DELFT UNIVERSITY OF TECHNOLOGY

MASTERS THESIS

FairData: A system to explore datasets as a tool to let data-centric researchers become aware of bias

Author:
Tomas HEINSOHN HUALA

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Web Information Systems Group
Software Technology

October 11, 2021

FairData: A system to explore datasets as a tool to let data-centric researchers become aware of bias

by

Tomas Heinsohn Huala

to obtain the degree of Master of Science
in Computer Science
at the Delft University of Technology,
to be defended publicly on Friday October 15, 2021 at 01:30 PM.

Student number:	4326318	
Project duration:	November, 2019 - October, 2021	
Thesis committee:	Dr ir. Jacky Bourgeois,	TU Delft, supervisor
	Prof. dr. ir. Alessandro Bozzon,	TU Delft, supervisor
	Dr. ir. Ujwal Gadiraju,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`

TU Delft

Abstract

Data-centric research is all about trying to obtain useful insights on products that its researchers trying to create/design by collecting usage data on their prototypes. This type of research, usually performed by data-scientist, make use of machine learning techniques in order to improve any system or product that makes use of automated decisions. Now that other researchers, in this case researchers from design studies, would like to perform data-centric research. Therefore, they must be made aware of the issues that could be introduced by using machine learning to automate decisions in electronic devices.

To address the issues around bias, we thought that it would be good to make them aware through some type of system where they should be made aware of bias. So we proposed that there was a need for a system where the user could explore bias present in the dataset or introduced by machine learning practices. To be able to find biases we had to decide which bias toolkit which led to use of the AIF360 toolkit that we eventually used to create FairData. FairData lets the user explore bias in datasets that include machine learning predictions by going through and iterative process consisting of a exploration stage and a result stage. The exploration part contains the parts for analysing the datasets by going through the attributes and making small modifications to the dataset. At the results page of this system it will show a change in bias represented by multiple metrics to support the users in becoming aware of bias.

The experimentation part of this thesis showed that our system is able to inform two different groups of participants (compute science & industrial design) about the bias present in the dataset. The experiment made it clear that even though participants have different behaviour when analysing/exploring the dataset, they all tend to become aware of the bias by looking at the different metrics provided by FairData. It also seems that it is possible to support users in trying to reduce this bias as they can see the effects of the modifications done to the dataset.

Acknowledgements

I would like to thank my supervisors Jacky Bourgeois and Alessandro Bozzon for helping me throughout the stages of this research and providing the opportunity to come to the result presented in this thesis. I also would like to thank Agathe Balayn for her help in the machine learning related parts and for her help during the experiment. Despite the outbreak of COVID-19 which introduced some challenges during this thesis, I also would like to thank you all for your patience and your understanding at these trying moments. Social distancing and thus being at home all the time brought its own complexities which did not make things easier. Finally, I also would like to thank my family and friends for their support throughout this thesis. Especially my parents and my brother for always supporting and encouraging me during my thesis project.

Tomas Heinsohn Huala
Schiedam, October 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data-centric research is a common type of research that often combines data with Machine Learning techniques in order to create/improve systems or products. The data used for this type of research often consists of datasets that contain personal data which is either obtained by letting other people use their prototypes/systems or retrieved from databases owned by companies/governments. This idea of data-centric research in the creation of new products is now also being considered by researchers that do not have the required skills or tools to usefully process their data.

Nowadays, practices of data-centric research are being used in different domains, but could be able to introduce situations where the data that is obtained and used could lead to cases where certain people or groups of people can be treated unfairly [17]. Seeing that there is effort to make the data scientist aware of this issue in Machine Learning practices, it might be useful to also introduce the issue of bias to the another group of data-centric researchers.

Due to the introduction of IoT (Internet of things) devices and similar kind of products, designers and other researchers also became interested in creating/designing such products. To successfully do this they needed to improve on existing practices of user experience design by obtaining more usage data, this new form of user experience design can thus be seen as data-centric research. These new type of practices is something that was also proposed by [10], as they propose that analysing sensor data would help product designers. The only problem is that (most) designers do not have the (programming) expertise to analyse/process the collected data which could make it very hard for designers/researchers without this expertise to do data-centric research.

They would need tools and possibly also experts in data analysis to identify usable usage patterns of interest. Researchers in [8] claim that for creating IoT products or similar, designers need to have a close collaboration with ethnographers and computer scientists. As that might not always be possible it would be better to look for a tool where these other type of practitioners of data-centric research would be supported to use those machine learning techniques themselves. Hence, it might be dangerous if we don't warn those practitioners about the dangers of using machine learning practices without paying attention to bias.

Therefore, we must introduce them into the concept of bias being introduced by either misrepresentations in their datasets or introduced by the machine learning algorithm they apply. There are already multiple cases where bias has a big impact and could be very harmful to certain groups of people. The COMPAS case is such an example where a recidivism risk prediction tool had a biased outcome towards African-American people [3]. This was extremely harmful as it influenced decisions in setting people free during multiple stages of the U.S. criminal justice system.

Based on the need of these practitioners of data-centric research and on the issues related to bias in machine learning, there is not such a system which lets those practitioners become aware of bias in data-centric research. This bias can be either in the dataset or being introduced by the machine learning techniques applied. So in this thesis it is important to find out if it is possible to build our own application which focuses on letting both types of practitioners become aware of bias in data-centric research and if such a system would work.

1.1 Problem Statement

The goal of this thesis is to create some type of exploratory system that supports practitioners of data-centric research to become aware of the potential bias introduced by their dataset. To carry that out, we first have to look why practitioners of data-centric research should become aware of bias and how we could find unfairness in data-centric practices. Therefore, we also have to investigate which type of tools we could use for finding bias.

Secondly, to build some kind of platform which is able to guide the users through the process of finding bias, without having them to write a single line of code can be pretty complex. We have to create a system which makes people aware of this bias in a relatively easy and intuitive manner, despite the fact that some of them might not have a lot of experience with bias in such scenarios.

Eventually, we also need to analyze if the application we create does support those practitioners of data-centric research to become aware of bias. For this experimentation we need to come up with a method to test if the system achieves its goal which can be used to draw conclusions in this research.

1.2 Research questions

With the help of the section 1.1 and the challenges that we discussed, it was possible to come up with a main research question and several sub research questions. Also motivated by the fact that there is not yet such a system that could support designers/researchers in exploring their dataset and be informed by potential bias.

Main Research Question: How can practitioners of data-centric research (data-scientist or not) be supported to become aware of bias with the help of an application that 'explores' datasets?

In order to answer this question, I split it up into three sub research questions:

RSQ 1: Why should the practitioners of data-centric research be informed about potential bias and how can bias be found in datasets obtained by data-centric research?

To answer this question, we first will look into the reasons why it is important to find bias in data-centric research. It will introduce the practitioners of data-centric research and look into the dangers of bias in different contexts. After that we will explore the different tools with which we could detect bias in a dataset and see how we could incorporate this into a system where the user should be informed about possible bias.

RSQ 2: What is required to create an application which helps to explore datasets that is able to find potential cases of bias located in datasets?

This question will look at the design of the exploratory system as a whole and will determine what is needed to create this application. With the help of the previous research question, I will be able to establish a process to find biases by asking for the right data. By knowing what is needed to calculate the potential bias, we can design a system which lets the user provide the right information in a convenient way.

RSQ 3: How does the application inform the user about potential bias located in the dataset and does it support the practitioners of data-centric research to reduce this bias?

The reason we created this system is to see if we can make people aware of bias in datasets or bias introduced by machine learning, so they will be aware that using personal data in data-centric research could have consequences. To answer this question we want to get insights on how these practitioners become aware of bias and if they are able to reduce it. For this we will run an experiment in order to obtain that information.

1.3 Contributions

Answering the research question will provide three main contributions:

Contribution 1: A process that is used in the application to explore datasets obtained by data-centric research and what is needed to calculate potential bias metrics. These are multiple steps where the user should provide information to be able to go to the next stage.

Contribution 2: A prototype application through which practitioners of data-centric research can analyze their data and become aware of bias present in the dataset. This will be a web application where the user could upload a dataset and could iteratively obtain more information about the bias present in the dataset.

Contribution 3: Insights on the effectiveness of the exploratory system that is made. We will use a slightly adapted version of the system covering a certain scenario to see if the user will be informed well enough by the system and if it does support both groups of practitioners of data-centric research. Those practitioners will be asked to shape the data with the help of our tool so we could observe what those practitioners would when the UI and fairness metrics will be available to them.

1.4 Outline

The structure of this thesis is as follows. After the introduction we will focus on the needs of practitioners of data-centric research and bias in machine learning in chapter 2. Also we will look into the different tools already present that focus on finding bias in the use of machine learning. Subsequently, we will look into the design of our system to be able to 'explore' datasets in chapter 3. In chapter 5 we will discuss the actual implementation of FairData. Furthermore, in chapter 5 we will discuss the experiment done to see if the system does where it is designed for.

Finally, we will provide a discussion of the results in chapter 6 and summarize our findings of this thesis in chapter 7.

Chapter 2

Related work/Background

In this chapter, we will discuss the related work and give an extensive background on the topic of finding bias and what is relevant to be used in the exploratory system. First we will look into literature of data-centric research to get a better understanding of our target group and why non-datascientists might want to do this type of research. After that, we will shortly look at the definition of bias and a couple of examples on consequences of bias in different applications. Lastly we will look into the different tools available for detecting bias and we will decide which one we would like to use for our system.

2.1 Practitioners of data-centric research

This section focuses on the practitioners of data-centric research. Data-centric research is broad term for describing research which makes use of data in order to aid the development of a new product in which the data helps to understand the target (audience). In this section, we will look at the practitioners of data-centric research with or without being a data-scientist to see why those practitioners that aren't data-scientists might need a system to get useful insights into their data.

