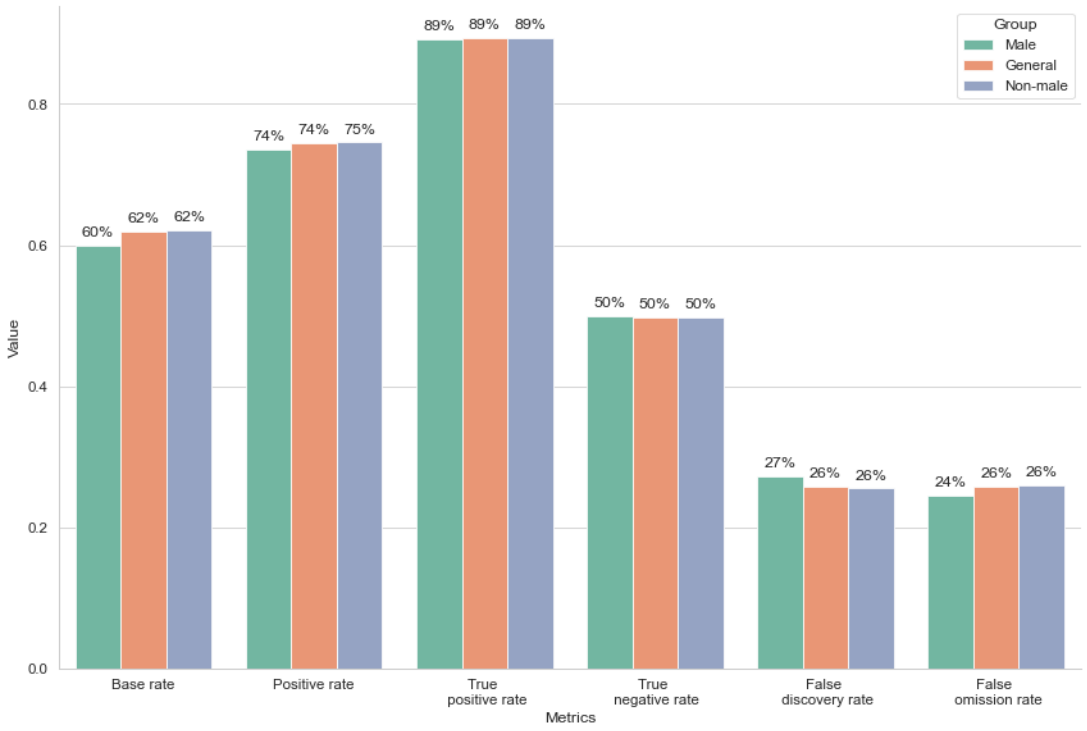


Figure 5.7: Statistical metrics of the fairness through unawareness model with regard to the sensitive attribute of gender

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✓	✓	✓	✓
Difference	1pp	1pp	<1pp	<1pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✓	✓	✓	✓
Difference	<1pp - <1pp	1pp - 2pp	0.9pp	0.022

Table 5.8: Fairness definition compliance of the fairness through unawareness model with regard to the sensitive attribute of gender

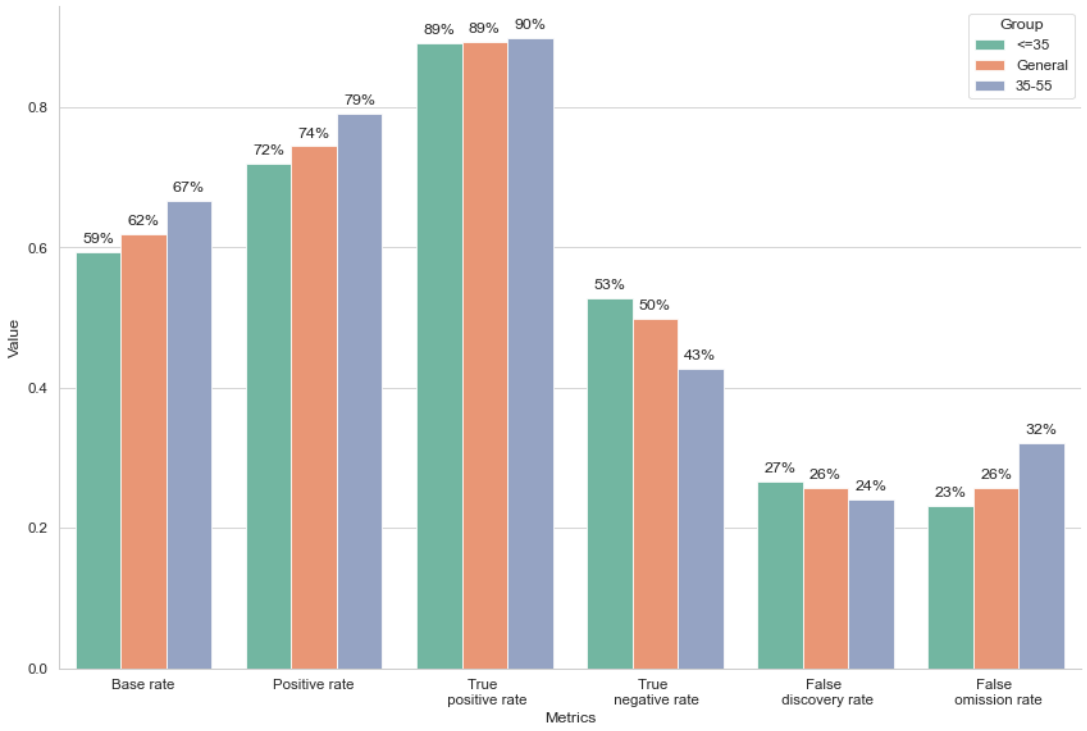


Figure 5.8: Statistical metrics of the fairness through unawareness model with regard to the sensitive attribute of age range

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	✓	✗	✓
Difference	7pp	3pp	10pp	1pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	✗	✓	✓
Difference	1pp - 10pp	3pp - 9pp	0.11pp	0.015

Table 5.9: Fairness definition compliance of the fairness through unawareness model with regard to the sensitive attribute of age

Sensitive attribute - age range The split on the sensitive attribute of age range tells yet another story compared with the base model results. The results of the statistical measures are shown in figure 5.8 and whether they satisfy the fairness definitions in table 5.9. The model satisfies the same fairness definitions as the base model with the added bonus that the differences in the definitions it did not satisfy shrunk. The differences in accuracy between the groups are also very minimal with 0.11% as the prediction accuracy for the group under 35 is 74.34% and 74.23% for the group of people between the ages of 35 and 55. Only the false negative to false positive ratio is increased both in the difference between the groups and their absolute values with 0.339 and 0.354 respectively for the group under 35 and the group between 35 and 55. This is undesirable as this means more false negatives than false positives than in the base model, while false negatives are less undesirable. However, the increase is from a very small value to a slightly larger value, but still well below the set range for equality.

Sensitive attribute - index of multiple deprivation The results for the index of multiple deprivation show nearly all the same trends as for the base model. These trends are shown in figure 5.9 and table 5.11. When using the fairness through unawareness model, the true positive rate becomes more equal than in the base model, satisfying the equal opportunity fairness definition. A trend now appears in the false negative to false positive ratio that increases with a higher index of multiple deprivation, making it not satisfy the treatment equality as it seems to increase for the middle range of the IMD index. It is also noteworthy that the accuracy across the different groups starts to vary much less than in the base model, especially with fewer groups with much higher accuracy. In terms of fairness, this could be seen as a positive evolution, but for the model itself, this is less positive. Table 5.10 shows the fairness definitions that the fairness through unawareness model satisfies and which it does not.

IMD range	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
Sample size	187	179	183	171	167	152	128	126	119	94
Accuracy	71.7%	76.0%	75.4%	78.4%	68.9%	73.0%	75.8%	72.2%	75.6%	77.6%
TN/TP	0.293	0.265	0.364	0.423	0.444	0.414	0.409	0.346	0.208	0.235

Table 5.11: Test set characteristics when split on the index of multiple deprivation in the fairness through unawareness

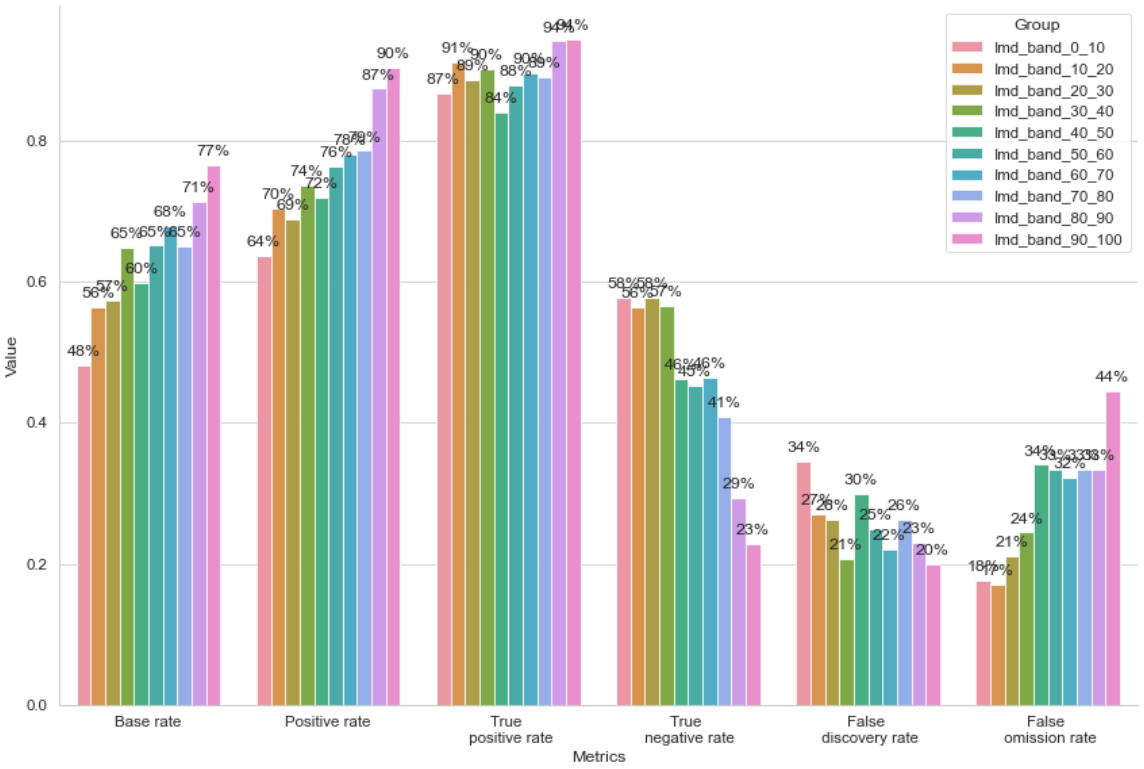


Figure 5.9: Statistical metrics of the fairness through unawareness model with regard to the sensitive attribute of the index of multiple deprivation

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	?	✗	✓
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	✗	✓	✗

Table 5.10: Fairness definition compliance of the fairness through unawareness model with regard to the sensitive attribute of the index of multiple deprivation

5.2.2 Suppression

Suppression is a pre-processing technique that builds further upon fairness through unawareness. The concept of suppression was already discussed in section 3.3.1. The implementation used for suppression is from the Fairlearn package [49]. More specifically, the class CorrelationRemover from the fairlearn.preprocessing package was used. Like fairness through unawareness, it removes the sensitive attributes from the data set and addresses residual correlation in the useful features with the sensitive attributes. One implementation of suppression is to remove the features that are highly correlated features with the sensitive attributes. This is not the technique used in this section.

The method used in this section will make use of linear regression in order to remove the correlation from all the remaining features with sensitive attributes. Removing this correlation results in a new set of features. The method introduces a new hyperparameter α. α determines how much of the correlation should be removed; if $\alpha = 1$ then all correlations will be removed so that only the new set of features will be used. In the case that $\alpha = 0$, the new set of features will not be used and rather the old set of features without the sensitive attributes are used. Any value for α in between 1 and 0 will create a weighted average for each feature based on the original and the new feature set.

Hyperparameter tuning can become tricky if a lot of the features are highly correlated with the sensitive attributes. In the case of the OULAD data set, this will be to a much lesser extent, so using an $\alpha = 1$ will work well. In the case of the SIMON-test, this can become more difficult as the features are more likely to be highly correlated with the sensitive attribute. Working with $\alpha = 1$ can remove too much information from the attributes. Therefore it is recommended to see if lower values of α would result in better performance with minimal fairness trade-offs.

Suppression is a pre-processing technique that is focused on the individual fairness definition. It is not designed to ensure any kind of group fairness, as will be clearly visible in the model evaluation. The idea for suppression is to decrease the dependency of the model on any sensitive attribute.

The second aspect that is important to remember is that suppression works with the sensitive attributes of the person. In order to use suppression in the pipeline of the AI system, these sensitive attributes need to be collected from everyone using the system. This is not a problem in certain applications, such as those used inside of an institution. However, in an application that is out in the open space, it might be more difficult to collect this personal information and not deter people because of it.

Model evaluation

The first somewhat surprising element when using the suppression technique in this instance, is the slight increase of 0.27 percentage points of the accuracy compared to when solely fairness through unawareness was used. With an accuracy of 74.57%, the suppression model is 0.53 percentage points lower than the base model, where the sensitive attributes were included as features. This is not surprising as the base model did have more features.

There is also a trend becoming clear in the statistical measures of the system. While the positive rate is constant between this model and the two previously discussed, other measures show a steady trend. The true positive rate dropped slightly when leaving out the sensitive attribute in fairness through unawareness and dropped even more when this form of suppression was applied. The same happened with the true negative rate. The opposite happened with the false discovery rate and the false omission rate. Here, the values steadily increased at first when using fairness through unawareness and continued when applying the suppression.

Sensitive attribute - disability The distribution of the statistical measure when split on the sensitive attribute of disability is shown in figure 5.10. From this figure, it can be determined whether or not a certain group's fairness definition is satisfied, as is noted in table 5.12. Compared to the model that only used fairness through unawareness, there are now two definitions, namely predictive parity and conditional use accuracy equality, where, in agreement with the user, it can be said that they are satisfied. The accuracy across both groups is much closer together when the split is made on disability, with the accuracy of the group of people without a disability being 74.6% and for the group of people with a disability becoming 73.9%. The similarity in accuracy from a gap of 4.6 pp in fairness through unawareness to only 0.76pp is especially an improvement. However, the gap is still larger compared to the base model. Accuracy is often at least one of the group fairness definitions that a model should satisfy, making it valuable. The false negative to false positive ratio increased, to much higher values for both groups, with 0.667 for people with a disability and 0.426 for people without a disability. The difference between both groups is also a lot higher than in fairness through unawareness, and it is also significantly higher compared to the base model. The statistical measures used in the group of fairness definitions not yet mentioned show very similar differences compared to fairness through unawareness. This is not surprising as suppression is not a technique that focuses on satisfying fairness through unawareness.