In both Industrial Design and Computer Science, user experience in design is a form of data-centric research where they obtain data about the use of a product. Involving multiple users in the design of new products could help the designer as they obtain more crucial information on the use of that product with the help of different sensors. The only problem is that practitioners without that data-scientist background would also like to be able to obtain the right information when they do this type of research. For a long time, there already was a need for methodological frameworks to help these type of designers to obtain better insights on findings in design research [13].

For instance, solutions in the form of frameworks were provided by [20], where they propose a 'palette' for design-inclusive UX research. This 'palette' showed the different roles of design activities in UX research and had the purpose of giving researchers ideas on what they would like to achieve with their research. On top of that, it could be also used to validate the inclusion of design in UX research. The roles of design activities as presented by [20], are very useful for determining why designers would do UX research and why they would need to use the system created in this project. Especially the second role of providing rich prototypes to study aspects of experience and interaction showed what design research involving the user headed towards to.

Other research even proposes the collaboration of ethnographer, ML experts and designers in order to create successful IoT products [8]. The expertise of these three groups should help to maximize the gain from patterns located in the data which

help the designer to be more successful in designing new products. However, designers may not want to be dependent on being required to collaborate with these other groups in order to design more useful products.

Most of the data-driven research incorporating the user is performed by data scientists. In [6] for instance, researchers from Philips use a baby bottle accompanied by sensors to obtain information on the use of this baby bottle and even provide a framework to both use quantitative sensor and subjective qualitative sensor data. This research was part of a larger project [19] involving different practices, of creating new IoT products, is what they call user experience research combined with design and they showed that combining these aspects has its benefits.

So seeing that using data-driven research creating those IoT products can be very useful, it might be useful if there were systems that help those non-data-scientists to analyze the data they obtain. In [7] they came up with an architecture for doing analysis on personal IoT data which focuses on being GDPR complaint (Global Data Protection Regulation). Such a system could help the practitioners of data-centric research in such a way that it offers a platform where they could not only protect the personal data obtained, but also can make use of different algorithms to do the data analyses they need.

Alternatively, there is also data-driven research performed by data-scientists that uses smartphone data in order to get insight into human mobility. For instance, it could be used to model the choice of catering locations on a campus with the help of Wi-Fi traces and see which locations are preferred by certain groups of people [9]. Similar research looks at human mobility trajectories in (a part of) New York also with the help of Wi-Fi probe request data [18]. This data helped them to get a better understanding of mobility patterns in cities. In both of these cases the research have to process a lot of data in order to get something useful and meaningful out of it.

The practitioners of data-centric research could definitely need some type of system to help those that aren't data-scientists. This can be some kind of GDPR-compliant system as spoken about earlier, but the most important thing to start with is that they will be able to 'explore' their data in order to get more insights into it.

2.2 Bias research and examples

As seen in previous section, a lot of data-centric research is eventually focused on extracting useful information out of the unstructured data they collected, so it could be used for specific purposes. However, the data collected by research does not always fully represent real life situations and thus could it misrepresent specific groups of people. This 'inclination or prejudice for or against one person or group, especially in a way considered to be unfair,' (Oxford Dictionary) is known to be called bias.

There could be many reasons why bias is located in a dataset or machine learning model. A very important fact is that many applications of machine learning use data generated by humans or is collected by a device that is created by humans. Therefore, bias could be introduced by just simple mistakes which are related to the person who is responsible for that data. On the other hand, it could also be a cause of existing inequalities or institutional racism and sexism which is just being highlighted or increased by an algorithm or application [14].

Bias in Machine learning is a well-known and important topic in the field of machine learning. Different types of bias have been identified in subsequent research and as AI became more popular it also introduced more issues related to it. Bias is

either found in the datasets used or coming from the algorithm applied to train a machine learning model.

For instance, the case of the recidivism risk prediction tool called COMPAS which seemed to be biased against African-American people [3]. This tool helped U.S. judges and parole officers in making decisions about a defendant its sentence. The problem with COMPAS was that its machine learning algorithm took prior arrests and friend/family arrests into account to measure the level of 'riskiness' or 'crime'. This was seen as pretty unfair as minority communities are controlled and policed more frequently, so it is likely that they have higher arrest rates.

Furthermore, there are also some cases where the usage of open source datasets could lead to fairness issues when using those for systems that include machine learning. Like the image datasets ImageNet and OpenImages which seemed to have some form of bias that when it comes to the representation of all different types of people [17]. The use of this kind of datasets could be harmful to different types of people (in terms of gender and race) as those datasets could be used to train image recognition software.

For this reason, in the EU they already have some regulations that try to prevent discriminatory decisions made by AI systems [15]. Although this might not cover all the different situations, it is important that these type of situations are avoided as much as possible. These regulations are also coming from GDPR as bias is highly related to misrepresentations of personal data.

So in order to avoid this, practitioners of data centric research, should be made aware of the impact their data might have. AI-systems which have a biased outcome/decisions could have a devastating effect on somebody's life. Therefore, both the bias in the original dataset as the bias introduced by modifications should be addressed by the exploratory system. By doing this we will be able to introduce them into the concept of bias and let them experience how their actions affect bias located in a dataset.

2.3 Tools for finding bias

Due to these harmful effects of bias in different applications, we wonder what kind of system we could create. That's why, this section will discuss the different tools available for exploring data and/or addressing bias in datasets. This is both for the purpose of finding how this system could stand out and for the purpose of finding out which tools we could reuse in our system.

A comparison was made based on the state-of-the-art tools that focus on finding bias and possibly also mitigating it. In order to make a decision on which tools we eventually want to use, we defined a couple requirements to decide which one fits best a larger system to let practitioners become aware of bias. The UI's of the different tools might also help to provide new useful ideas for a new system.

A first important requirement is that it should be possible to integrate this tool into some kind of web application. Working with datasets, it is best to create a system which isn't limited by the user its PC. Secondly, we also want this tool to be able to work with different datasets that include predictions so we don't have to implement different machine learning techniques into the system which would make it a lot more complex. Lastly, it would be also good if the tool has clear and not too advanced metrics. The bias metrics that will be used in the system should be understandable to all the practitioners of data-centric research.

	Usability	Dataset support	Useful metrics
TensorFlow, Fairness Indicator			x
Aequitas	x	x	x
FairLearn		x	x
AIF360	x	x	x
FairVis		x	x
DiscriLens			x
What-If			
FairSight			x

TABLE 2.1: Comparison table of the tools for finding bias

The first toolkit we looked at was the TensorFlow toolkit which is an open source library focused on developing and training ML models. Besides that, it has an additional library for reviewing fairness issues called fairness Indicators [1]. The only issue with this library is that its representation is a separate view which makes it harder to integrate into a web application. Similar to this there is also the toolkit called Fairlearn [5] that also provides a component to extensively visualize fairness of a trained ML model. This component also has the problem that it is some kind of Python widget which is not very convenient as we want a tool that could be easily integrated into a complete system.

Additionally, we also looked at the bias audit toolkit called Aequitas [16]. This toolkit focuses on both data-scientists and policymakers as it provides support for creating extensive bias reports which are also understandable for non-datascientists. It requires datasets that already include predictions and has it own plot which could be easily integrated into the type of system we want.

Another toolkit called the IBM AI Fairness 360 open source toolkit (AIF360) offers similar opportunities for finding bias [4]. This toolkit was made for the increasing importance of detecting fairness in machine learning models, but can also be used to calculate bias without an actual ML model trained. Seeing the demo of this toolkit, it shows that it could be very useful seeing the simple visualizations of the fairness metrics in which they show the difference of mitigating the bias. This idea could be very helpful in a system where the user could see the impact of their actions.

Furthermore, there are tools which focus more on the visualization of the fairness of machine learning models. Tools like FairVis [2] and DiscriLens [21] visualize the bias present in machine learning models and focuses on the exploration of the subgroups present in the dataset. Both tools provide a wide range of metrics and visualizations in order to give them insight into specific subgroups. Unfortunately these tools are already complete systems, although their visualizations could help in the design of our system.

There is also the What-If Tool [22] which is more about exploring different models by tweaking the parameters and let the user measure the performance of those models by multiple ML fairness metrics. This tool is a visual part of the TensorFlow framework and therefore also not very useful to integrate into a larger system. Similarly, there is also the tool called FairSight [1] which also includes the mitigation of the potential bias found. This tool offers full support to all the stages from data exploration to the outcomes of the machine learning models, but might be a bit complex to be used by all practitioners. So despite the relevance of these visualization

[1] https://www.tensorflow.org/tfx/guide/fairness_indicators

tools in terms of exploring the dataset, we need something which is more accessible for the practitioners of data-centric research in general.

Based on the requirements we had and the different tools existing that cover bias, it was clear that it might be best to use the AIF360 toolkit in a system where we let the user become aware of bias. This AIF360 toolkit was chosen as it will provide a framework with which it will be possible to calculate bias metric on different datasets. Besides that, it is more advanced and complete than the Aequitas tool as this tool is also able to mitigate bias. With the help of this toolkit we will be able to create an exploratory system that offers the practitioners to explore their dataset and inform them about bias located in a dataset.

2.4 Summary

Data-centric research is an area of research where both data-scientists and designers want to learn more about the use of a prototype or product by obtaining usage data. Both groups find out that it is very useful to do this type of research as you really get to understand how people interact with certain concepts of products. For instance, the effectiveness of the studies done by data-scientists [19] showed that it is beneficial to involve the user into the design this way. To cover up for the fact that normal designers won't be able to do certain data analysis, we saw that an architecture as shown in [7] will be the solution into helping those designers to effectively do their research.

Furthermore, we saw that the effects of bias in certain applications could be devastating to those affected by it. The fact that one person or certain groups are discriminated against by tools like COMPAS is dangerous and should be avoided in the future. Also, as regulations like GDPR try to prevent bias, the practitioners of data-centric research should be made aware of the risk of unfairness in a dataset or algorithms. Seeing these examples of where bias could lead to and noticing the need of practitioners of data-centric researchers, it is important to make those practitioners become aware of the dangers of bias. This should not be done by only telling them, but by showing them what could be the consequence of their actions and what aspects to focus on in a dataset.

Since our main target group will be data-centric researchers that are not actual data-scientists, we have to use metrics which are understandable and give clear indications of potential bias. In order to calculate biases in the exploratory system made we decided to use the AIF360 toolkit as it was well documented and provided the opportunity to calculate the bias with only the model predictions. As we want the users to become aware of the potential bias located in a dataset, we need to let the user explore a dataset by offering them several opportunities to learn more about it. For instance, let them make simple modifications to a dataset to see the difference in fairness when applying such changes.

Chapter 3

System Design

In this chapter, a detailed specification of the complete exploratory system is provided. Starting by determining what is needed in order to be able to specify all the parts that are required for this system. Also, with help of section 2.3, we are able to define and implement a process of exploring the dataset and finding bias. The system that will allow practitioners of data-centric research to explore datasets will be called FairData.

3.1 Design overview

In the previous chapter, we decided that we want a system for all practitioners of data-science where they will be able to become aware of bias by exploring datasets. Based on the different tools available and on what we want to achieve with a FairData, we decided to use the AIF360 toolkit as its bias metrics seems useful in clearly showing the potential bias that can introduced by Machine learning practices.

So to let practitioners of data-centric research become aware of bias, we want a system where the user will be able to explore a dataset to find out if there is bias present in the dataset and to learn what causes this bias. This will require an application where the user could go through multiple iterations of analysis and will be able to modify the dataset through basic modifications like filter and deletion of attributes. The result of these modifications should represent this bias through multiple visualizations.

3.2 Requirements for exploratory system

Now be able to define the requirements for the exploratory system. These requirements evolved over time by looking at what we need to analyse and 'explore' bias in datasets obtained by data-centric research. It is also inspired by what is possible by using the AIF360 toolkit and how we could extend on that functionality. The requirements are split up in multiple categories . We will start with the functional requirements and end with the non-functional ones.

3.2.1 Basic needs for dataset

First we go over some basic requirements for the datasets that were going to use in FairData, based on how we want to use those datasets:

- The user must be able to upload their own dataset to the system.
- The user must be able to select a dataset which is already uploaded to the system.

- The system should let the user do multiple exploratory sessions.
- The system must generate metadata when a dataset is uploaded to the system.
- The system should generate distributions on the different attributes in order to inform users about a certain attribute

These requirements focuses on some simple actions that a user should be able to do in this system. An important aspect in this is that the user must be able to upload datasets that are not too big in size and consists of personal data. After that the dataset is uploaded it should be able to detect the structure of the dataset in terms of attributes and their types.

The metadata generated should also consist of statistical information about the attributes which serves the purpose of informing the user on the distribution of each attribute its values. This metadata also consist of the amount of empty values located in each attribute and basic information like the size and shape of a dataset.

3.2.2 Basic dataset Modifications

This section focuses on the basic modifications to be performed on a dataset which will help to see changes in the bias. This will help the user to try out different stuff in order to try to reduce any potential bias.

- The user should be able to reshape the dataset by filtering out on numeric or string values.
- The user should be able to rename attributes.
- The user should be able to delete attributes from the dataset.
- The user should be able to substitute values if values are not present within certain attributes

All these actions are the basic modifications that are the minimal actions a user could take to reduce bias in a dataset. The first three actions are independent of the dataset itself and can be performed with every attribute. The substitute action on the other hand is dependent on what the system finds after the user uploads a dataset. This action is only available when an attribute in the dataset has rows that don't have any value which can be detected by the system.

3.2.3 Bias actions & results

The most important requirements of FairData are all about gathering the right information, so it is possible to calculate some bias and show the results of that. For that we came up with the following requirements:

- The user must provide information to be able to calculate any potential bias.
- The system must be able to calculate the bias with the provided information by the user
- The system must be able to generate the visual output of the bias calculations.
- The user should be able to select different kind of metrics in the results screen.

- The user should also be able to select dataset distribution plots to see how the dataset itself changed during the exploratory process.
- The results page of the system should also show metrics on a trained machine learning model based on the resulting dataset.
- The system must show the exploratory results per version of the dataset based on any modification made to the dataset.

To be able to calculate bias, we need to make use of the AIF360 toolkit. This toolkit requires additional information about the attributes besides the dataset itself which has to be provided by the user as that information is different for each dataset. We need both the primary attribute and the protected attributes as seen in the instructions of the AIF360 toolkit, for each of those we need to indicate different groups which can only be provided by the user. The way we will try to obtain this crucial info from the user will be discussed in section 3.5.5.

After the bias is retrieved from the AIF360 toolkit, we are able to show the result of the bias metrics in a result screen. With this results screen the user should be able to select the metrics they would like to see and to select additional distribution plots. Both of these visual representations will show how the dataset changed over time and make the user aware of what the consequence of their actions might be.

For that reason, we also saw that it could be useful to see the difference between datasets versions based on the dataset modifications made by the user. This idea was further evolved into a data version timeline based on the CHAMELEON interface designed by [12] which was made for data iteration in Machine Learning. We made use of this timeline so we can see the difference between two versions of the dataset, where each version represents one dataset modification. To make this work we had to calculate all the results for each version of the dataset.

3.2.4 Non-Functional requirements

Finally, we also have some non-functional requirements:

- The user interface should be an web application accessible through an internet browser.
- The average inexperienced user must be able to explore their dataset within an hour.

This type of system is clearly something you want to do in a browser as you don't want to do the calculations on your own computer. Besides that, it should be accessible for multiple practitioners of data-centric research. Also, the UI of this system should be clear and helpful for the user so exploring their dataset should not take more than half an hour.

3.3 Mock-ups

With the help of the requirements and based on the info the AIF360 toolkit needs, we continued the process of designing FairData by having multiple iterations of mock-ups. The first iteration of mock-ups were more complex than the final ones, as it also included an advanced analysis stage which would let the user also instrument Machine learning algorithms. This turned out to be too complex as a system for practitioners of data-centric research to become aware of bias. At this point it was

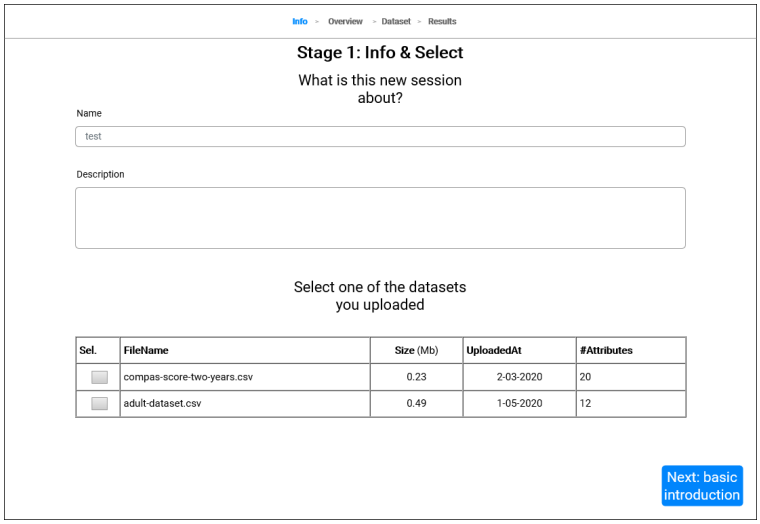
decided to use datasets that include the machine learning outcomes to step away from an actual data analysis including ML techniques and focusing on exploring the bias. It is important to note that we already started implementing the system at this point to test certain functionality and to build parts of the system that we needed anyway.

After that, we decided that it might be better to focus more on exploring our dataset in a way that the users of this system will be able 'explore' the bias present in a dataset. This second set of mock-ups had 5 stages including the results to explore a dataset. Besides an uploading and results stage, it had also separate stages for the basic modifications, more info on the attributes and to provide information for the bias. It was sufficient to see the bias present in a dataset, but we wanted the user to be able to better analyze the attributes of the dataset in a more convenient manner.

Eventually, we came up with a final set of mock-ups that focuses more on analysing each attribute separately. This was for the purpose of having a system where the user could explore the bias by themselves rather than having this fixed process where the user has to follow the instructions of every stage. For our final version of the system we also made use of the expertise of a data-scientist to see how it would interact with the issue of bias through a coding example working with the AIF360 toolkit. This gave us some additional insights into how we should represent the bias to the user which eventually resulted in the mock-ups discussed in the next sections.