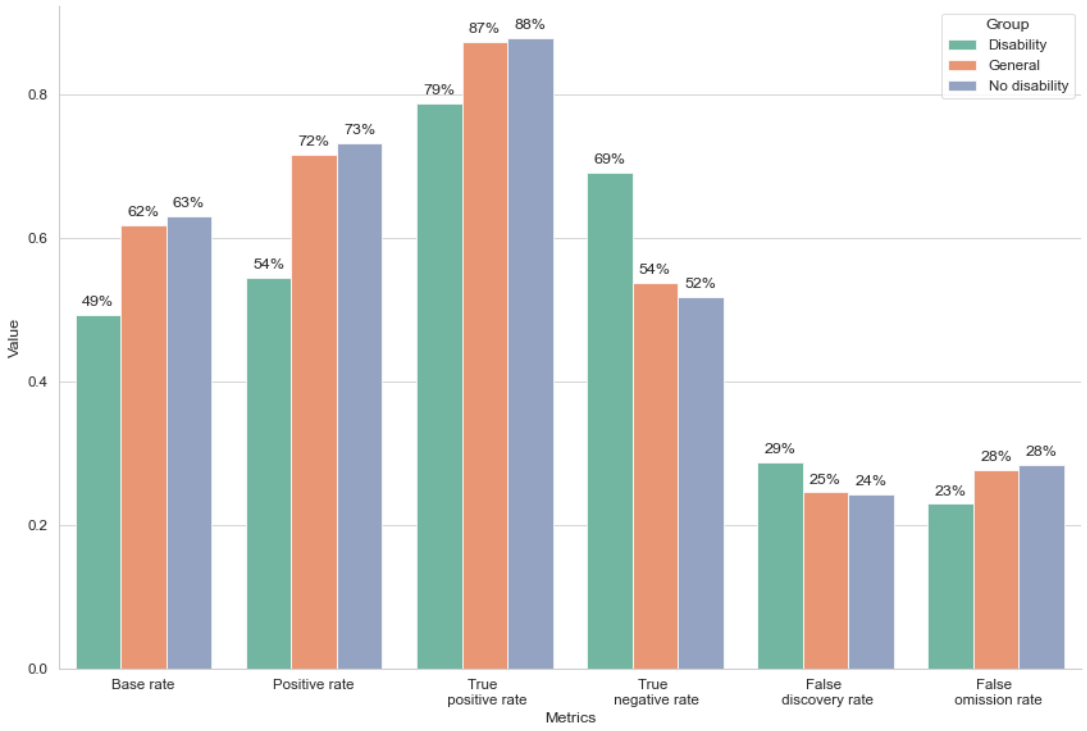


Figure 5.10: Statistical metrics of the model with suppression with regard to the sensitive attribute of disability

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	?	✗	✗
Difference	19pp	5pp	17pp	9pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	?	✓	✗
Difference	9pp - 17pp	5pp - 5pp	0.76pp	0.241

Table 5.12: Fairness definition compliance of the fairness through unawareness model with regard to the sensitive attribute of disability

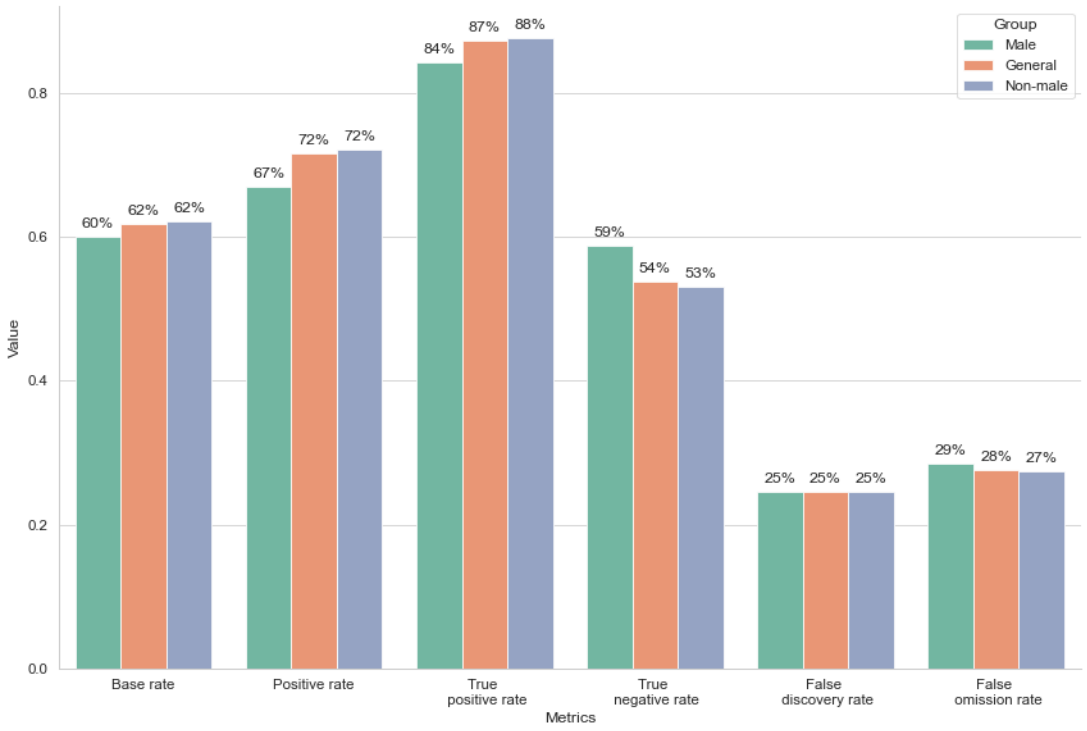


Figure 5.11: Statistical metrics of the model with suppression with regard to the sensitive attribute of gender

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	**?**	✓	✗	**?**
Difference	5pp	< 1pp	6pp	4pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	**?**	✓	✓	✗
Difference	4pp - 6pp	<1pp - 2pp	0.51pp	0.141

Table 5.13: Fairness definition compliance of the model with suppression with regard to the sensitive attribute of gender

Sensitive attribute - gender From figure 5.11 and table 5.13, the statistical measures and the compliance with the chosen group fairness definitions can be found when split on gender as a sensitive attribute and using the model with suppression. In this instance, fewer of the fairness definitions are satisfied, unlike when the split is made on disability, where more definitions were satisfied compared to the fairness through unawareness model. Positively the disparities between the different statistical measures are still fairly small and could be seen as acceptable in certain situations. Another positive factor when using suppression is the accuracy of both groups becoming more similar compared to both the base model and the fairness through unawareness models. The accuracy for the group that identifies as male becomes 74.1% and 74.6% for the group that identifies as not male which is an increase for both groups compared to the fairness through unawareness model. The difference and the values themselves for the false negative to false positive ratio increased significantly with the new values becoming 0.571 and 0.430 for the group that identifies as male and not male respectively. This is not beneficial as, in this use case, false negatives are far worse than false positives. When looking at the distribution itself, it is clear that there are more false negatives in the model than in the fairness through unawareness model, but also fewer false positives which on its own is positive, as that means higher accuracy.

Sensitive attribute - age range As shown in figure 5.12, there are larger and smaller disparities ranging from 13 to 2 percentage points between the different statistical fairness measures when split on the sensitive attribute of age. Again different behaviour can be seen in the group fairness definitions that are satisfied for the split made on age compared to the previously discussed splits, as noted in table 5.14. All the statistical measures are fairly similar to the fairness through unawareness model. The only fairness measures significantly different are the differences in the false omission rate, which is now just 6 percentage points making it low enough for conditional use accuracy to be possibly considered as satisfied. The same behaviour continues that the differences in accuracy lower further with 74.54% for the group of people under the age of 35 and 74.62% for people between the ages of 35 and 55. The important difference for treatment equality, namely the false negative to false positive ratio increased from 0.015 to 0.087. These are, however, still significantly small numbers to satisfy the definition. These values are, however, again increased compared to fairness through unawareness, with the ratio being 0.476 and 0.389 for the group under 35 and the group between 35 and 55 respectively.

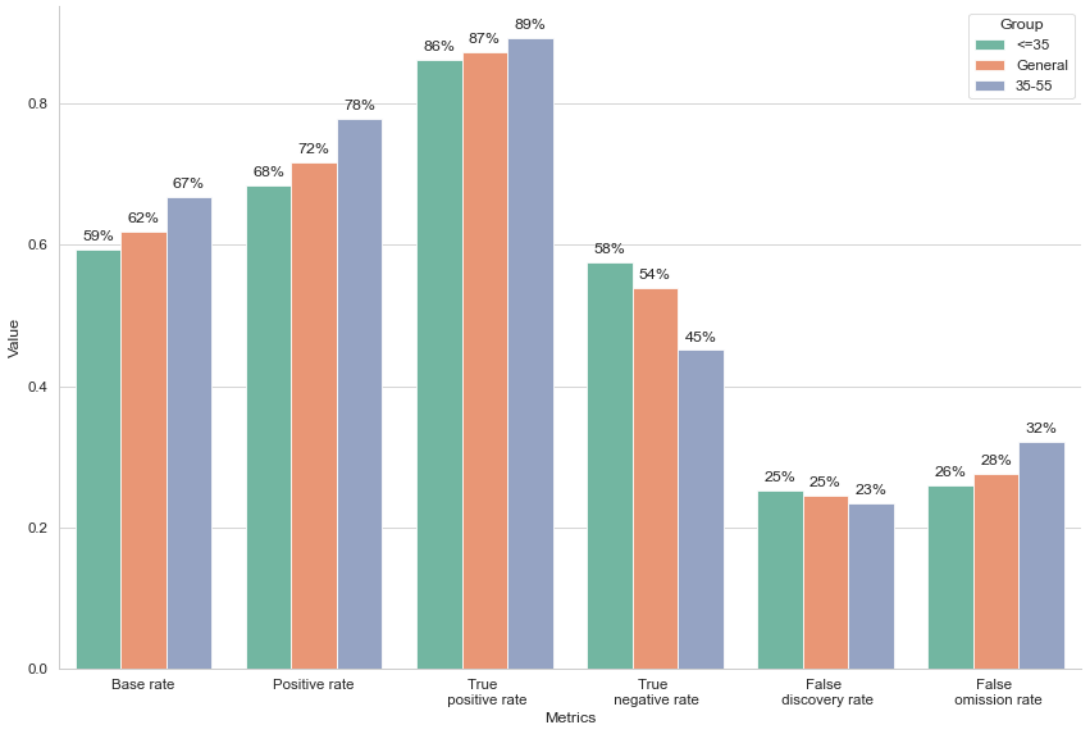


Figure 5.12: Statistical metrics of the model with suppression with regard to the sensitive attribute of age range

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	✓	✗	✓
Difference	10pp	2pp	13pp	3pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	?	✓	✓
Difference	3pp - 13pp	2pp - 6pp	0.08pp	0.087

Table 5.14: Fairness definition compliance of the model with suppression with regard to the sensitive attribute of age

Sensitive attribute - index of multiple deprivation The graph containing the statistical measures for the groups split on their index of multiple deprivation is shown in figure 5.13. The table containing the specific accuracy and false negative to false positive ratio for each group is noted in table 5.15. The fairness definitions that the suppression model satisfies and does not satisfy are included in table 5.16. The same trend continues for this split that the suppression model is not beneficial for the group fairness definitions, except for the accuracy. For the false negative to false positive ratio the unfairness increases in this instance as the ratio is a lot lower for people who live in a region with a higher index of multiple deprivation than people who live where it is lower.

IMD range	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
Sample size	187	179	183	171	167	152	128	126	119	94
Accuracy	71.1%	76.0%	76.5%	78.4%	68.3%	73.7%	77.3%	71.4%	75.6%	79.8%
TN/TP	0.688	0.483	0.593	0.423	0.559	0.379	0.381	0.385	0.208	0.118

Table 5.15: Test set characteristics when split on the index of multiple deprivation in the model with suppression

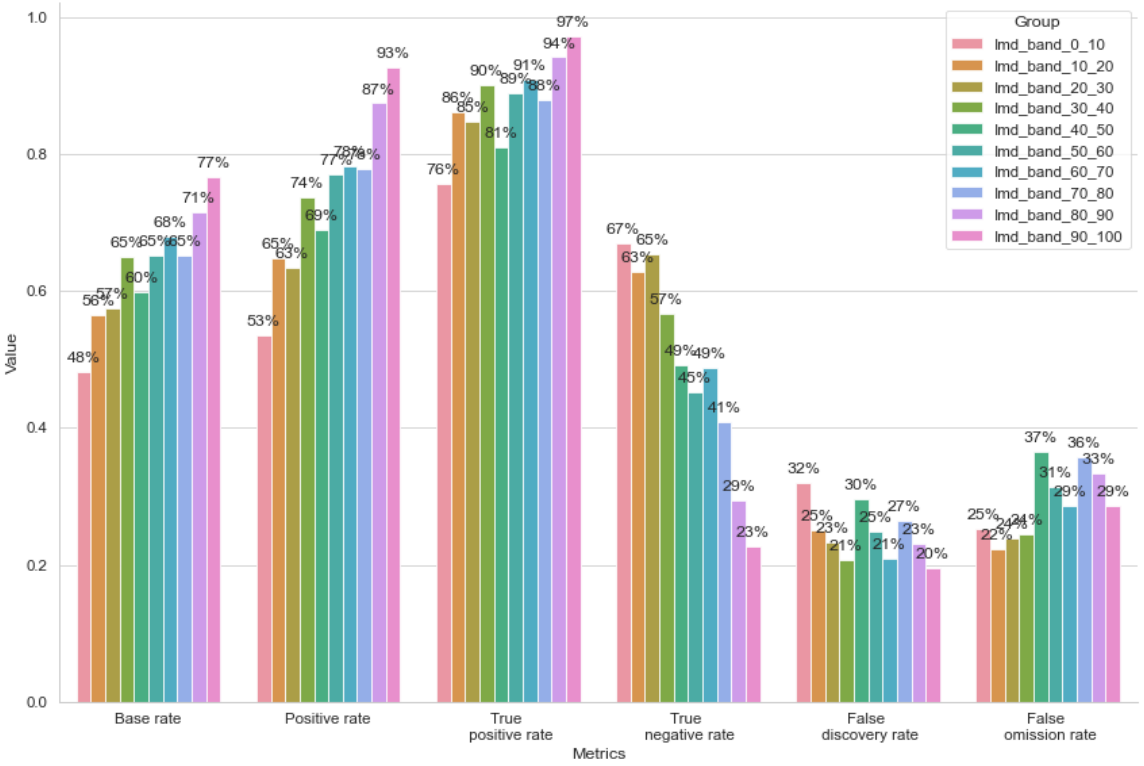
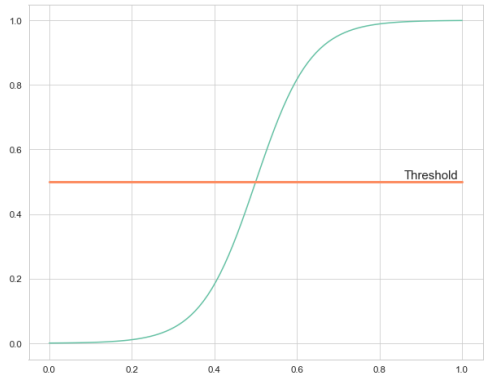


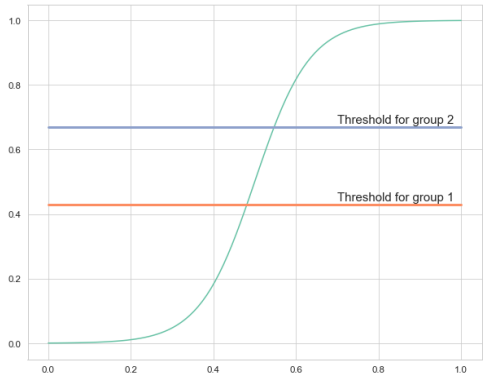
Figure 5.13: Statistical metrics of the model with suppression with regard to the sensitive attribute of the index of multiple deprivation

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	✗	✗	?
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	✗	✓	✗

Table 5.16: Fairness definition compliance of the model with suppression with regard to the sensitive attribute of the index of multiple deprivation



(a) Basic decision threshold in logistic regression



(b) Decision threshold in logistic regression optimised per group

Figure 5.14: Decision thresholds in logistic regression

5.2.3 Threshold Optimiser

The threshold optimiser is a very basic post-processing technique. This post-processing technique can work with a given fairness constraint (a statistical measure which should be held similar across groups). The threshold optimiser takes the result from a normal classification or regression model and then adjusts it according to the sensitive attributes of the sample.

In the case of classification the threshold optimiser shifts the decision threshold. In figure 5.14a the basic decision threshold in logistic regression can be found. The curve is fitted based on the training data. The features are the input x of the system from which the y-value in the range [0,1] for the sample is returned. If the y-value is higher than the threshold, the prediction will be 1; otherwise, it will be 0. In basic logistic regression this threshold, normally at 0.5, is set for all groups equally. In the threshold optimiser this threshold is split into one threshold for each group. These thresholds can be placed at different values.

The main consequence of these thresholds per group is that causal discrimination is no longer prevented even when removing the sensitive attributes from the feature set. These sensitive attributes will determine the group in which a sample will fall, thus directly influencing the y-value and, consequently, the prediction. The sensitive features can be added to the feature set since causal discrimination will not be prevented by not including them in the feature set. Sometimes including these features can increase the fairness of the model. In order for the pre-processing technique to function it is necessary for this information about a person's sensitive attributes to be collected.

When using the threshold optimiser it is necessary to know which fairness definition needs to be satisfied. The threshold optimiser will work with the constraint of keeping the statistical measure coupled to the fairness definition equal across the different groups. Secondary to the fairness definition, it can also be possible to choose what metric the system tries to optimise. This cannot conflict with the chosen fairness definition, but it can further increase equality.

The implementation of the threshold optimiser used is from the open-source package Fairlearn [49]. It accepts 3 different types of fairness definitions. The first one is statistical parity so that the model will ensure a similar positive rate across the different groups. The other group of possible statistical measures are the true/false positive/negative rate. The last possibility is satisfying equalised odds, which simply means satisfying both the true positive and true false rates.

The objective as that the threshold optimiser is optimising an accuracy or a rate. The accuracy could be simply the accuracy on the entire test set, but it could also be chosen to use the balanced accuracy. Balanced accuracy is the average of the accuracy of the different groups defined by the sensitive attributes. This differs from the normal accuracy because, in this case, it would be a weighted average between the groups; with balanced accuracy all groups are equal in the calculation. Another possibility is maximising the overall positive rate, true positive rate or true negative rate. It is important to match the objective and the constraint because otherwise, certain trivial predictors, such as only predicting the negative class, could result in the solution.

A problem when using this implementation of the threshold optimiser is the definitions of the groups. The groups are created by combining different sensitive attributes. This means that if two samples differ in one sensitive attribute, they will be divided into different groups. Because the data set used in this dissertation is fairly largely split, this split will result in fairly small groups making the operation far more difficult. In the implementation it was even to this extent that the IMD band could not be included as a sensitive attribute in order not to get a trivial predictor.

Model analysis

Due to the nature of the technique of threshold optimisation it is necessary to control for causal discrimination. These showed between 15 and 21 instances of causal discrimination depending on what sensitive attribute was controlled. This means that in around 1% of instances causal discrimination occurs. If this application was used in a setting where it was purely informative for the user and about that same user

then this causal discrimination could be accepted. However, if this system were to be used as a factor in a decision pipeline, this causal discrimination is not acceptable and perhaps grounds for a lawsuit as the difference in the decision is based on a sensitive attribute.

The second element investigated was whether or not it would be beneficial to include the sensitive attributes as features in the system. This did increase the overall accuracy by 0.7 percentage points compared to when they were not included in the feature set, leading to an overall accuracy of 74.9%. It even improved the accuracy slightly from when fairness through unawareness or the model with suppression was used. Another bonus of adding the sensitive attributes to the feature set is that the occurrence of causal discrimination was lowered. This is one very important aspect which should always be kept as low as possible. Overall it is only beneficial to include these sensitive attributes as features in this technique.

Perhaps due to the small group sizes, it often happened that certain configurations of the objective and constraint led to trivial predictors. This is something to be on the lookout for with these types of post-processing techniques. Eventually, it was chosen for accuracy as the objective to maximise and to satisfy the true positive ratio for the equal opportunity fairness definition, which is the most crucial in the imaginary context for the application.

Sensitive attribute - disability The distribution of the statistical measures when split on the sensitive attribute of disability can be seen in figure 5.15. From this figure, it can be determined whether or not a certain group's fairness definition is satisfied as is noted in table 5.17. Although the threshold optimiser was set to ensure an equal true positive rate across the groups, it is not equal between the group of people with a disability and the group of people without a disability. This is most likely because the groups for which this constraint must be satisfied are a combination of the sensitive attributes. On the positive side, the difference of 6 percentage points is the lowest difference from all the models yet tested when the split is made on disability. This, coupled with the lowest true negative rate seen for all the models, improved fairness when looking at the definitions for predictive equality, equal opportunity and the combination of them both in equalised odds. The accuracy difference is the largest compared to the previously discussed models. The accuracy for the group of people with a disability drops to 68.7%, while the accuracy of the group of people who do not have a disability stays fairly the same at 75.4%. An advantage of tuning the true positive rate is that the false negative to false positive ratio also becomes fairly similar between the groups. This with 0.323 and 0.284 for the group with a disability and the group with a disability, respectively.

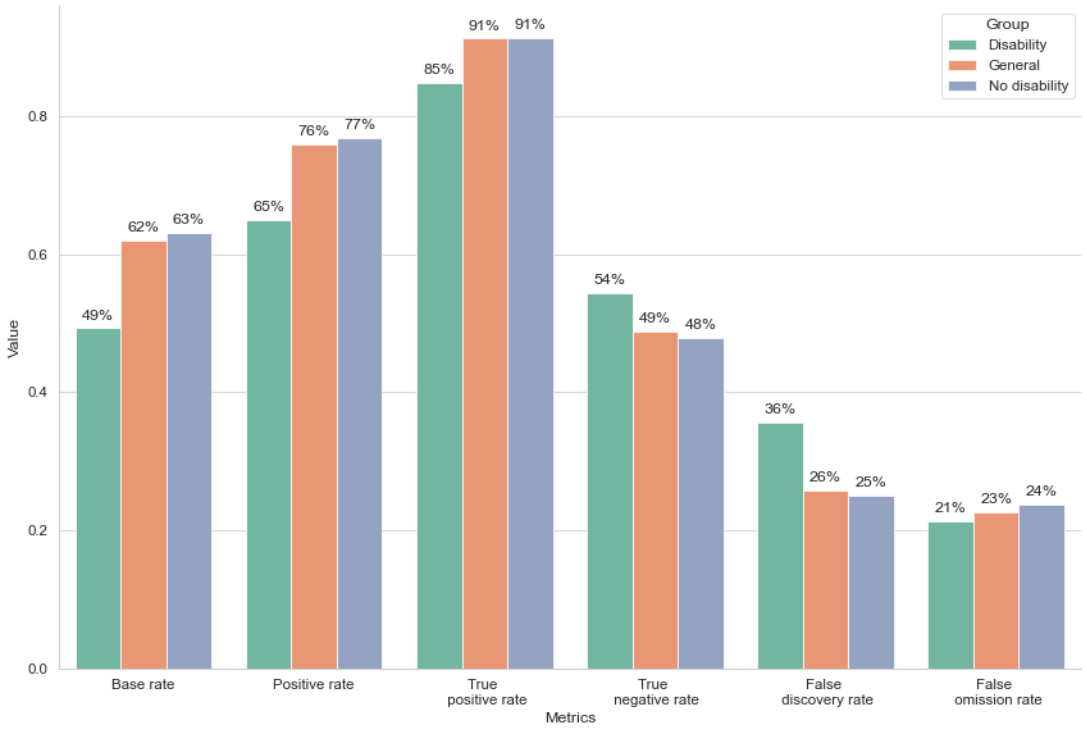


Figure 5.15: Statistical metrics of the threshold optimisation model with regard to the sensitive attribute of disability

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	✗	✗	✗
Difference	12pp	11pp	6pp	6pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	?	✗	✗	✓
Difference	6pp - 6pp	11pp - 3pp	6.7pp	0.039

Table 5.17: Fairness definition compliance of the threshold optimisation model with regard to the sensitive attribute of disability

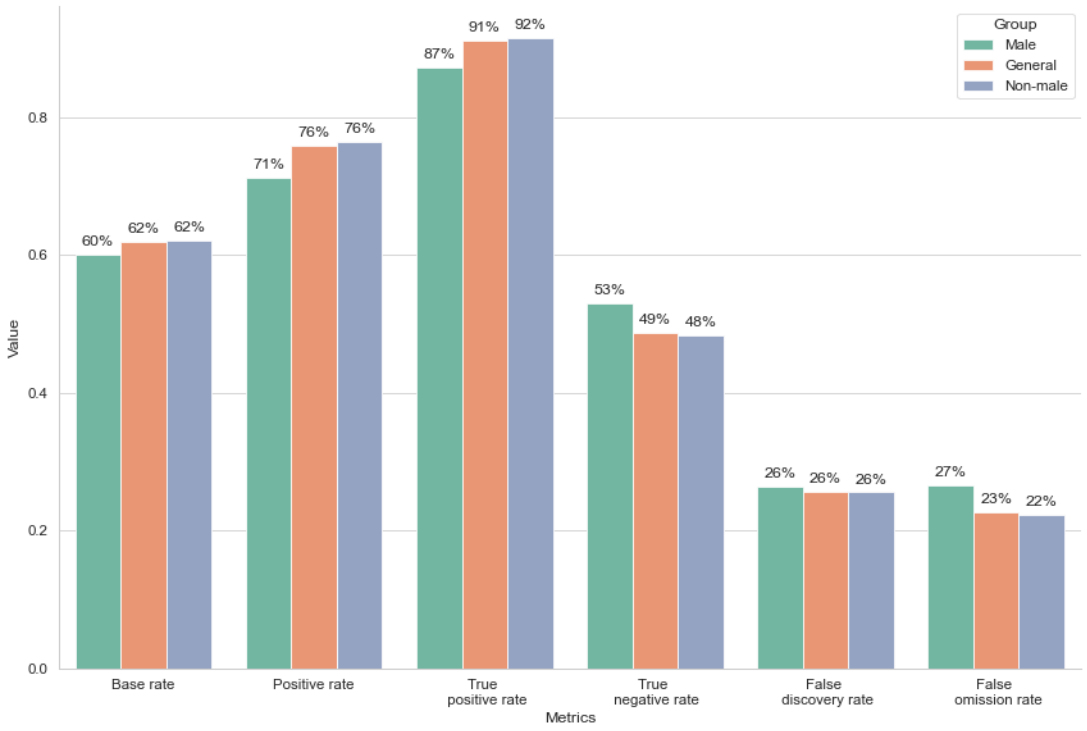


Figure 5.16: Statistical metrics of the threshold optimisation model with regard to the sensitive attribute of gender

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	**?**	✓	**?**	**?**
Difference	5pp	< 1pp	5pp	5pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	**?**	✓	✗	✗
Difference	5pp - 5pp	<1pp - 3pp	6.55pp	0.138

Table 5.18: Fairness definition compliance of the threshold optimisation model with regard to the sensitive attribute of gender

Sensitive attribute - gender From figure 5.16 and table 5.18, the statistical measures and the compliance with the chosen group fairness definitions can be found when split on gender as a sensitive attribute and using threshold optimisation. The use of the threshold optimiser was not beneficial for satisfying the accuracy constraints when the split is made on gender. This is most likely because the base rates are very similar between these groups, and the technique works on much smaller groups. This means that when using this technique, this group split needs to include some compromises. However, this is still relative as most fairness definitions can be considered satisfied. Just like the split on disability, the accuracy difference takes a big hit. The accuracy for people who identify as male dips down strongly to 68.8%, while the accuracy for the group that identifies as not male increases ever so slightly compared to the past models achieving 75.4%. The false negative to false positive rates for this split is 0.406 and 0.268 for the groups that identify as male and not male, respectively. This creates a large discrepancy between both groups. The ratio is more desirable than the values returned from the model with suppression, but it is far from a positive evolution compared to the other models discussed.

Sensitive attribute - age range As shown in figure 5.17, there are larger and smaller disparities between the statistical measures ranging from 12 to 1 percentage points when split on the sensitive attribute of age. Like disability, the disparities between the statistical measures like true positive and true negative rates were lowered. Especially beneficial is the very low true positive as the threshold optimiser used as a constraint. As shown in table 5.19, this leads to equal opportunity to be satisfied, which is one of the best definitions to satisfy in this context. As has been the trend when using age as the split, the accuracy between both groups remains very similar. When using the threshold optimiser the accuracy for the group under the age of 35 becomes 75.1% and 74.4% for the group between the ages 35 and 55. The difference in false negative to false positive ratio is close to the threshold of 0.100 in order to be considered close with 0.262 and 0.367 for the group under 35 and from 35 to 55, respectively. However, this is the largest difference in this ratio for all the previously discussed models.