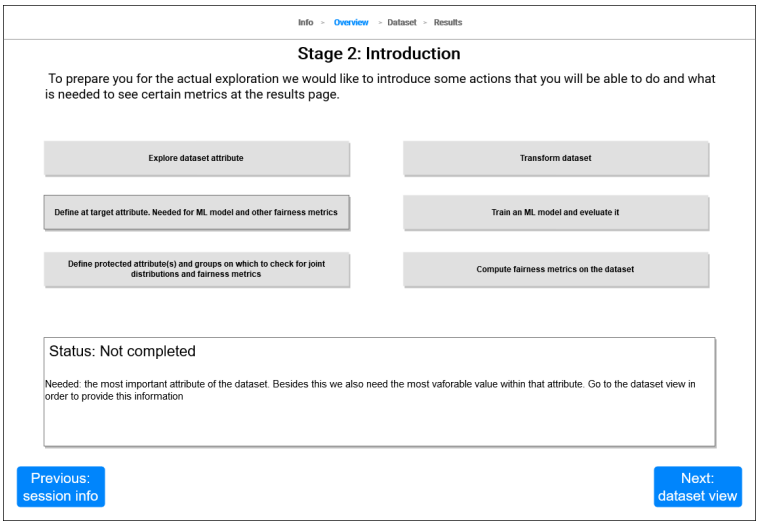
3.3.1 Session info & select a dataset

This is the first stage of the process where the session info should be provided and where a dataset should be selected. For the selection of the dataset, we choose to use a simple table of the available datasets based on the uploaded datasets.



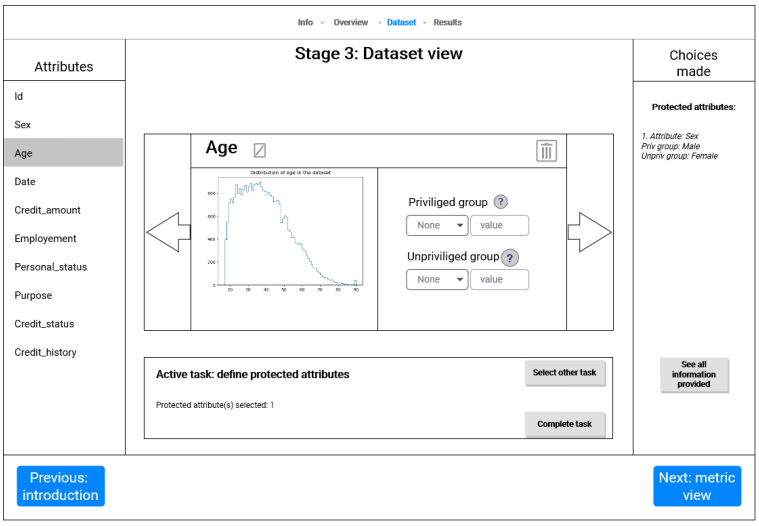
3.3.2 Overview

At this stage it is possible for the user to see all the things they would be able to see or do when doing an exploratory session. By clicking on the button they will see the tasks they will need to fulfil in order to get results out of it. The resulting metrics will be available for the user in the results screen.



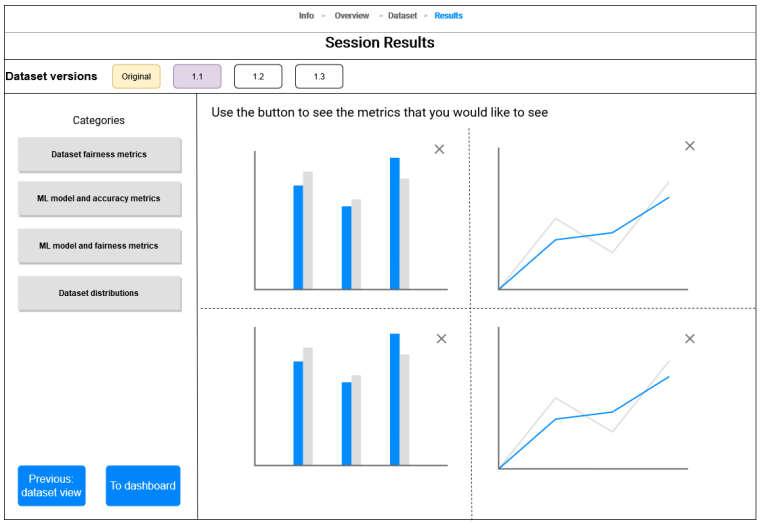
3.3.3 Dataset view

The most important stage of the exploratory process is the dataset view where the user will be able to provide data which is going to be used to calculate the eventual metrics. All the information will be supplied by following certain tasks depicted by the previous mock-up.



3.3.4 Metrics view (results)

Eventually we arrive at the metrics view where the user will be able to select the metrics they would like to see based on the information they provided. Besides that, it will also be possible for the user to select different versions of the dataset based on the possible modifications made to the dataset. In that way, it becomes possible for the user to see the impact of their actions.



3.4 Architectural design

With the help of the mock-ups and the requirements, we are able to come up with more detailed specification of FairData. Starting with the system architecture providing an overview of all the important components in the system. On top of that, we will also provide a flow diagram and a data model

3.4.1 System Architecture

Based on the requirements and on what is needed to make it work smoothly, we decided to use a layered architecture like shown in figure 3.1. The presentation layer shows two UI's, one for exploring a new dataset and one as one central dashboard containing all the uploaded datasets and exploration sessions per user. Besides that, we also have a business layer containing the DatasetProcessing service that is responsible for all the calculations done in the process. The following two layers shows how the data is stored in this system which will be contained by one database.

Starting with the dataset dashboard, this is just a central place providing the opportunity to upload datasets and to start a new exploration session. The upload part is also the part where all the general metadata will be obtained by the Dataset Processing service. If a dataset is successfully, it will be possible to start a new exploratory session which will navigate the user to the DatasetProcessing UI. The dataset processing UI will cover the rest of the exploratory process dividing it into different stages. This part of the system will also make use of the DatasetProcessing service to retrieve relevant data and calculate the results.

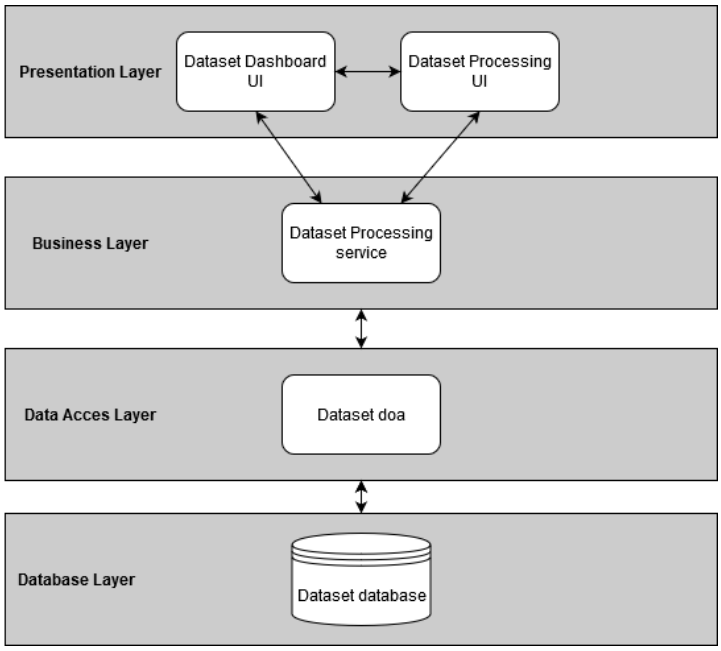


FIGURE 3.1: System Architecture

3.4.2 Flow diagram

In order to visualize the process flow of the system, we created the following flow diagram in figure 3.2. The process visible shows the three different actions the user could start with which is either upload a new dataset or go to the exploratory parts of the system. The process flow part of a new exploration session shows the different stages needed in this process in order to get to the results part. Important to notice is that it shows the possibility to make multiple iterations with a certain dataset.

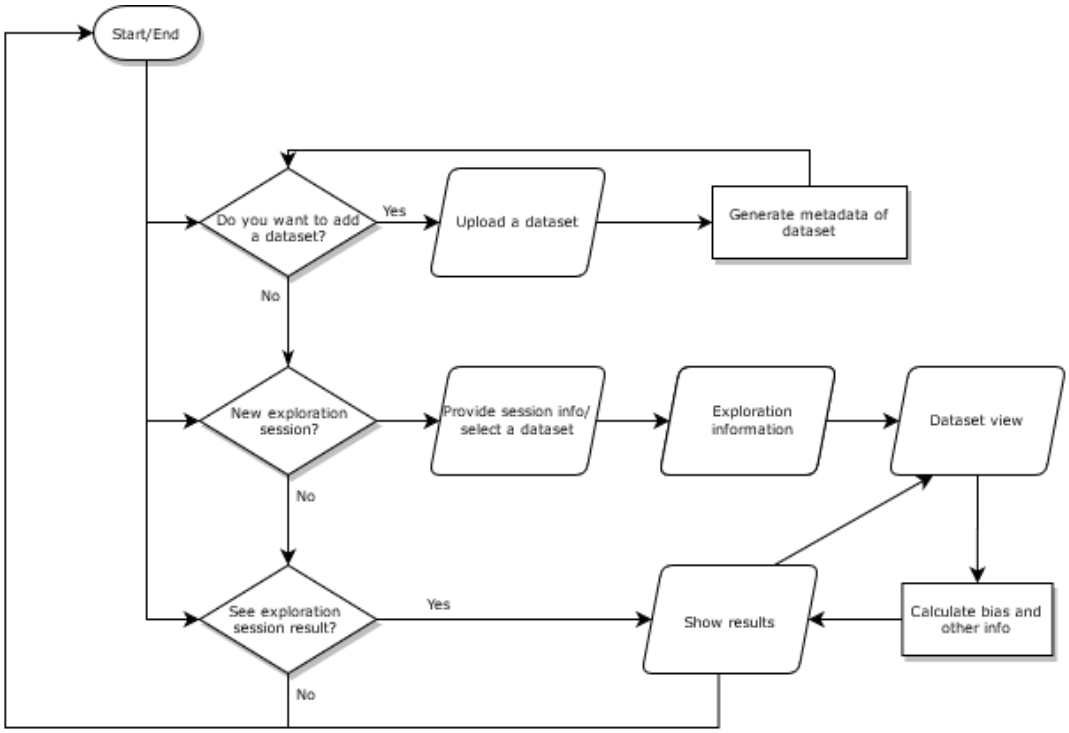


FIGURE 3.2: Flow diagram of the system

During the creation of the flow diagram, we thought that it might be useful to have a dynamic object (called DataSelection) considering a dataset which is an extension of the metadata and tells the system what part of that dataset the user would like to use. It will be an extension of the metadata objects defining what data should be retrieved. Furthermore, these DataSelection objects could help to depict the different states of the exploratory process and could be created in such a way that they are reusable when a user completes one iteration of this process.