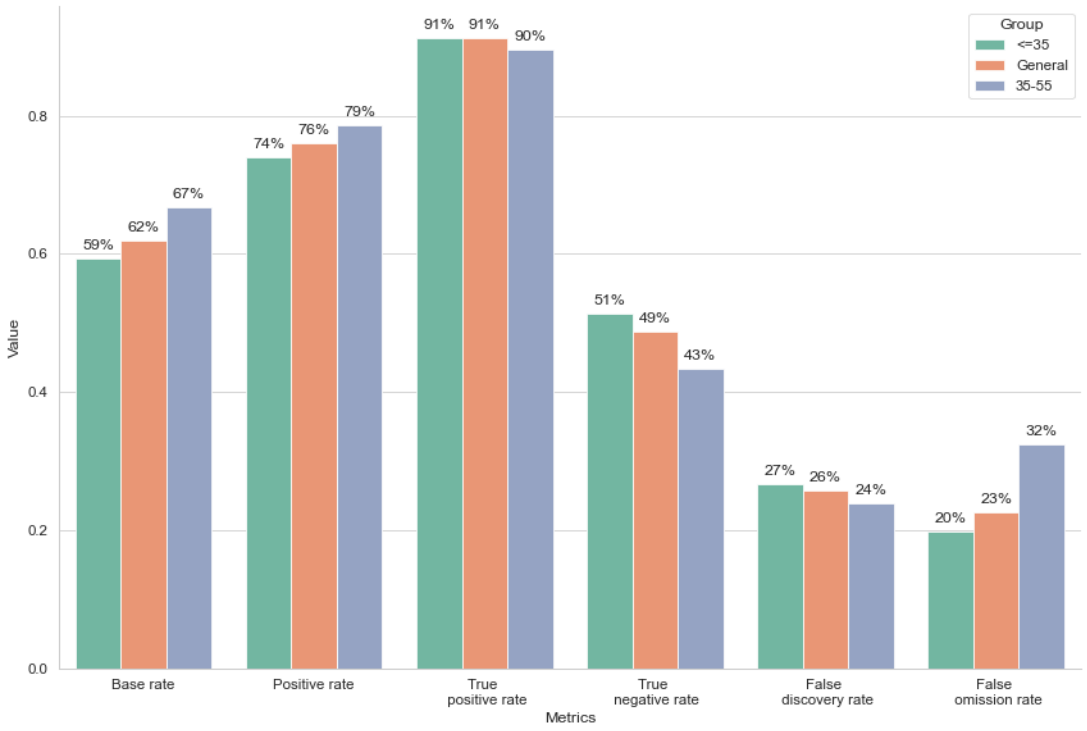


Figure 5.17: Statistical metrics of the threshold optimisation model with regard to the sensitive attribute of age range

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	**?**	✓	✗	✓
Difference	5pp	3pp	8pp	1pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	✗	✓	✗
Difference	1pp - 8pp	3pp - 12pp	0.628pp	0.105

Table 5.19: Fairness definition compliance of the threshold optimisation model with regard to the sensitive attribute of age

Sensitive attribute - index of multiple deprivation The graph containing the statistical measures for the groups split on their index of multiple deprivation is shown in figure 5.18. The table containing the specific accuracy and false negative to false positive ratio for each group is noted in table 5.20. The fairness definitions that the threshold optimiser model satisfies and does not satisfy are included in table 5.21. It is important to note that the index of multiple deprivation was not used to determine the groups used in the threshold optimiser. This made them too small and the problem impossible without a trivial classifier. The true positive rate varies slightly across the different IMD ranges, so it is up to the implementer to determine if this is sufficiently close to be deemed it fair. The conditional use accuracy equality is satisfied for the first time, making this an advantage if this technique were used. However this was deemed a less important fairness definition in the context. The accuracy varies between 70.7% and 80.9% for the different groups. These are higher than the previous models, which is also an important factor in determining fairness. However, the false negative to false positive ratio is not fair across the groups. The lower the index of multiple deprivation the higher this ratio seems to be.

IMD range	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
Sample size	187	179	183	171	167	152	128	126	119	94
Accuracy	73.8%	76.0%	75.4%	77.2%	70.7%	73.7%	75.0%	72.2%	74.8%	80.9%
TN/TP	0.382	0.333	0.294	0.44	0.371	0.242	0.240	0.129	0.111	0.059

Table 5.20: Test set characteristics when split on the index of multiple deprivation in the threshold optimisation model

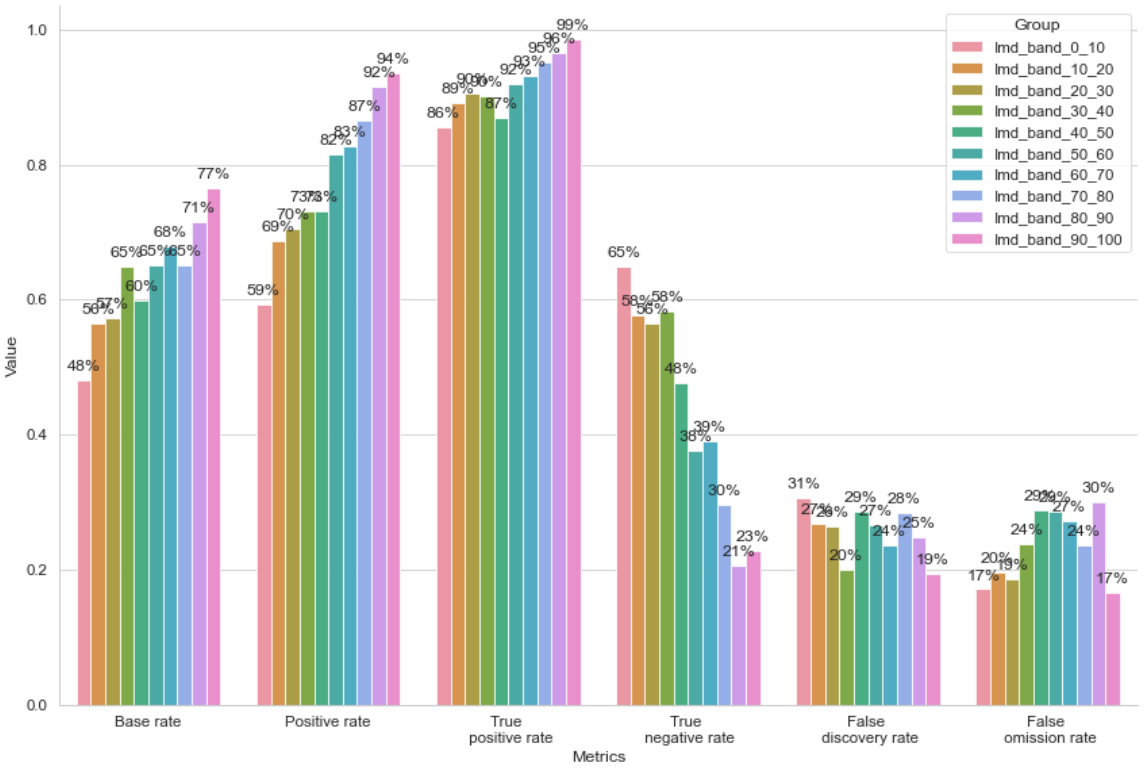


Figure 5.18: Statistical metrics of the threshold optimisation model with regard to the sensitive attribute of the index of multiple deprivation

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	✓	✗	?
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	✓	✓	✗

Table 5.21: Fairness definition compliance of the threshold optimisation model with regard to the sensitive attribute of the index of multiple deprivation

5.2.4 Adjusting the loss function

Adjusting the loss function is a very simple form of an in-processing technique. The goal of machine learning algorithms is to minimise a certain loss function. In classification, the inputs of this loss function are the predicted labels on the training set and the actual labels of the training set. The more labels are predicted incorrectly, the higher the loss. Sometimes a loss function can work even with the probabilities of how certain a model is of its prediction. This means that optimising the loss means minimising the number of incorrect predictions and becoming less certain in the cases of incorrect predictions.

In nearly all implementations of the loss function the loss function itself is impartial to the mistake that has been made, whether it was mistaking a negative for a positive or the other way around. However this is not the case in most fairness applications, where one type of incorrect prediction has a worse result than the other way around. As was discussed in this case; a false negative is far worse for the person than a false positive. This is something that should be conveyed into the system.

This could be done in two ways. The first possibility would be writing a hardcoded custom loss function to increase the importance of certain mistakes compared to others. This might seem a good method at first; however, the study of loss functions is vast and complicated. In most cases it is unnecessary to go to the lengths of creating a custom loss function for this purpose. The second method is simpler, but imposes some constraints. In many applications of loss functions it is possible to give a set of weights to the function itself. With these weights the importance of certain samples can be conveyed. In this situation, conveying that one mistake is worse than another can be done by increasing the weights of those samples that, when incorrectly predicted, leads to worse outcomes. In the case used here, this would mean increasing the weights of the samples that have a ground truth of positive. Incorrectly predicting those as false negatives is precisely the occurrence that should be avoided. Using this method with the adjusted weights in the loss function requires, first of all, that the implementation of the loss function accepts custom weights. The second constraint is that only the weights can be adjusted, so creating a real fairness definition is not possible. As a result, it is only possible to determine which samples are crucial to classify correctly.

After determining which samples need to have different weights, it is important to determine the value of the weights. The higher the weight the more the model changes to try and correct these samples correctly. The weights must be tuned while keeping the model's accuracy in mind. If the weights were set too large, the model might become trivial and only return positive predictions. The effect of updating the model

weights is in a way similar to the threshold optimiser technique. Changing the threshold optimiser shifted the decision boundary. While changing the weights of the samples for the loss function will change the output function on which the threshold is applied for the classification. However, unlike the threshold optimiser, changing the weights for the loss function works on the internals of the logistic regression and requires access to the model itself.

Model analysis

It is not possible to use the same implementation of logistic regression as was used for the previous models, as it is difficult to access the loss function. In order to solve this problem, the python library pytorch was used with the Binary Cross Entropy loss function. However, when using the implementation with the basic weight values, the accuracy on the test set was already slightly lower. This can be due to the difference in loss function used or simply the implementation.

In order to use this technique, a good weight for the samples of interest must be determined. The samples, in this case, are the samples which should be classified as positive. This is done in order to minimise the number of false negatives, which is the worst result possible. The decision was made to set the weight of these samples to three and the other sample weight for the loss function to one. The implementation for this technique follows fairness through unawareness in order to ensure that no causal discrimination occurs. When using the weights set to three, then the test set accuracy increases slightly from 74.3% to 74.4%. This increase is fairly insignificant given the size of the test. However, it does mean that no real harm has come to the model's capabilities by setting slightly different priorities.

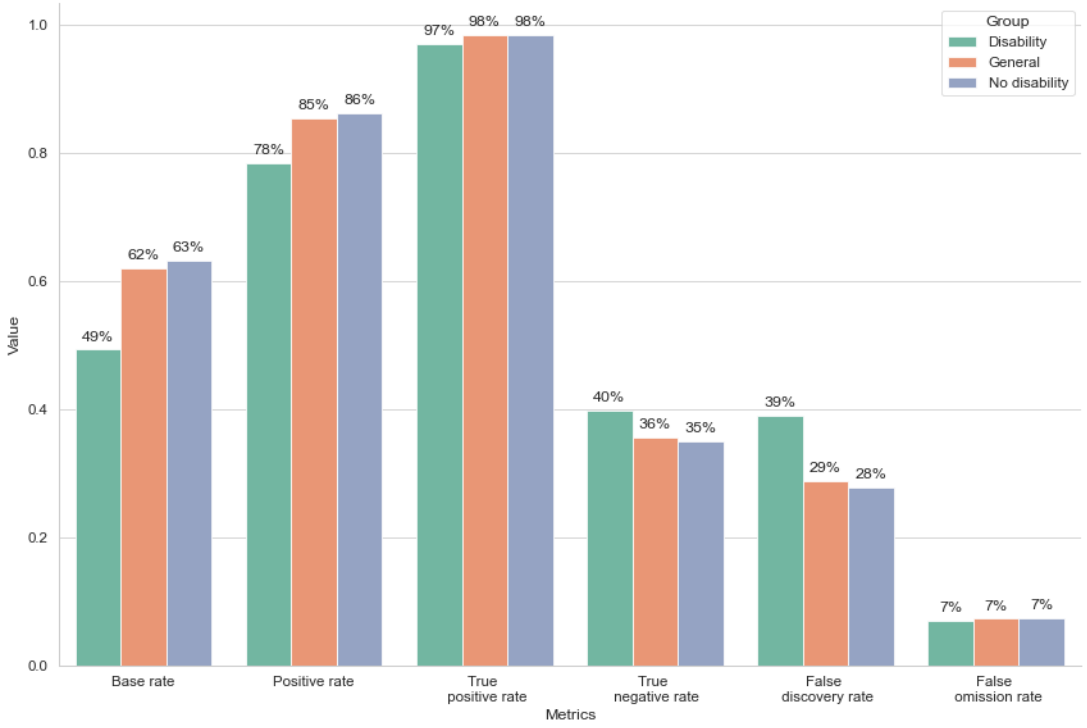


Figure 5.19: Statistical metrics of the model with a custom loss function with regard to the sensitive attribute of disability

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	✗	?	✓
Difference	8pp	11pp	5pp	1pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	?	✗	✗	✓
Difference	1pp - 5pp	11pp - <1pp	7.1pp	0.0061

Table 5.22: Fairness definition compliance of the model with a custom loss function with regard to the sensitive attribute of disability

Sensitive attribute - disability The distribution of the statistical measures when split on the sensitive attribute of disability can be seen in figure 5.19. From this figure, it can be determined whether or not a certain group fairness definition is satisfied as is noted in table 5.22. From previous models it was clear that the split on disability was the most difficult to satisfy. However, when using this technique of adjusting the weights in the loss model the most important fairness definition is satisfied, being equal opportunity. Given the accompanying low difference in a true negative then it is reasonable to conclude after consulting with the implementer that equalised odds is met. An important problem is the difference in accuracy, with the group of people with a disability having an accuracy of 67.9% and the group of people without a disability having 75.0%. This can be put into perspective as the false negative to false positive ratios are 0.0487 and 0.0426 for the group of people with a disability and without a disability respectively. These low values indicate that most of the incorrectly classified samples were false positives, which are not the worst to occur. The large difference in accuracy is most likely due to the base rates, as this model will favour classifying as positive, but the group of people with a disability has fewer positive samples than the other group.

Sensitive attribute - gender From figure 5.20 and table 5.23, the statistical measures and the compliance with the chosen group fairness definitions can be found when split on gender as a sensitive attribute and using the adjusted loss function. Similar to the other models, the statistical measures across the groups split on gender are very similar. The only value slightly high is the true negative rate; this is not surprising as adjusting the weights of the samples for the loss function made correctly classifying samples with the ground truth less important. However, it is important to note that the difference of 5 percentage points could still be considered acceptable. One advantage of this technique is the smallest difference in accuracy and false negative to false positive ratio as of yet. The result is fairly good, with an accuracy of 74.1% and 74.4% for the group of people who identify as male and not male. These are both fairly high in accuracy and also similar which are both strong advantages. The false negative to false positive ratios are very low as can be expected with the choices made. The group of people who identify as male has a false negative to false positive ratio of 0.0731 and those who do not identify as male have a ratio of 0.0395.

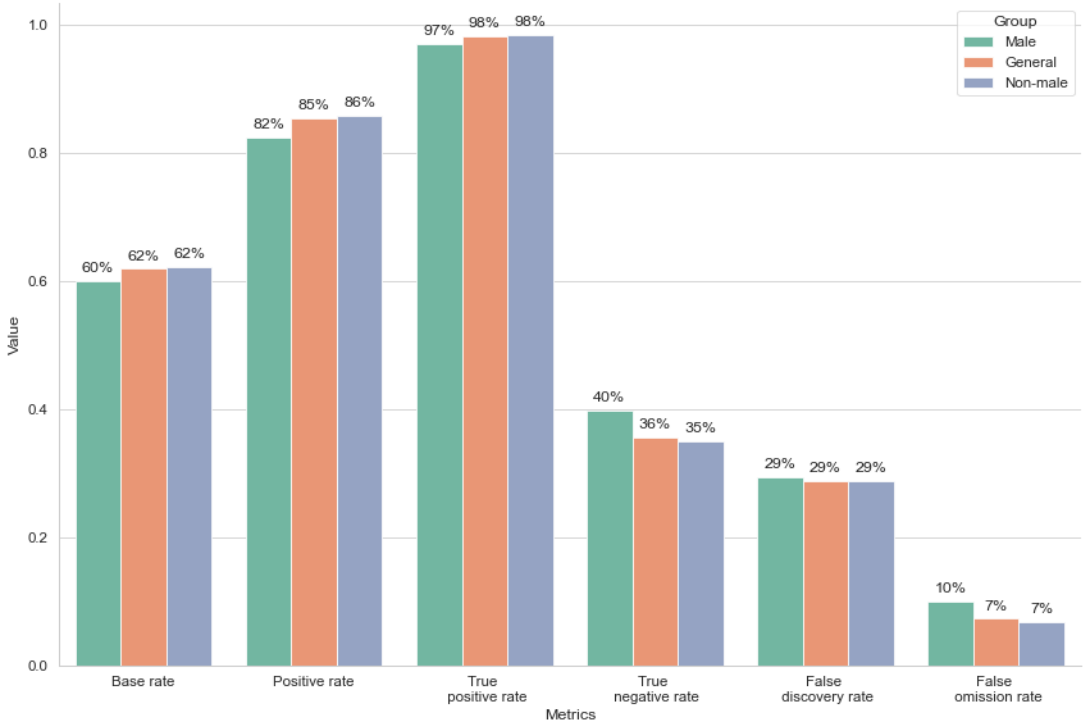


Figure 5.20: Statistical metrics of the model with a custom loss function with regard to the sensitive attribute of gender

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✓	✓	?	✓
Difference	4pp	< 1pp	5pp	1pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✓	✓	✓	✓
Difference	1pp - 5pp	<1pp - 5pp	0.28pp	0.034

Table 5.23: Fairness definition compliance of the model with a custom loss function with regard to the sensitive attribute of gender

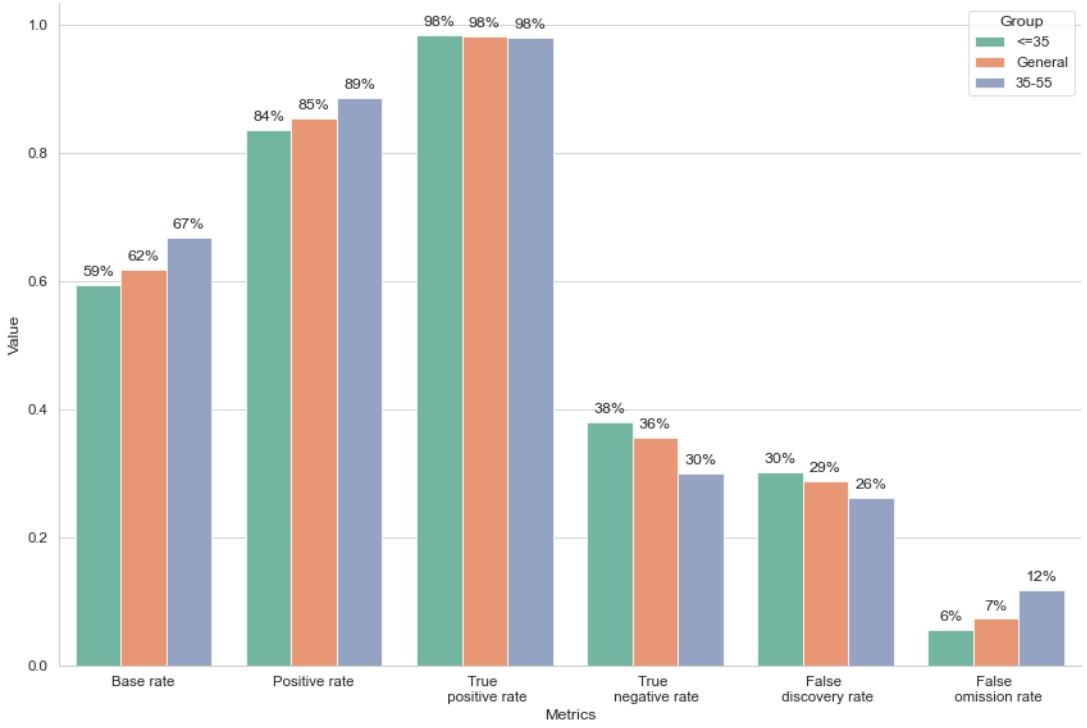


Figure 5.21: Statistical metrics of the model with a custom loss function with regard to the sensitive attribute of age range

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	?	✓	✗	✓
Difference	5pp	4pp	8pp	<1pp
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	?	✗	✓
Difference	1pp - 8pp	3pp - 6pp	1.55pp	0.0218

Table 5.24: Fairness definition compliance of the model with a custom loss function with regard to the sensitive attribute of age

Sensitive attribute - age range As can be seen in figure 5.21, there are somewhat larger and smaller disparities between the statistical measures ranging from 8 to less than 1 percentage points when split on the sensitive attribute of age. Especially beneficial in these metrics is the very low true positive difference, as can be seen in table 5.24. This leads to equal opportunity to be satisfied, which is one of the best definitions to satisfy in this context. The difference in true negative rate is still somewhat too high, leading to equalised odds, the best possible fairness definition to satisfy, not being achieved. There is somewhat a difference in the accuracy; however, it seems small compared to the difference in accuracy when the split is made on disability. The accuracy of both groups is still rather acceptable, with 73.8% for the group under the age of 35 and 75.4% for the group of people between the ages of 35 and 55. The false negative to false positive ratio is both small in value and difference between the groups with 0.0361 and 0.0579 for the group of people under the age of 35 and people between 35 and 55 respectively.

Sensitive attribute - index of multiple deprivation The graph containing the statistical measures for the groups split on their index of multiple deprivation is shown in figure 5.22. The table containing the specific accuracy and false negative to false positive ratio for each group is noted in table 5.25. The fairness definitions that the threshold optimiser model satisfies and does not satisfy are included in table 5.26. The technique of adjusting the weights for the loss function had a beneficial outcome when the split was made on the index of multiple deprivation. The true positive rate is very similar for all groups, with the biggest difference being 3 percentage points. This means that this implementation of the model satisfies equal opportunity. The true negative rate is not very equal; however, it might show a different story if the subgroups were somewhat larger. Lastly, the false negative to false positive ratios are fairly similar and small, as seen with this model. The highest percentiles of the index of multiple deprivation even have the ratio at 0. The other groups have almost the same values meaning that the model satisfies treatment equality.

IMD range	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
Sample size	187	179	183	171	167	152	128	126	119	94
Accuracy	66.3%	74.3%	73.8%	77.2%	73.7%	74.3%	78.1%	73.8%	75.6%	81.9%
TN/TP	0.050	0.045	0.043	0.054	0.047	0.054	0.037	0.065	0.000	0.000

Table 5.25: Test set characteristics when split on the index of multiple deprivation in the model with a custom loss function

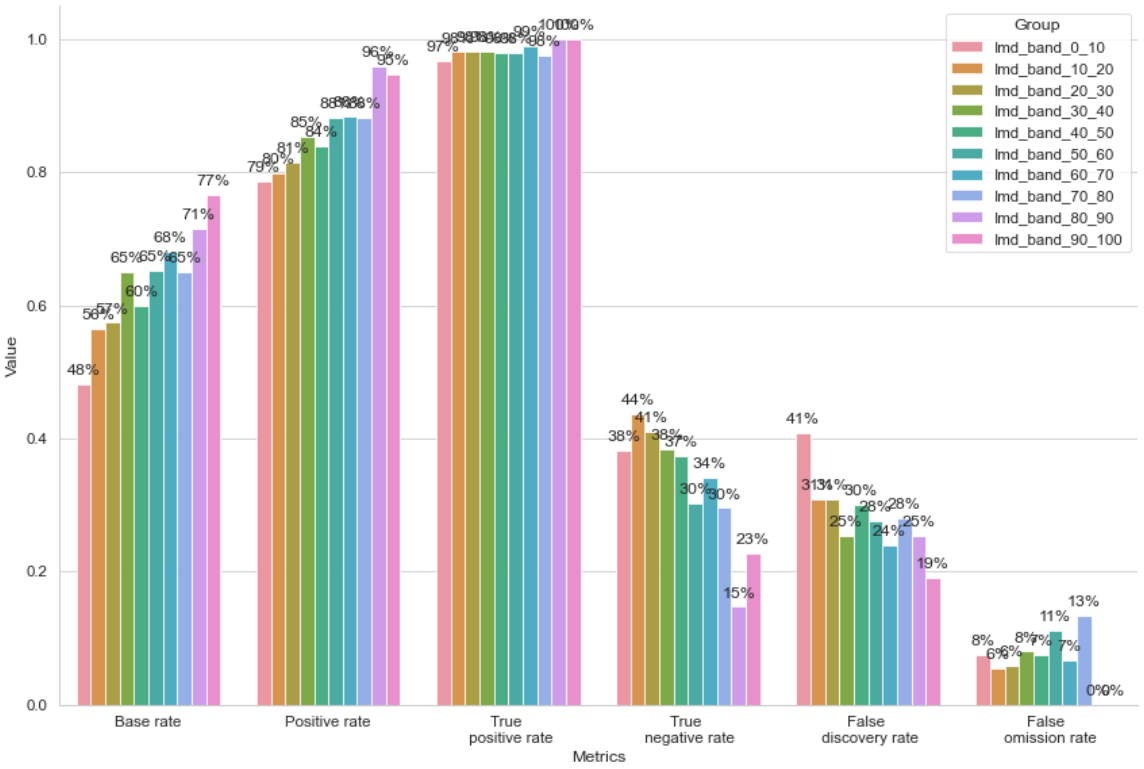


Figure 5.22: Statistical metrics of the model with a custom loss function with regard to the sensitive attribute of the index of multiple deprivation

Fairness definition	**Statistical parity**	**Predictive parity**	**Predictive equality**	**Equal opportunity**
Statistical measure	Positive rate	False Discovery rate	True Negative rate	True Positive rate
Satisfied	✗	✓	✗	✓
Fairness definition	**Equalised odds**	**Conditional use accuracy equality**	**Overall accuracy equality**	**Treatment equality**
Statistical measure	True Positive rate True Negative rate	False Discovery rate False Omission rate	Accuracy	False negative to False positive ratio
Satisfied	✗	✓	✓	✓

Table 5.26: Fairness definition compliance of the model with a custom loss function with regard to the sensitive attribute of the index of multiple deprivation

5.3 Summary

In this chapter, different techniques were discussed concerning increasing the fairness of a certain AI system. In order to know the effect of the different techniques, a baseline must be established. Two possibilities for a baseline exist, using all the information including the sensitive attributes as features or not using the sensitive attributes in the feature set to prevent causal discrimination.

The most difficult groups to ensure some kind of fairness between them are groups where the base rate differs strongly, as was discussed in section 3.3.3. It is also important to note that when changing the model to increase the fairness between two types of groups, groups based on a different split might see a decrease in fairness. This means that trying to tune an AI system to be fair for all sensitive attributes can be very difficult to achieve. In order to make this tuning less difficult, a certain range can be set, which constitutes the allowed difference between statistical measures and still be considered fair.

The different techniques discussed all work on different levels and thus their usability depends on the developer's access to the AI system itself. In-processing techniques can be very powerful, as could be seen in section 5.2.4. They convey what is important to the system itself, even when it is a very simple technique like adjusting the weights of the samples in the loss function. On the other hand, the pre-processing and post-processing techniques work differently because they simply change the properties of the data or the outcome in order to increase fairness. This makes these types of techniques more volatile to the data itself and tuning them more dependent on the quality of the data.

Something very important when determining what technique to use is the data that will be available. Both the suppression and threshold optimiser were dependent on the sensitive attributes of the sample in testing. This means that in order to use the AI system the sensitive attributes must be collected. This might make it more difficult to persuade people to use the application as certain sensitive attributes such as socioeconomic background are not information that people like to share. Tied to the sensitive attributes, it is also important to verify if causal discrimination could occur as a result of the technique. This is still very dangerous behaviour to occur in a model and should thus be checked if present. Of course, the systems that are agnostic to these sensitive attributes will not have any causal discrimination.

Chapter 6

Monitoring bias

An AI system which requires fairness cannot simply be put into production and left alone. There are two reasons why monitoring is important after an AI system which requires fairness starts being used. The first reason is universal for all AI systems and these relate to the bias of the training set, as discussed in chapter 4. The second reason arises from the fact that fairness is a social concept, and the social landscape can influence the use of the system.

In the white paper of the World Economic Forum, *How to Prevent Discriminatory Outcomes in Machine Learning* [16], this sentiment of monitoring the system's bias is encapsulated in the principle of Access to Redress. This means that the designers and developers of machine learning systems are responsible for the use and actions of their systems. Their responsibility is to try and redress those affected by disparate impacts and establish processes for the timely redress of any discriminatory outputs.