3.4.3 Data Model

The data model visualizes the several important entities needed in the system and is presented in figure 3.3. This data model is based on the process shown in the flow diagram based on what is needed for the AIF360 toolkit to calculate the bias.

The system will start with converting raw datasets into Metadata files which can be seen as unedited datasets that will be selected at the start of an exploration session. These Metadata files will become DataSelection objects when import actions are taken causing the dataset to change. These objects will contain Action objects that depicts dataset modifications like filter an attribute value or delete attributes. The DataSelection object also contains the representation of the attributes and other important info about the state of a dataset.

Furthermore, there is also the SessionResults object being made as a result of this exploratory process. This will at least contain the results of the bias measurements and comparisons between the important attributes. The comparisons will help to give a better insight on how the user its actions changed the distribution of the dataset.

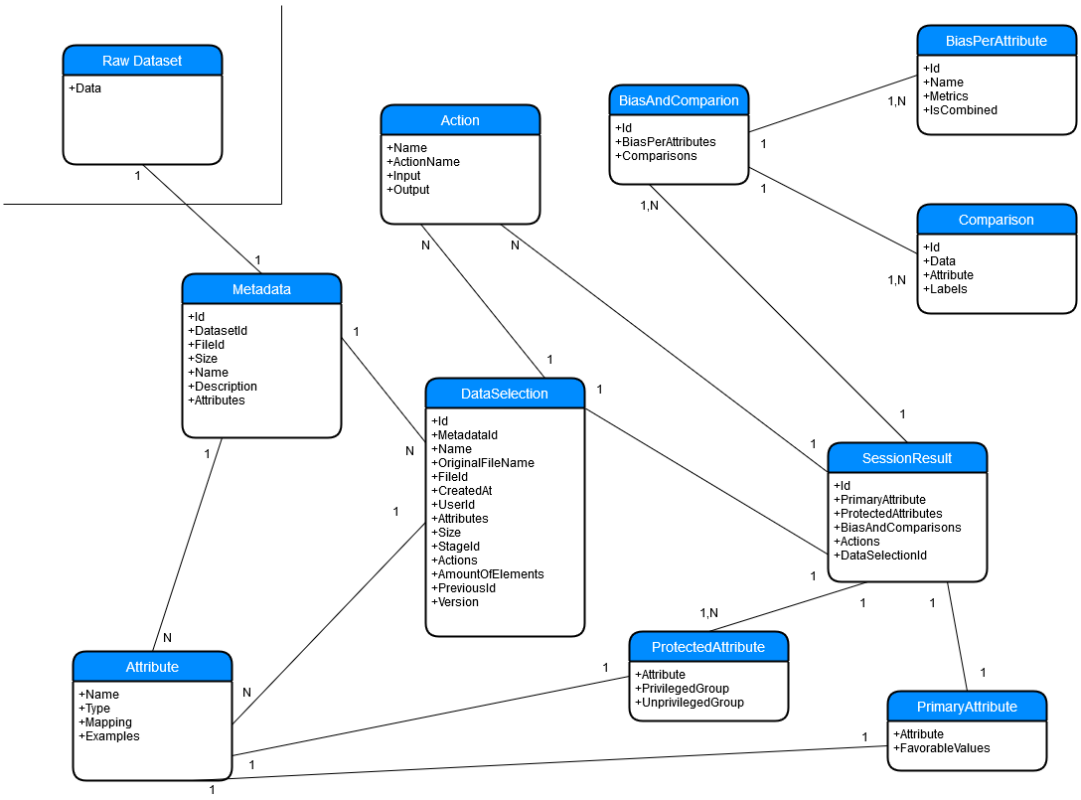


FIGURE 3.3: the DataModel of the system

3.5 Detailed design

This section will go more deeply into the details of the modules shown in the system architecture. Starting with the Dataset Dashboard UI which is a simple UI where the user can manage its exploration sessions. Secondly, we will dive in to the Dataset Processing UI which is the main UI component of this system together with the Data Processing service. Eventually, I will also define a more detailed specification of the several parts in the data access layer and the database layer.

3.5.1 Dataset Dashboard

Here we will focus on giving the user a separate UI to manage the exploratory sessions and let them upload their own dataset. So first, the user will have to start with uploading their dataset at the 'dataset' section of this UI. The backend implementation of uploading the datasets should therefore be able to handle CSV-type dataset files. When uploaded to the system the Dataset Processing service will create a Metadata object related to the uploaded dataset. This Metadata object could be seen as the original version of the DataSelection object and will generate simple distribution plots on the attributes of the dataset. These distribution plots will come in handy during the exploration sessions.

The dataset uploaded in this part of the pipeline will be read by the service part with the help of the Data Analysis library called Pandas[1]. This library is needed later on when working with the AIF360 toolkit, but also offers a well-structured data object called Dataframe. Besides that, it also offers simple ways to discover the

[1]https://pandas.pydata.org/

data types that are present in the dataset and offers basic statistical functionality that can be used at other stages of the process.

The second part of this UI is all about the exploratory sessions. The user could either start a new exploration session or they can see the results of an older exploration session which will be showed with the help of the Dataset Processing UI.

3.5.2 Dataset processing pipeline

This part of the system is responsible for exploring uploaded datasets and finding potential bias located in this dataset. In order to do that we will be running the user through several stages based on what information is needed and based on what they should be able to do. For the purpose of finding potential bias we looked at the AIF360 toolkit to determine what info is needed and to determine how we should change the dataset, so the toolkit could work with it. The several stages of this part of the system are depicted in figure 3.4 and will be discussed in the next sections.

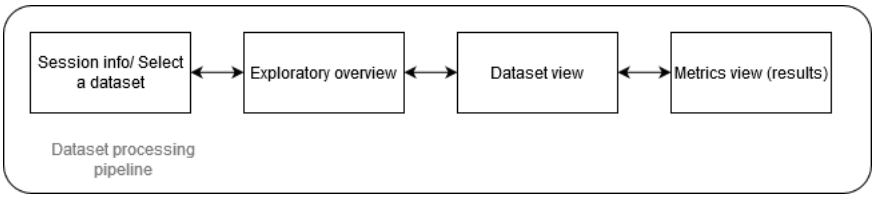


FIGURE 3.4: The Dataset Processing pipeline

3.5.3 Session info & select a dataset

This stage consists of two pars, first it is just about providing some basic information to be used to indicate a new exploratory session. The UI of this part will just consist of 2 simple input forms to provide a name to the session and to provide a description. The session name will also be the name of the DataSelection objects.

Secondly, the UI will let the user select the dataset they would like to explore. Based on the datasets uploaded in the Dataset dashboard, the user will be able to select one of the datasets. Besides that, it might be also useful to be able to select new versions of the datasets obtained by doing an exploratory session, so the user would be able to iteratively explore uploaded datasets.

3.5.4 Introduction stage

The purpose of this stage is to provide information to the user of all the actions it could take during this process. This is just to inform the user what they could do in the dataset view stage of this process. Instead of helping the users in taking certain actions we let them determine themselves what might be useful for them to see or to calculate.

3.5.5 Dataset view

Guided by the previous stage, we now let the user change the dataset itself by providing them an UI containing the several buttons discussed in the introduction stage. These buttons will be combined with a view that shows a form per attribute, based on the button pressed. Each of these buttons represents a certain task to be done and requires the user to provide the correct information. Below we will discuss all the actions to be taken by the user.

Primary attribute action:
The next two tasks are about providing the crucial info to the AIF360 toolkit. With the help of the AIF360 toolkit we identified that at least two parts of information are needed to calculate any potential bias. The first important part that is needed is the primary attribute of a dataset and its favorable value. This attribute is the result of ML predictions after training a model and testing it which we will not do in this system as spoken about earlier. That's why we will make use of datasets that already include a certain prediction like the COMPAS or German credit datasets.

Besides that it is also important to indicate what value of this attribute could be considered to be favorable when looking at the complete dataset. This part of the UI will let the user select one primary attribute selected out of the list of attributes. The favorable value(s) should be provided besides that.

Protected attribute action:
Furthermore, the attributes on which subgroups of people can be discriminated against are called protected attributes. Typically, these are attributes like race, age, sex and religion but this could differ depending on the dataset that is provided. The most crucial part of these protected attributes is that they should be split up into two groups in order to be able to calculate the bias. Besides that, it is also important to indicate which of these groups is privileged or unprivileged. For example, when picking race as protected attribute we could choose to pick "white/Caucasian" as the privileged group and decide to have the other set of race values (non-"white/Caucasian") to be the unprivileged group.

Explore & Transform actions:
These actions are all about exploring the dataset and let the user modify certain parts of the dataset. The explore action will focus on going through the attributes and try to learn more about this dataset. For every attribute we are able to see the distribution of its values, the percentage of missing values and also their relation to the protected attributes. This last piece of info will only be available after the user provided the task on the protected attributes.