The following sections will focus on the possible causes of bias arising that were not present during the creation of the AI system and the evaluation necessary to ensure that the AI system can still be considered fair.

6.1 Possible arising biases

The first biases which will be discussed are those simply inherent to the data itself; these are not necessarily directly related to the fairness aspect, but those biases can influence the fairness and lead to disparate impacts. Next, the following biases will be discussed on how they can arise after an AI system after it is brought into use: Historical bias, Temporal bias, Representation bias, Evaluation bias, Deployment bias and Emergent bias.

Historical bias (Definition 4.1.2) can only occur when the system has been up for quite some time and was never retrained. Historical bias depends on the biases of society getting into the data set. This bias only occurs during the lifetime of the application, if society has changed during that period and the model was not adjusted for it. This means that the model itself needs to be quite old in order for it to be outdated with the current society and that it did not undergo any updates to mitigate it.

Temporal bias (Definition 4.1.1) is the most obvious form of bias to occur after a system is brought into production. Temporal bias occurs when there are behavioural differences over time and for those differences to be systematic across groups. This is different from historical bias as the behaviour in question is on the platform on which the AI system operates and can thus happen much quicker.

Representation bias (Definition 4.1.6) and Evaluation bias (Definition 4.1.10) can be grouped in this section as the bias is introduced in precisely the same way but influences the model at different points. One possibility is that the user population changed over time, making the training and test set used no longer representative. Another possibility would be that the training set and test set are updated with the data gathered during the use of the system. However, this technique brings with it the danger of under-representing people who would receive a negative. This is due to the fact that those people are a lot less likely to be included in the data set. For example if the AI system predicts if someone were to repay their loan, then if the AI predicts that someone would not be able to repay their loan, they would not get a loan. This means that that person cannot be included in the data set as the ground truth remains unknown. A possibility would be that the system would systematically make a suboptimal decision in order to gather more data for the system to learn from and evaluate it [23]. However, this would mean "sacrificing" some people in order to gather more data, meaning that some people who would receive a negative prediction will receive a positive prediction as to gather the data. This raises an ethical debate prior to starting to use the AI system.

Deployment bias (Definition 4.1.12) is a type of bias that can occur fairly quickly inside an AI system. The moment the system goes into production the control moves away from the developer, leading to it being used in uncontrolled settings. When the AI system is not used in the intended fashion, unfairness can arise due to this mismatch. The article *Fairness and Abstraction in Sociotechnical Systems* [9] introduces the Ripple Effect which explains how implementing of a new software system can change the dynamics of the problem that the system tried to solve. This shift in dynamics can lead to the problem drastically changing and making the fairness notion used no longer relevant.

Emergent bias (Definition 4.1.14) occurs due to changes in society itself. This is a type of bias to which a system in the context of fairness is especially sensitive. Unlike historical bias, emergent bias not related to the data gathered but rather to the context itself. This means that due to societal changes the fairness definition or implementation to achieve fairness are not in line with society anymore.

6.2 Methods to monitor and predict bias

The best way to monitor the system's performance is to use the system's data of during its operation. This means that data collection should continue while the system is in use, which requires permission from the organisation itself. Another possibility would be not using all the data but only using samples. However, there is no strong reason to do this, as the compute time will be minimal to evaluate the AI system.

This monitoring will reveal if the system satisfies the new data's desired fairness definition and makes it possible to compare the training set and test set with the user population. It is most important that the fairness definition is still satisfied. However, it is also interesting to see if the use population matches the population reflected in the training and test set. If these results are starting to diverge, then it can be a sign that the model may become unfair in the future due to the changing use population if the fairness definition used uses group fairness. On the other hand, if individual fairness is ensured, then a changing use population should not affect satisfying the definition as it is not sensitive to how the groups are.

The work of D'Amour et al. takes an interesting approach by creating theoretical environments in order to simulate the behaviour if an AI system is implemented [50]. These simulations do not use real-world data but rather serve as a type of thought experiment. As mentioned before several biases can occur after the implementation of an AI system. These simulations aim to give the developer a sense of how the real-world might evolve. Certain situations were already implemented for the paper and offered some food for thought about the evolution of AI systems when taken into the real world. These simulations are not sufficient to accurately predict the real world's evolution as they are too simplified for that.

Chapter 7

Conclusion

In this dissertation the different aspects relevant to create a fair AI system was discussed. Fairness in AI requires insight both into the social concepts at play where the AI system would function and a technical view on how to achieve that fairness. This means that both fronts are always evolving, requiring the applications to evolve with them.

The first element that was discussed is the social framework in which the AI system would function. That social framework determines the concept of fairness. The first indication of what society deems fair is given by the legislation, for example through the rights that an individual has and the discrimination laws. The current legislative framework in Europe is not yet equipped to handle the arrival of AI systems, even though they are already here. Therefore, the European Union have made a proposal for specific AI legislation. The goal of that proposal is to strengthen the use of AI in the European Union while protecting the people in order to prevent harm. Quite a lot of technical aspects required for high-risk AI systems are also included in the proposal.

The second aspect in the social framework is the motivation for businesses to implement these AI systems and why the extra effort needs to be put in in order to make the adoption of the technology a success. More and more research is arising about the economical impact these technologies can have with the World Economic Forum also taking an interest in the implementations of AI systems.

Once the position of the AI system in society has become clear it is necessary to determine what is fair in the particular use case. Unlike humans, a computer cannot sense if something is fair, but rather it requires some mathematical concept to adhere to. That mathematical concept needs to be chosen with

great care as a system that does not have the correct fairness requirements will not be fair. Different fairness definitions were discussed, although the list provided was not exhaustive. New fairness definitions are still being created making it impossible to claim that every type is included.

In most cases it is best to combine multiple of these fairness definitions rather than just selecting one. One definition does often not encompass the fairness required for the system. Combining these definitions needs to be done with thought as certain incompatibilities can arise.

The next step is looking at the data, how it is collected, what is present and how it is processed. The AI systems discussed in this dissertation depend on the information of people, meaning that the data will be collected from people. This is one important factor that makes it almost likely that biases will occur in the data sets. Being aware of biases in the data set can gives clear indications of what characteristics to investigate when developing a fair AI system. The list of these different types of biases is quite large, where some types are more likely to occur than others. This concept of biases in the data is important and data quality is also one of the aspects the EU proposal for AI focuses on.

After all these decisions and analyses the AI system can be made. In the creation of the AI system the fairness definitions chosen can play an important role. Four techniques were discussed in that chapter, including two pre-processing techniques, one in-processing technique and one post-processing technique. An analysis ensued of the effects of the techniques. This analysis is done by looking at the statistical measures of groups that were split based on a sensitive attribute. This analysis also clearly showed that it is difficult to satisfy fairness definitions across all these different groups as the properties of each group differed.

Finally the concept of monitoring the AI system after it is taken into production is discussed. This is necessary because that is the moment a model can truly be evaluated as it functions in the real-world. But it is also necessary as the environment in which the application is used will be ever changing and certain changes could negatively impact the performance in terms of fairness. This concept of persistent monitoring is also voiced in the EU proposal for AI.

Chapter 8

Bibliography

[1] J. Larson and J. Angwin, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.," May 2016.

[2] DataRobot, "Trusted ai," Aug 2021. https://www.datarobot.com/platform/trusted-ai/.

[3] Vlaamse adviesraad voor Innoveren en Ondernemen, *Vlaamse beleidsagenda artificiële intelligentie*, vol. 5. 2018.

[4] Council of European Union, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," 2021. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206.

[5] B. Ruf and M. Detyniecki, "Towards the right kind of fairness in AI," *CoRR*, vol. abs/2102.08453, 2021.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *CoRR*, vol. abs/1908.09635, 2019.

[7] T. L. Quy, A. Roy, V. Iosifidis, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *CoRR*, vol. abs/2110.00530, 2021.

[8] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, "A clarification of the nuances in the fairness metrics landscape," *Scientific Reports*, vol. 12, no. 1, 2022.

[9] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the Conference on Fairness, Accountability, and*

Transparency, FAT* '19, (New York, NY, USA), p. 59–68, Association for Computing Machinery, 2019.

[10] M. Szczepański, "Economic impacts of artificial intelligence - european parliament," Jul 2019.

[11] M. Fierens, E. Van Gool, and J. De Bruyne, "De regulering van artificiële intelligentie (deel 1) - een algemene stand van zaken en een analyse van enkele vraagstukken inzake consumentenbescherming," *Rechtskundig Weekblad*, vol. 2020-2021, p. 962–980, Feb 2021.

[12] European Union, "Charter of fundamental rights of the european union," 2012. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT.

[13] K. L. Wong and A. S. Dobson, "We're just data: Exploring china's social credit system in relation to digital platform ratings cultures in westernised democracies," *Global Media and China*, vol. 4, p. 220–232, Jun 2019.

[14] I. M. Enholm, E. Papagiannidis, P. Mikalef, and J. Krogstie, "Artificial intelligence and business value: A literature review," *Information Systems Frontiers*, 2021.

[15] R. V. Loon, "How corporate c-levels can be the guardians of ethical ai," Jun 2020.

[16] World Economic Forum Global Future Council on Human Rights 2016-18, "How to prevent discriminatory outcomes in machine learning," tech. rep., World Economic Forum, March 2018.

[17] M. Schäfer, D. B. Haun, and M. Tomasello, "Fair is not fair everywhere," *Psychological Science*, vol. 26, no. 8, p. 1252–1260, 2015.

[18] B. Paaßen, "European ai alliance - a review of the machine learning literature on fairness," Aug 2018.

[19] S. Verma and J. Rubin, "Fairness definitions explained," *Proceedings of the International Workshop on Software Fairness*, 2018.

[20] H. Hofmann, "Statlog (German Credit Data)." UCI Machine Learning Repository, 1994.

[21] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, (New York, NY, USA), p. 560–568, Association for Computing Machinery, 2008.

[22] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, "Fairness through awareness," *CoRR*, vol. abs/1104.3913, 2011.

[23] A. Chouldechova and A. Roth, "A snapshot of the frontiers of fairness in machine learning," *Communications of the ACM*, vol. 63, no. 5, p. 82–89, 2020.

[24] S. Baert, M. Lamberts, and P.-P. Verhaeghe, "Het terugdringen van arbeidsmarktdiscriminatie in de vlaamse sectoren: academische visie en instrumenten," p. 9–15, Oct 2020.

[25] T. Miconi, "The impossibility of "fairness": a generalized impossibility result for decisions," *arXiv: Applications*, 2017.

[26] S. Beijne and S. Sibie, "Studeren met een functiebeperking - beleidsdocument 2020-2022," Mar 2020.

[27] Fonteyne, Lot, *Constructing SIMON : a tool for evaluating personal interests and capacities to choose a post-secondary major that maximally suits the potential.* PhD thesis, Ghent University, 2017.

[28] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Frontiers in Big Data*, vol. 2, 2019.

[29] Y. Liu, C. Kliman-Silver, and A. Mislove, "The tweets they are a-changin': Evolution of twitter users and behavior," 2014.

[30] C. B. Barshied, "The progress of medical labor: Gender shifts, generational differences, and the coverage continuum in obstetrics and gynecology," 2016.

[31] L. J. Sanna and N. Schwarz, "Integrating temporal biases: The interplay of focal thoughts and accessibility experiences," *Psychological Science*, vol. 15, no. 7, pp. 474–481, 2004. PMID: 15200632.

[32] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, (New York, NY, USA), Association for Computing Machinery, 2021.

[33] H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht, ""blissfully happy" or "ready tofight": Varying interpretations of emoji," 2016.

[34] S. C. Bates and J. M. Cox, "The impact of computer versus paper–pencil survey, and individual versus group administration, on self-reports of sensitive behaviors," *Computers in Human Behavior*, vol. 24, no. 3, pp. 903–916, 2008. Instructional Support for Enhancing Students' Information Problem Solving Ability.

[35] C. Ziems, J. Chen, C. Harris, J. Anderson, and D. Yang, "VALUE: Understanding dialect disparity in NLU," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Dublin, Ireland), pp. 3701–3720, Association for Computational Linguistics, May 2022.

[36] P. Bickler, J. Feiner, and J. Severinghaus, "Effects of Skin Pigmentation on Pulse Oximeter Accuracy at Low Saturation," *Anesthesiology*, vol. 102, pp. 715–719, 04 2005.

[37] Y. Dong, O. Lizardo, and N. V. Chawla, "Do the young live in a "smaller world" than the old? age-specific degrees of separation in a large-scale mobile communication network," *CoRR*, vol. abs/1606.07556, 2016.

[38] J. C. Doidge and K. L. Harron, "Reflections on modern methods: linkage error bias," *International Journal of Epidemiology*, vol. 48, pp. 2050–2060, 10 2019.

[39] M. Merler, N. K. Ratha, R. S. Feris, and J. R. Smith, "Diversity in faces," *CoRR*, vol. abs/1901.10436, 2019.

[40] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020.

[41] B. W. Carlson, "Problem of causality," Jan 2019.

[42] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Trans. Inf. Syst.*, vol. 14, p. 330–347, jul 1996.

[43] R. Baeza-Yates, "Bias on the web," *Commun. ACM*, vol. 61, p. 54–61, may 2018.

[44] D. Agarwal, B.-C. Chen, and P. Elango, "Explore/exploit schemes for web content optimization," in *2009 Ninth IEEE International Conference on Data Mining*, pp. 1–10, 2009.

[45] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific Data*, vol. 4, no. 1, 2017.

[46] G. A. Abel, M. E. Barclay, and R. A. Payne, "Adjusted indices of multiple deprivation to enable comparisons within and between constituent countries of the uk including an illustration using mortality rates," *BMJ Open*, vol. 6, Nov 2016.

[47] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, and et al., "Bias in data-driven artificial intelligence systems—an introductory survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2020.

[48] L. R. Olsen, "Multiple-k: Picking the number of folds for cross-validation," Nov howpublished=https://cran.r-project.org/web/packages/cvms/vignettes/picking_the_number_of_folds_for_cross-validation.html 2021.

[49] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Tech. Rep. MSR-TR-2020-32, Microsoft, May 2020.

[50] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern, "Fairness is not static: Deeper understanding of long term fairness via simulation studies," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAccT '20, (New York, NY, USA), p. 525–534, Association for Computing Machinery, 2020.

Appendix A

Statistical metrics

A.1 Basic metrics

- True Positive (TP):

  These are the cases where the true and predicted classes are both the positive class

- True Negative (TN):

  These are the cases where the true and predicted classes are both the negative class

- False Positive (FP):

  These are the cases where the true class is the negative class, and the predicted class is the positive class.