For the transform task there will be several actions possible to execute on the attributes for instance changing its type, renaming, filtering, deleting an attribute from the dataset and substituting missing values. The first two actions mentioned do not really change the dataset and thus won't have too much influence on the results section of this exploration session. However, the actions of filtering, deleting and substituting are can make a difference between being a dataset with or without bias.

The action of filtering will let the user remove a part of the attribute values from a certain attribute which will either be a range of values when working with numbers or it will be a single attribute value in case of an attribute that consists of string/categorical values. Besides that it is important for the user to keep track of

missing values located in attributes as this could badly influence the outcomes of the bias calculation. This is because the AIF360 toolkit will remove rows from the dataset that contain so-called 'null' values before calculating the bias. So it is important to either remove those attributes or to substitute those values by providing a value or to select from a list of mathematical functions (depending on the type of this attribute)

When the user continues to the next stage, all this information is stored into a new DataSelection object to also point to the dataset file as shown in 3.3. The dataset file itself will be stored as a 'Parquet' file which is an Apache Hadoop's columnar storage format but can also be used by other well known systems. The main benefit of using this type of storage is that it is very good at storing large volumes of data and that it does not take too much time to load this dataset when needed [23].

3.5.6 Bias results:

Lastly, we arrive at the bias results stage of the system. This part will show the results of the finding potential bias similarly as the AIF360 toolkit demo shows. For each metric used, a diagram will be shown indicating how it scores compared to a bias threshold. Biases will both be calculated per protected attribute and all the protected attributes combined, because both will be relevant to the situation.

Alternatively, there will be also the possibility to see how the distribution of protected attributes with respect to the primary attribute changed. This will help the user to gain additional insights on how the bias changed.

3.5.7 Dataset Processing service

For the dataset processing service I found out that the web framework Flask [2] was very suitable as we need to use Python because of the AIF360 Toolkit. This service will handle all the important calculations related to the uploaded dataset and will also store all the data showed in figure 3.3 in a MongoDB. This database is used for both normal JSON data as for the storage of figures and other files. We choose this type of database as it is simple and can also store different types of files.

[2] https://github.com/pallets/flask

Chapter 4

Implementation

In this chapter, we discuss the implemented exploratory FairData by going over the different parts of the system. First we will cover the Dataset Dashboard part of our tool as it crucial to first upload a dataset before you can explore it. After that, we will look at the Dataset Processing UI and see how the exploratory session are implemented.

4.1 Dataset Dashboard

This is the central part of the dataset exploration system where the user could log into, manage their datasets and exploratory sessions. The Dashboard is created with the help of a simple template [1] as it was added to the system at a later stage of the design and we didn't want to spend too much time on creating a simple dashboard style UI.

After logging in, the user will arrive at the introduction of the system as a whole which will shortly describe what the user could do. Besides that, the Dataset Dashboard contains two other pages called 'Datasets' and 'Explore' that will be further explained in the next sections

4.1.1 Datasets page

The purpose of this page is to manage a user its datasets so they can be used in the exploratory sessions. This part of the UI is actually very important as it will try to detect the structure of the dataset and it will generate metadata to be used in the exploratory sessions. As can be seen in figure 4.1, this page will open a new form for uploading a dataset when clicking 'New File'. When the user uploads its CSV-file, the system will try to load this CSV and save the Pandas Dataframe in the database. The back-end will return the ID of this dataset so it can be used when the user completes the rest of the form. If the user successfully uploads a new dataset and if the name for this dataset is supplied, then they will be able to start the generation of metadata.

[1] https://github.com/akveo/ngx-admin

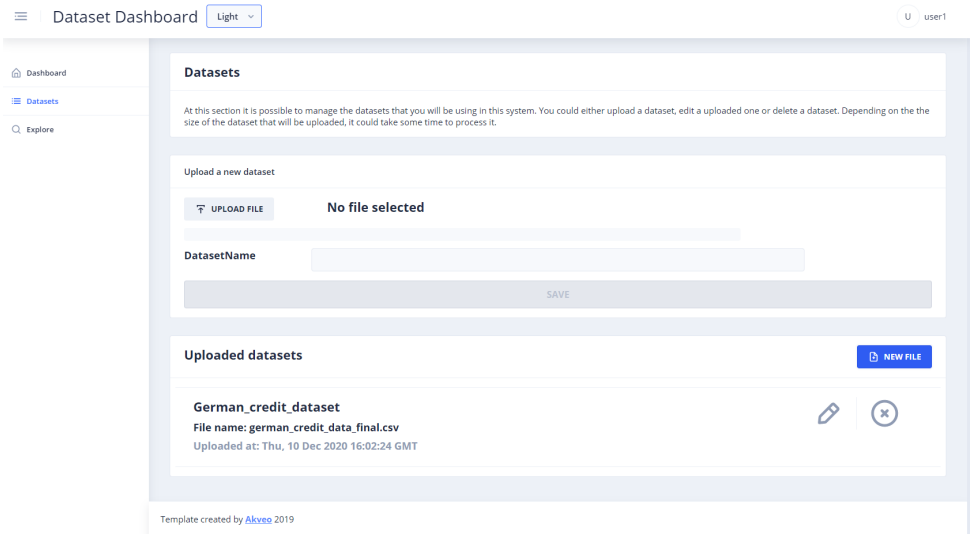


FIGURE 4.1: The datasets page of the Dataset Dashboard UI

The part of generating the metadata is performed by the Dataset Processing service and will start with loading the uploaded dataset into a Pandas Dataframe. With that we were able to retrieve the attribute objects as seen in the Data model and we are able to obtain the distributions per attribute. Eventually this is combined into a single Metadata object and saved to the system so it can be selected during an exploratory session. Uploaded dataset are connected to a single user so they will only be available to that user and these datasets will be retrieved every time they go to this page.

4.1.2 Explore page

Exploring your datasets starts with going to the Explore page which is simpler in terms of back-end. This page will provide the opportunity to start a new exploratory session or to see the results of an older session. Exploratory sessions are linked to user that is currently logged in. Besides that, it is possible to remove older sessions as can be seen in figure 4.2.

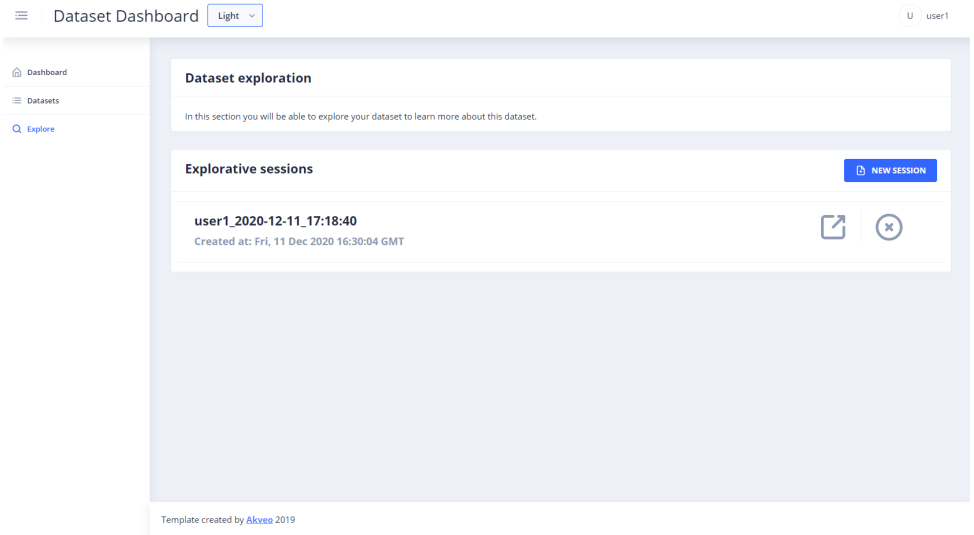


FIGURE 4.2: The explore page of the Dataset Dashboard UI

4.2 Dataset Processing UI

The second part, the dataset processing UI is the most important part of the system as it contains the actual exploratory sessions. Compared to the design chapter we will see several differences in the UI as the eventual system is focused more on the evaluation off the our tool which means that several parts are simplified. Also, as seen in section 3.5.2, the user is able to do multiple iterations of stage 3 & 4, so they can play with the potential bias present in the dataset and find a way of reducing it.

4.2.1 Stage 1: Info & Select

This first page of the exploration session is pretty simple as it is about providing session info and selecting the dataset to work with. In figure 4.3 we see the first part where the user is able to provide a session name and a description which will be used as attributes on the exploration session objects. Furthermore, there is also the Dataset table which lets you select a dataset based on all the datasets linked to this user. The datasets visible could also be a resulting dataset from another exploration session.

FIGURE 4.3: The exploration session info form

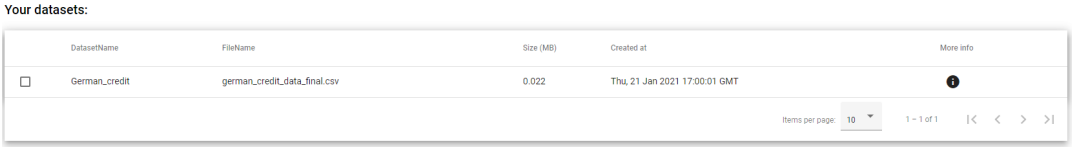


FIGURE 4.4: The table with a user its own datasets

4.2.2 Stage 3: Dataset view

This stage of the exploratory process lets the user 'explore' their dataset by being able to look at all the attributes and by being able to provide new info or modifying the dataset. The interface of this stage changed a lot because of the purpose of this tool and the fact that both type of users should be able to do the minimal exploration steps to become aware of bias. That's how we eventually introduced the attribute 'card' as seen in the UI in figure 4.5. This is a simple way to show the all the info about the selected attribute and let the user provide information/modify each attribute.