- False Negative (FN):

  These are the cases where the true class is the positive class, and the predicted class is the negative class.

- Actual positives (P):

  The cases where the actual class is the positive class. This equals to the sum of the true positives and false negatives.

  $$P = TP + FN$$

- Actual Negatives (N):

  The cases where the actual class is the negative class. This equals to the sum of the true negatives and the false positives.

  $$N = TN + FP$$

A.2 Derived metrics

- Accuracy:

  This is the fraction of classes which were correctly classified.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

- Misclassification rate (MR):

  The complement of the accuracy. This is the fraction of classes which were incorrectly classified.

$$\frac{FP + FN}{TP + FP + TN + FN}$$

- Base Rate (BR):

  The fraction of actual positive classes in the entire set.

$$\frac{P}{P + N}$$

- Positive rate (PR):

  The fraction of all samples predicted to be in the positive class.

$$\frac{TP + FP}{TP + FP + TN + FN}$$

- Negative rate (NR):

  The fracion of all samples predicted to be in the negative class.

$$\frac{TN + FN}{TP + FP + TN + FN}$$

- Positive predictive value (PPV):

  The fraction of positive cases correctly predicted to be in the positive class out of all cases predicted as positive. This is also referred to as precision.

$$\frac{TP}{TP + FP}$$

- False discovery rate (FDR):

  The fraction of negative cases incorrectly predicted to be in the positive class out of all cases predicted as positive.

$$\frac{FP}{TP + FP}$$

- Negative predictive value (NPV):

  The fraction of negative cases correctly predicted to be in the negative class out of all cases predicted as negative.

$$\frac{TN}{TN + FN}$$

- False omission rate (FOR):

  The fraction of positive cases incorrectly predicted to be in the negative class out of all cases predicted as negative.

  $$\frac{FN}{TN + FN}$$

- True positive rate (TPR):

  The fraction of positive cases which are correctly predicted as positive out of all positive cases. This is also referred to as sensitivity or recall.

  $$\frac{TP}{TP + FN}$$

- False positive rate (FPR):

  The fraction of negative cases which are incorrectly predicted as positive out of all actual negative cases.

  $$\frac{FP}{TN + FP}$$

- True negative rate (TNR):

  The fraction of negative cases which are correctly predicted as negative out of all negative cases.

  $$\frac{TN}{TN + FP}$$

- False negative rate (FNR):

  The fraction of positive cases which are incorrectly predicted as negative out of all actual negative cases.

  $$\frac{FN}{TP + FN}$$

A.3 Confusion matrix

Table **??**omething shows how the confusion matrix will be used in this dissertation.

Accuracy: $\frac{TP+TN}{TP+FP+TN+FN} * 100$		Predicted		
		Positive	Negative	BR = $\frac{P}{P+N}$
True	Positive	TP	FN	TPR = $\frac{TP}{TP+FN}$
	Negative	FP	TN	TNR = $\frac{TN}{TN+FP}$
		FDR = $\frac{FP}{TP+FP}$	FOR = $\frac{FN}{TN+FN}$	
		PR = $\frac{TP+FP}{TP+FP+TN+FN}$	NR = $\frac{TN+FN}{TP+FP+TN+FN}$	

Table A.1: Confusion matrix and statistical measures of the test set.

Appendix B

Sensitive attribute distributions

B.1 In the OULAD data set

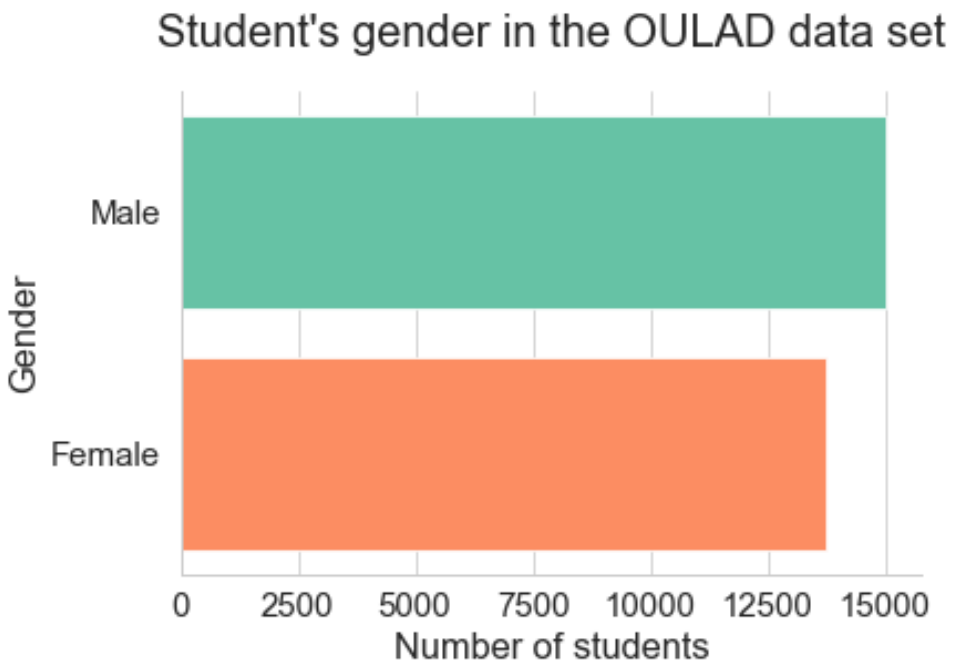


Figure B.1: Gender distribution in the OULAD data set

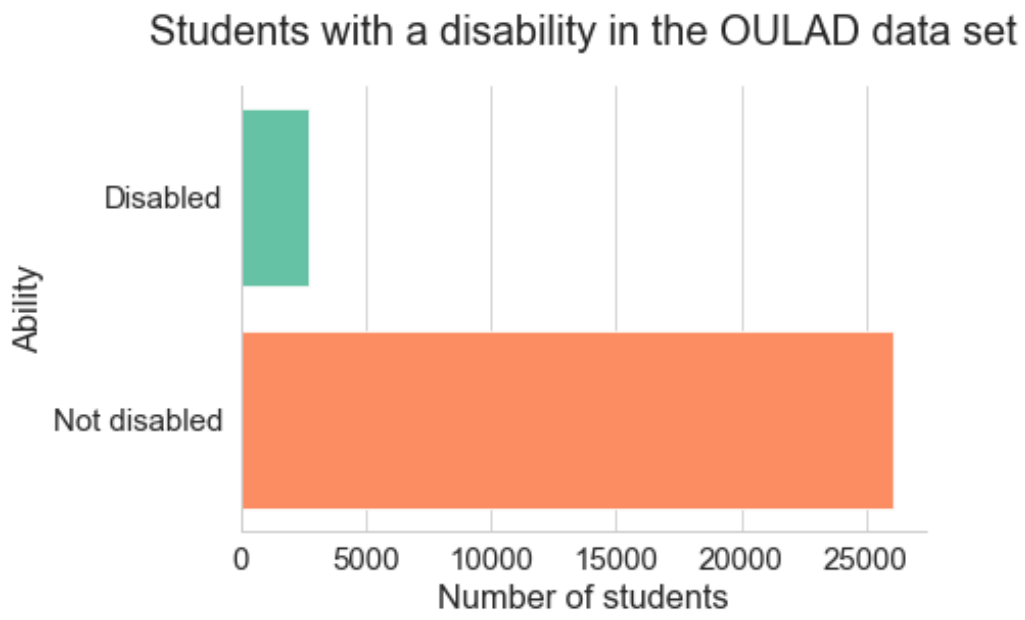


Figure B.2: Proportion of people with a disability in the OULAD data set

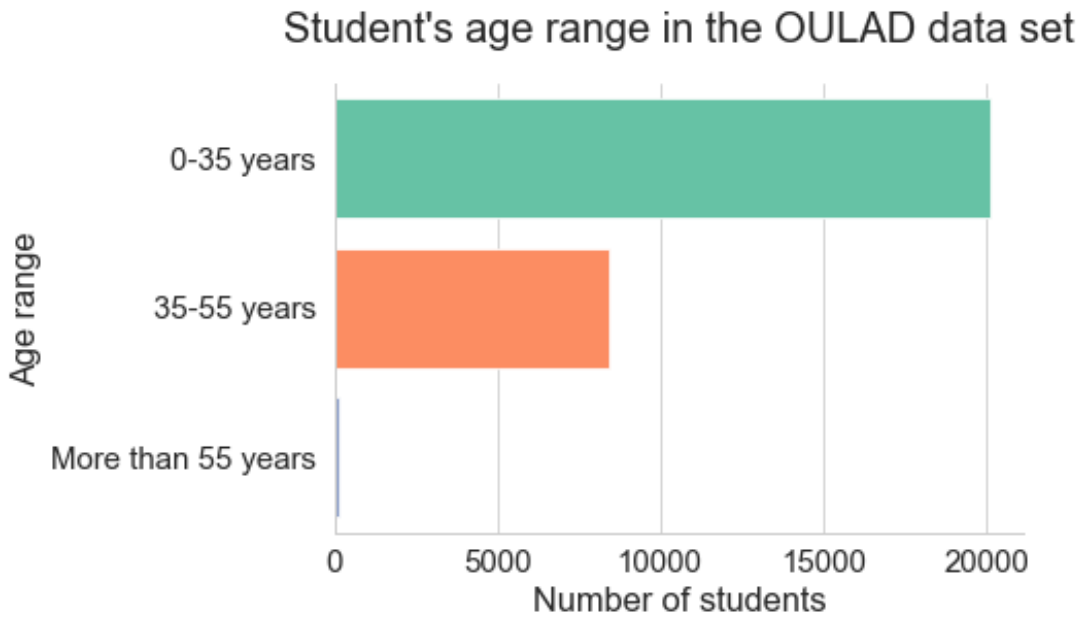


Figure B.3: Distribution of people's age in the OULAD data set

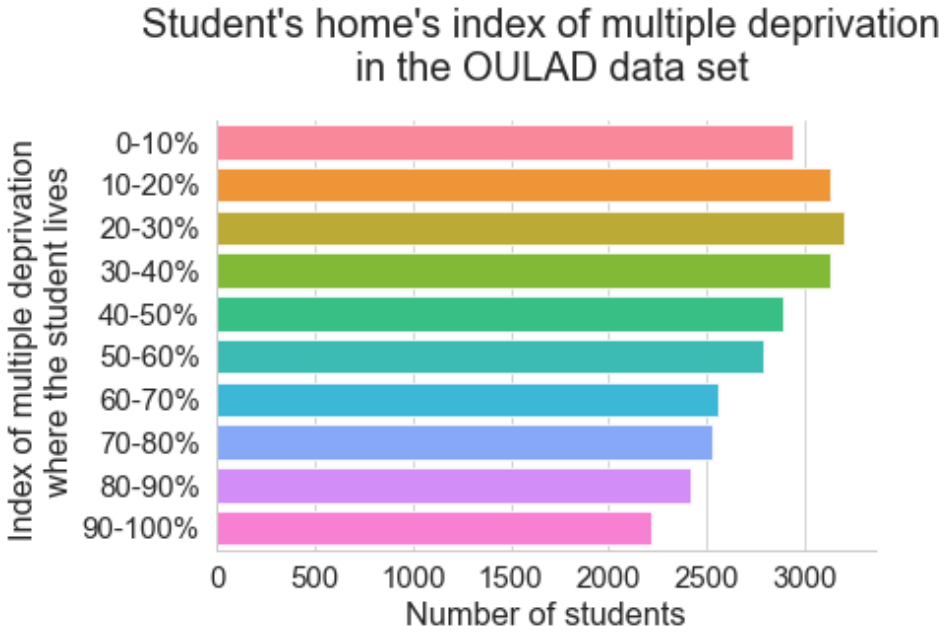


Figure B.4: Distribution of people's homes' index of multiple deprivation in the OULAD data set