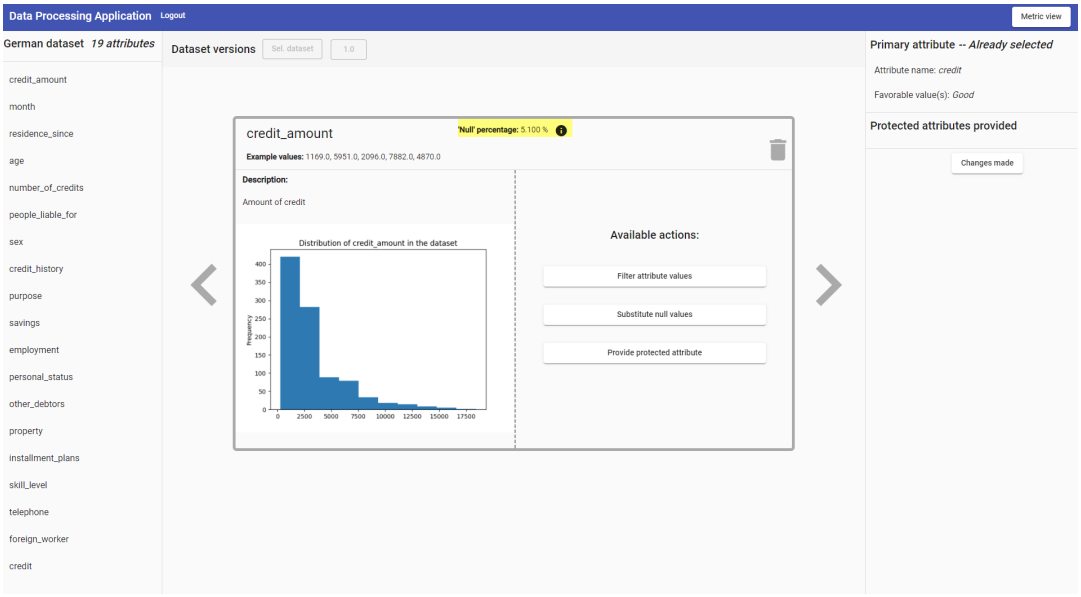


FIGURE 4.5: The dataset view of a new session

Besides the attribute card, there is a column at the left for selecting the attributes and there is a column at the right for seeing what is provided by the user and the system. The purpose of these two columns is to guide the user and to inform them on their own actions. Moreover, there is also the Dataset versions toolbar as spoken about earlier is section 3.2.3 which also has its purpose of informing the user on how many versions of the dataset there will be. In the following sections we will introduce the different actions available in this interface.

Primary attribute

As you can see in the image, the primary attribute is already supplied as it is very crucial to the calculation of bias. In a more open version of FairData this would also be something the user has to provide by having an additional form for implementing ML techniques. More information on the provided primary attribute will be given in the next chapter.

Protected attributes

This form is visible in figure 4.6 and lets you supply the protected attributes. Based on the attributes their type, it is possible to create different forms to split up the attribute into two groups. In the case of integer or float type attributes it is also possible to select a range of values as one group. The default form is always about providing one or more values per group which is also the case in figure 4.6 where Male is the privileged group and Female the unprivileged one.

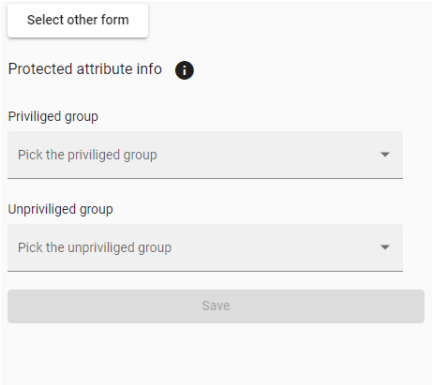


FIGURE 4.6: The protected attribute form

The implementation of these types was very important as it is crucial to the bias calculation and eventually we implemented it as a separate ProtectedAttribute object which made it easier to use throughout the system. The most challenging thing with this attribute is that the it has to split up those attributes into the two groups, as there are many possibilities related to the attributes their types. For this reason, we decided to automatically convert string types into categorical ones.

Filter an attribute

Filtering an attribute can be done as seen in the form of figure 4.7 which also differs depending on the type of attribute. For the numerical types it is possible to give either a range of values or a single value. In the case of the categorical types it is possible to select one of the attribute values. Both the filtering and substituting are actions which will be saved in the Action objects that are used to indicate the different dataset versions.

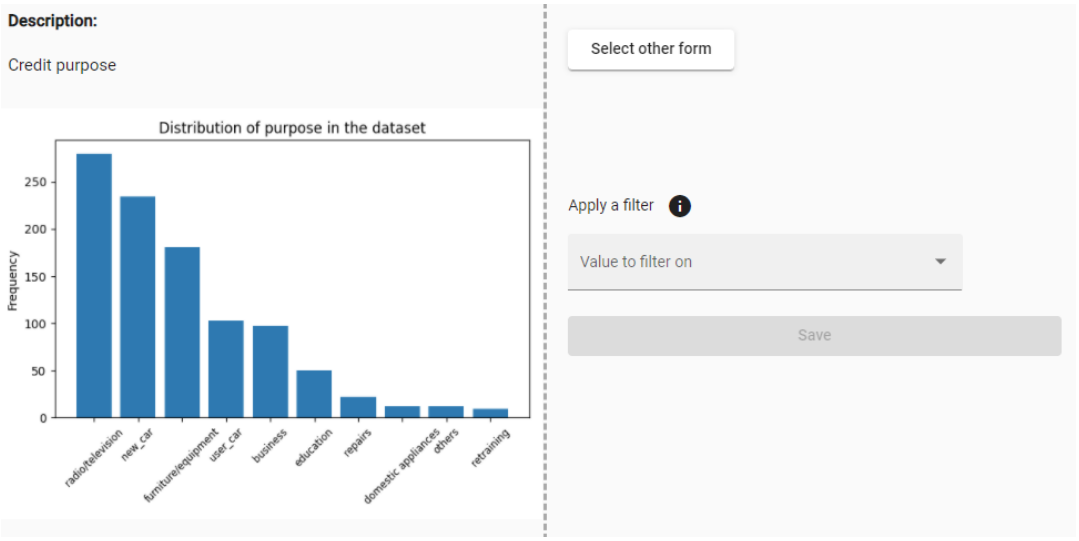


FIGURE 4.7: The form to apply a filter on an attribute

The implementation of this action was relatively easy as it is something which is already supported by the Pandas Dataframe. On the other hand, it may result in the issue that too large part of an attribute is filtered which reduces the dataset its size drastically. Even with the distribution plots visible for the user they could easily make the mistake of filtering the wrong values.

Substitute a null value

During the analysis of a newly uploaded dataset the system also checks for the amount of 'null' values in a particular attribute. This percentage on how many 'null' values is shown at the top middle of the attribute card and could be fixed by using the substitution form seen in figure 4.8. There are two different options with this form: First one is just providing a single value to replace all those null values. Secondly it is also possible in the case of numerical types to select a mathematical function like average to substitute those 'null'-values with.

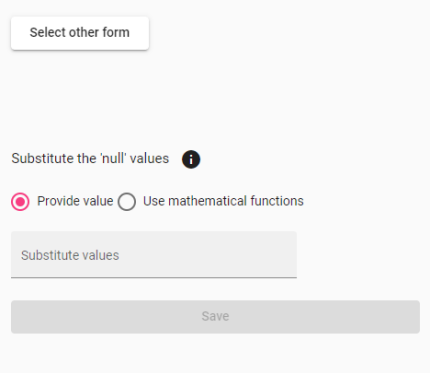


FIGURE 4.8: The null substitution form

Similarly to the action of filtering attribute values, this was also relatively easy to implement with the help of the Pandas Dataframe. This action also has some significant influence on the bias especially when this attribute is correlated to any of the protected attributes.

4.2.3 Stage 4: Metrics view

The last stage is all about the results of the exploration session. If there is enough information provided, the system will be able to show different types of plots available to the user. For instance, all the plots in the metrics view require at least one protected attribute to be provided. Besides the plots, it is also possible to make use of the dataset version toolbar which can be seen in figure 4.9. It is possible with this toolbar to select the different dataset versions the user would like to be represented in the plots they would like to see. The different types of metrics will be discussed in the following sections.

FIGURE 4.9: The dataset versions toolbar

Bias metrics

The first group of metrics are the bias metrics that are calculated with the help of the AIF360 toolkit. In figure 4.10 an example is shown of how such metrics will look like. The bias metrics consists of at most two values based on the versions selected in the dataset versions toolbar. All the bias metrics are calculated per dataset version for both each of protected attributes and all those protected attributes combined. That's why when clicking on of the bias metrics the user will be able to select on of these options.

FIGURE 4.10: Two different bias metrics

The implementation of the AIF360 toolkit in the Dataset processing service was pretty difficult as we had to shape the dataset in such a way the toolkit could handle it. It also required specific input about the attributes that should be included in the calculation and it required information on the attributes that are string types. As seen with the protected attributes, the user will have to split up those in to the designated privileged and unprivileged groups which had also to be done by the back end of the system.

Especially, before automatically converting string types into categorical ones, it was very sensitive to errors due to the fact that the toolkit can't include string types into their bias calculations and thus everything had to be transformed before calculating the bias. Besides this, there was also the issue of 'null' values and the fact that every row containing one will be removed by the implementation of the AIF360 toolkit. As this could lead to an almost empty dataset we decided to remove the columns instead to save the bias calculations.