B.2 In the BBB course

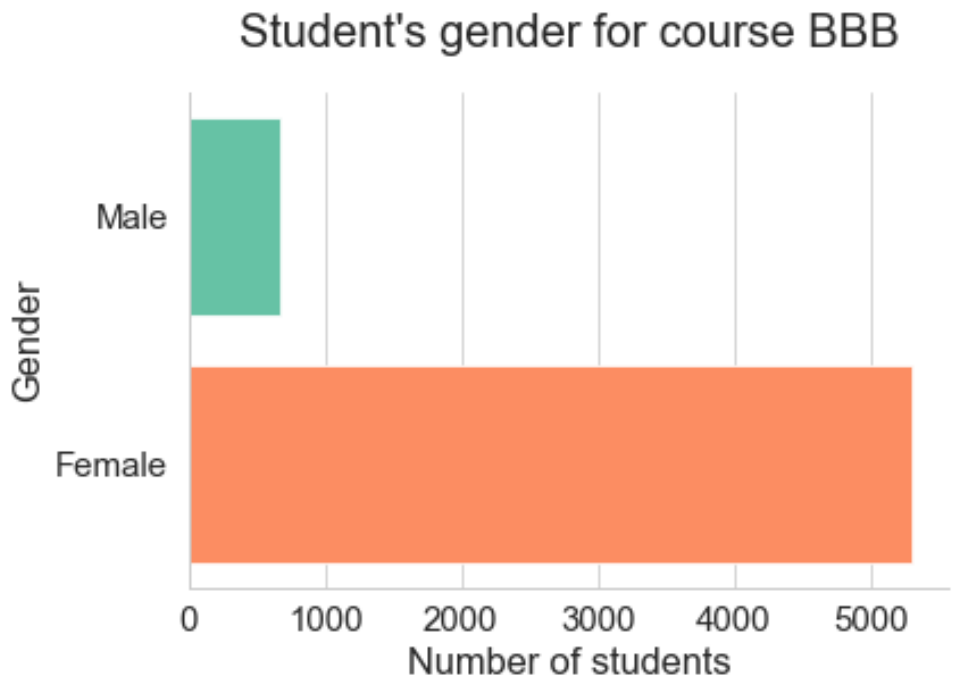


Figure B.5: Gender distribution for the BBB course

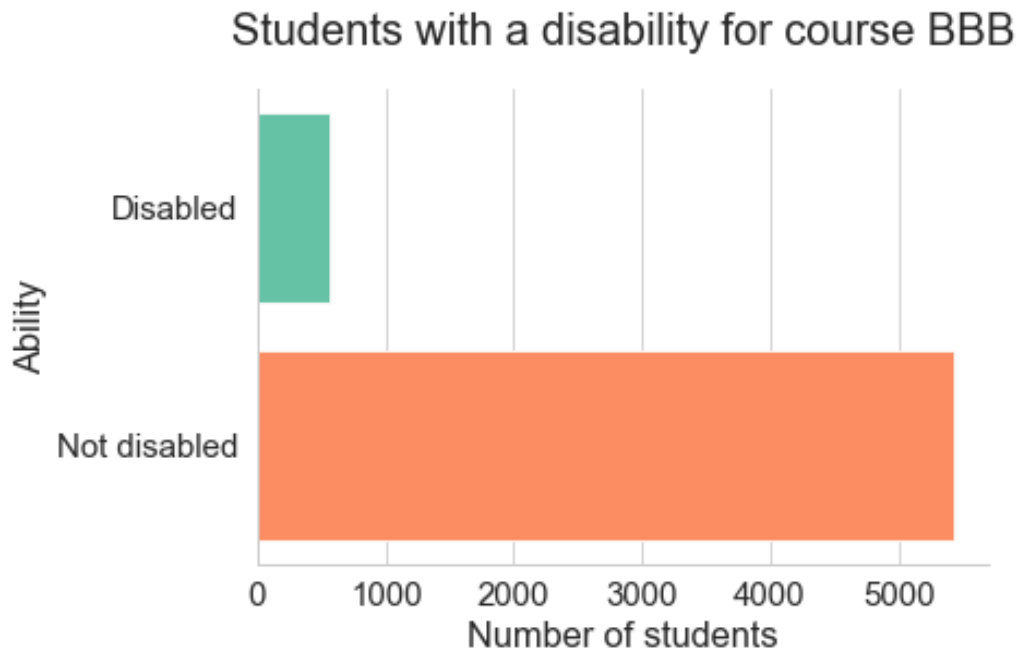


Figure B.6: Proportion of people with a disability for the BBB course

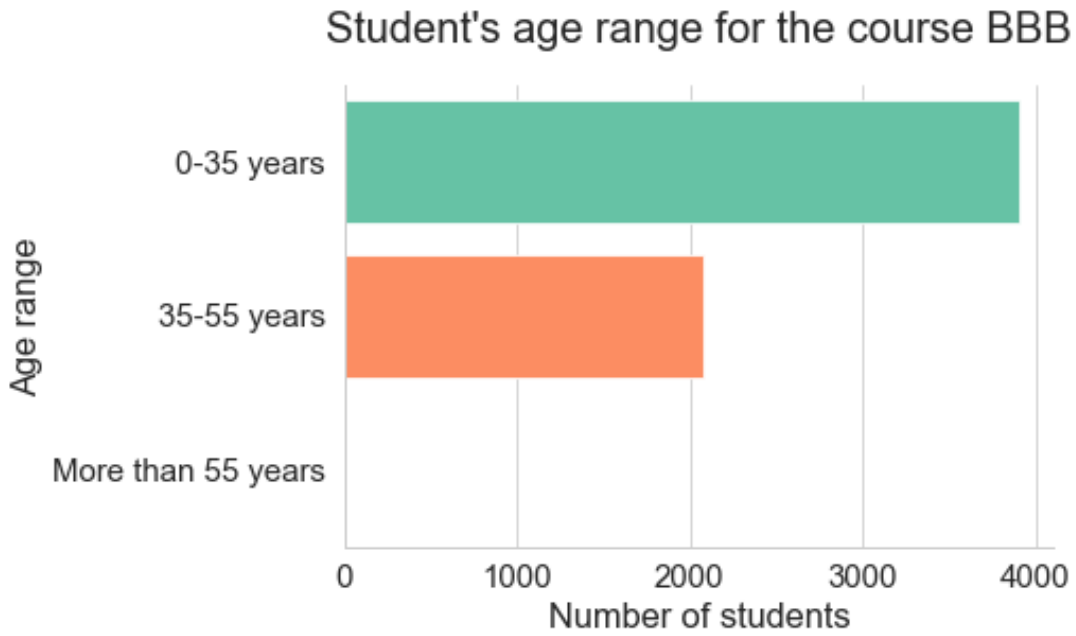


Figure B.7: Distribution of people's age for the BBB course

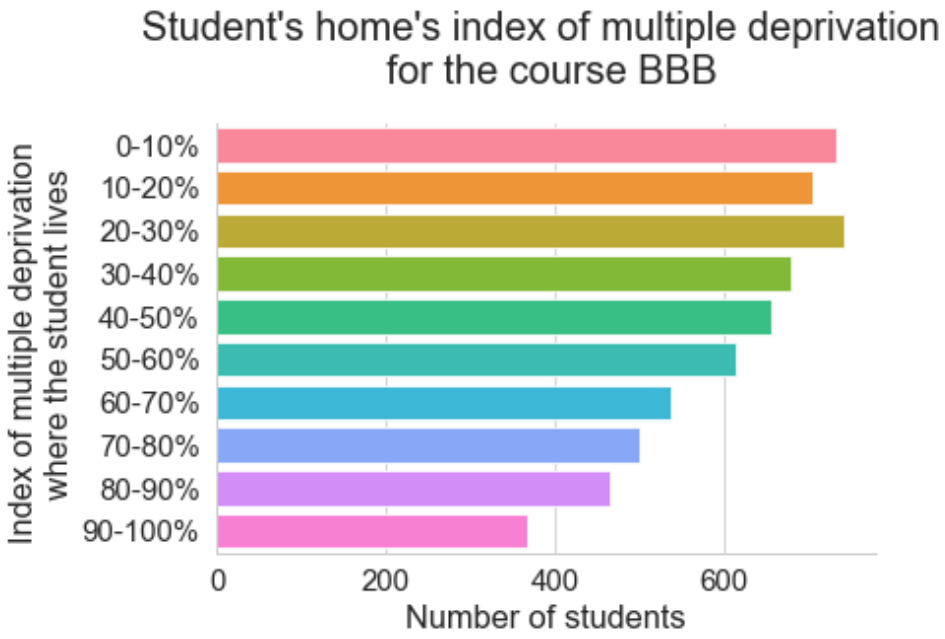


Figure B.8: Distribution of people's homes' index of multiple deprivation for the BBB course

B.2.1 People who dropped out from course BBB

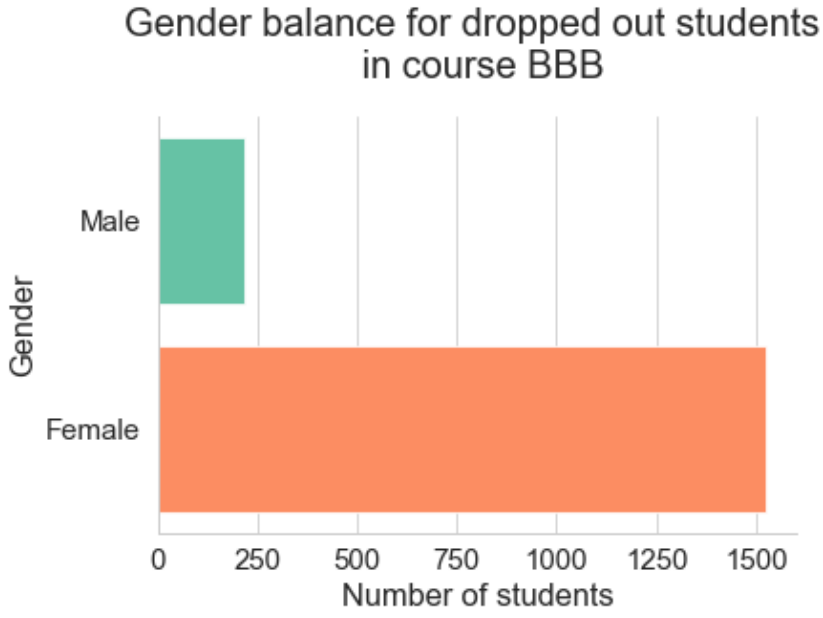


Figure B.9: Gender distribution of people who dropped out for the BBB course

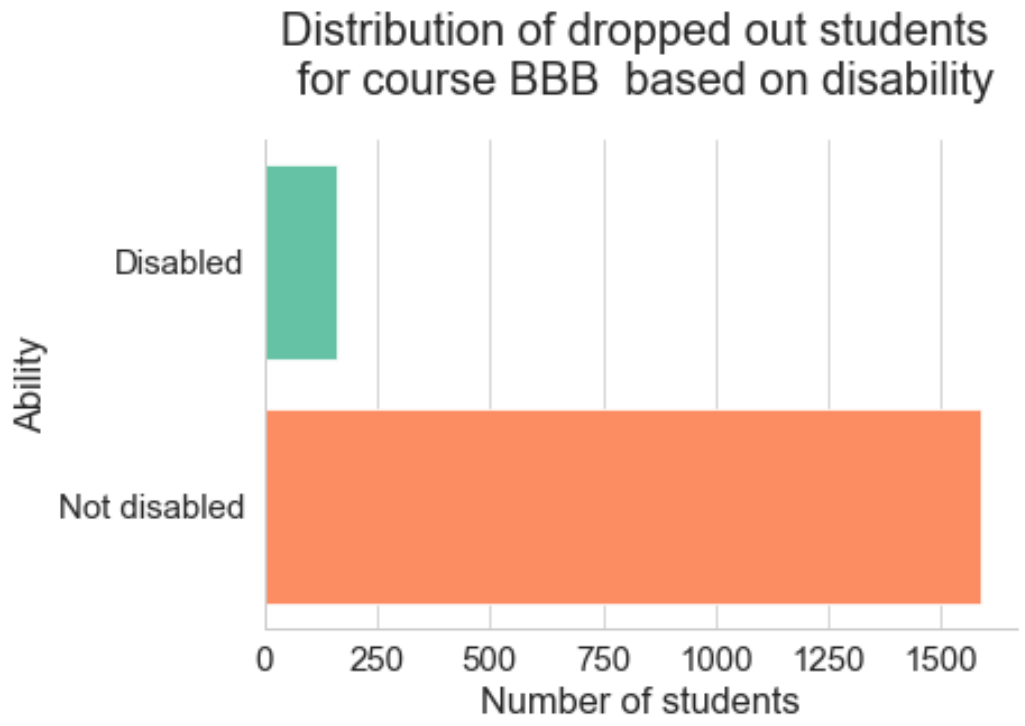


Figure B.10: Proportion of people who dropped out with a disability for the BBB course

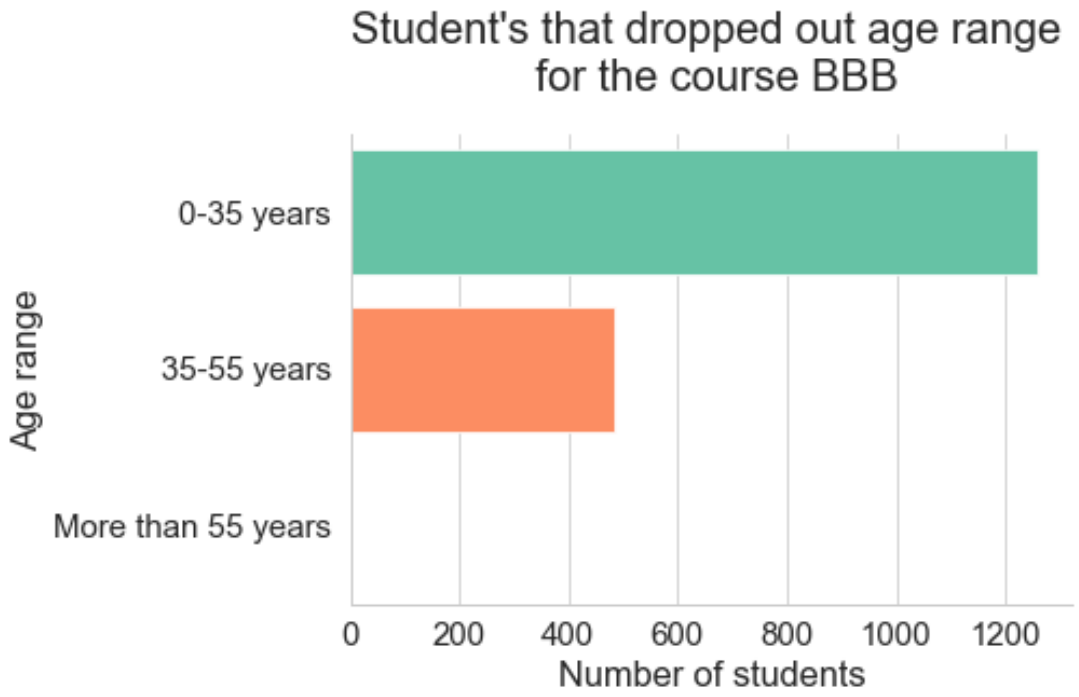


Figure B.11: Distribution of people's age of who dropped out for the BBB course

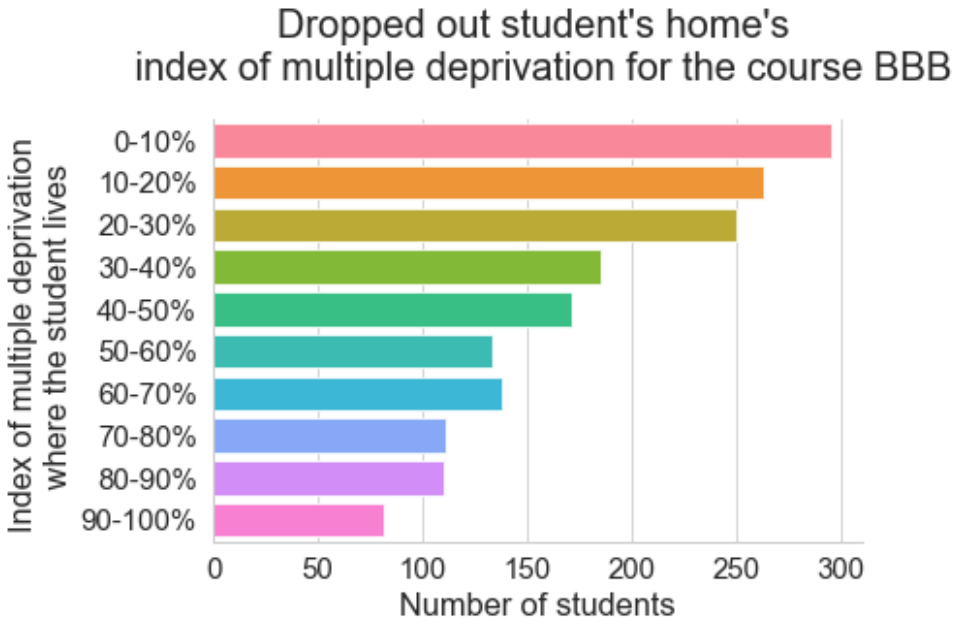


Figure B.12: Distribution of people's homes' index of multiple deprivation of who dropped out for the BBB course