Machine Learning metrics

Secondly there is the Machine Learning metrics which makes use of the Bias metric structure and data to calculate different metrics based on a ML model. This is to both illustrate the use of this toolkit with Machine learning and to give the user additional information on the potential bias in a dataset as seen in figure 4.11. It's important to note that this part of the system when looking at the back end implementation, is fixed in the sense that it uses a specific classifier. This would be different in a system where the user would be able to instrument the Machine learning model themselves.

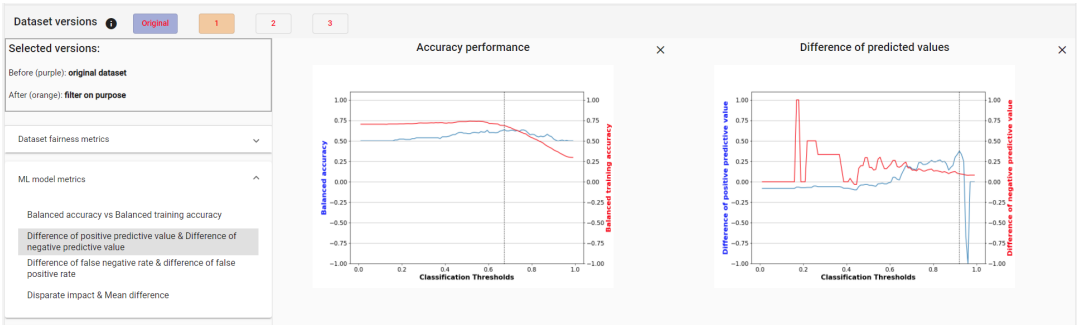


FIGURE 4.11: Two different metrics based on a ML model

The output of the Machine Learning metrics is different because the metrics are calculated for each value of an array of thresholds. These thresholds are used to compare with the prediction scores so the AIF360 toolkit could calculate any potential bias based on that learned model. The reason we already create the plots in the back end is that these resulting arrays of metric values could slow down the user interface.

Distribution plots

The final type of plots is a lot easier in terms of implementation, as it will show the distribution of the protected attributes compared to the primary attribute. For the purpose of showing how these groups of protected attributes are changed by the dataset modifications performed on the dataset, an example of these type of plots is shown in figure 4.12. So the main purpose of this plot is to give additional information on how the dataset changed with the attributes which are sensitive of bias.

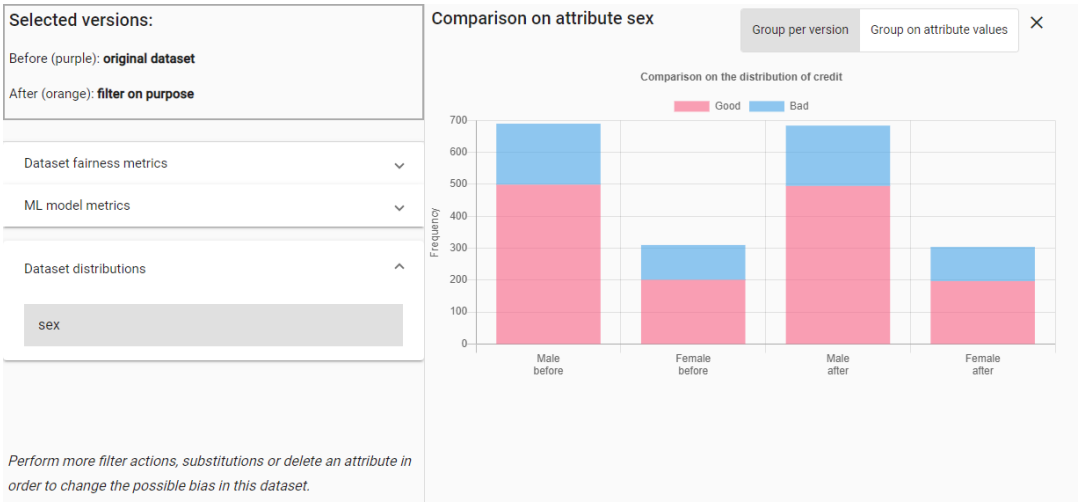


FIGURE 4.12: A distribution plot on the attribute sex

Just like the bias metrics we only have to save a few values, so the graph can be created in the front end. The back end implementation of this part just uses a predefined attribute count function to gather all the required information. The corresponding plot is also able to group the stacked histogram in two different manners as can be also seen in figure 4.12

Chapter 5

Experimental setup & results

This chapter discusses the plan for doing experiments with FairData in order to gain insights to be able to answer the third sub-research question. This sub-research question is "How does the application inform the user about potential bias located in the dataset and does it support the practitioners of data-centric research to reduce this bias?". So the experiment is about the qualitative aspects of our tool to 'explore' datasets as we used it to observe what different practitioners of data-centric research would do with the dataset in the context of designing a machine learning algorithm. In the following sections we provide an explanation on what the actual context is and how it is evaluated. In the last section we also present the results of the experiments.

5.1 Context

To gain insights on how the user could interact with unfairness issues introduced by a dataset, we wanted to do an experiment where we can observe the interaction of practitioners of data-centric research with the exploratory system we made. Thus, it is not about evaluating the FairData, but more like using this system to see how practitioners think and interact with the issue of bias being present in datasets. Hence, we have to observe their behavior during and interview them afterwards to be able to answer the third sub-research question.

Therefore, based on a dataset that includes the machine learning model output, we decided to create a specific scenario of helping a bank to analyse data for the design of a machine learning service for loan attribution. The dataset used contains financial data about the bank its customers and the decision made by the machine learning algorithm they designed. The bank wants participants of the experiment to inform them about potential fairness issues related to their algorithm and the dataset. The creation of this scenario was inspired by the German Credit dataset that will be explained in next section.

5.2 German Credit dataset

The dataset used in the scenario of the experiments is called the German Credit dataset [11] which is a dataset used to classify people as good or bad credit risks based on financial personal data. This dataset was chosen as it was very suitable for the type of experiment we wanted to do, due to the fact that the principle of loan attribution is common knowledge for all practitioners of data-centric research.

We also wanted to use this dataset since it has only a thousand records which will save time calculating bias and is relatively easy to work with both for the user and

our system. Besides that, it contains all the type of attributes we need for this experiment and it provides enough attributes that can be possible protected attributes. The only issue with this dataset is that filtering out certain subgroups in the dataset will remove to much records which will result in training a model based on a handful of people.

For this experiment we modified the German Credit dataset in such a way that it would benefit the experience of working with FairData. First of all we did not want the participant to spent time on preparing the dataset before using it to calculate any bias, for example the original dataset uses numbers to differentiate between the categories in an attribute. This is also to reduce potential errors as those numerical values could introduce bugs or incorrect results when the participant has to use them as integers. They are also used for indicating privileged and unprivileged groups when providing the protected attributes.

Also, at one of the attributes called 'credit_amount' we introduced 'null'-values into that column so the user would also be able to use the substitution action. This provides more opportunities to the participant when they find and change the bias besides the simple filter and deletion functionalities. On top of that, it is also important to the user to see that substituting these missing values has its effect on resulting bias measurements.

Additionally, the experiment only focuses on the exploration stages of the system as it is the most relevant to the sub research question that we want to answer. Therefore, we only give the participant access to the DataProcessing UI as they wouldn't need to change the dataset to be used. Furthermore, we also replaced the first two stages with one welcoming screen which has the purpose of loading the dataset and we already supplied the user with the primary attribute. This is the attribute called 'credit' which is the most important attribute of this dataset.

5.3 Procedure

For this experiment we involved multiple practitioners of different groups of students/researchers which includes designers (Industrial design) and data-scientists (computer science). These groups represent the different people involved in data-centric research practices as we have seen in section 2.1. We look at both groups as we want to observe how they interact with datasets that contain bias. On top of that, we also want to see the difference between those groups as it helps to see how they could be supported when doing data-centric research.

Both group of participants are students/researchers from the TU Delft recruited as they had some affinity with the topic of machine learning. In the case of the computer science students, it was necessary that they have experience working with machine learning techniques and analysing data. The industrial design participant on the other hand were recruited if they have some experience with machine learning through following some elective courses. An overview of the participants involved in this experiment is given in the table below.

The task of those participants is to explore/analyse the dataset by using the FairData and see how their interactions with the system makes them aware of the issues with this dataset. After a short introduction on the context and the experiment, we asked them to analyse the data and think out loud about the actions they take. This is both for the purpose of supporting them if they need help and for the purpose of observing their reasoning when they use the system.

Participant	Faculty	Education	Experience
CS1	Computer Science	Master Student	ML/Datascience courses
CS1	Computer Science	Master Student	ML/Datascience courses
ID1	Industrial Design	Master Student	ML electives
ID2	Industrial Design	PhD	ML electives

TABLE 5.1: Table about the participants of our experiment

One session took about 60 minutes which starts with 15 minutes of introducing the experiment in general and telling the participant about the design brief. This also includes a quick introduction to the tool itself so they can start using it right away. After that, the participant started analysing for 30 minutes and we observed the actions they take as they are also asked to think out loud. Eventually, we took 15 minutes to interview them about the experiment which helped to evaluate this experiment.

5.3.1 Observations

During the experiment, especially when the participants start with analysing, we focus on what actions they take and how they interact with the system. We are expecting to see significant differences between the two groups of participants which is primarily due to their experience on working with datasets and using those in combination with machine learning. Besides that, we expected that the design participants have more trouble with some of the concepts related to bias as this is a more advanced topic related to machine learning.

5.3.2 Interview/Questions

Towards the end of the experiment we interviewed the participants on multiple aspects related to this experiment. These questions will be about both their findings with respect to helping the bank and the actions they took during the experiment. The questions asked during the interview are listed below in separate categories.

About the scenario
These questions are focused on the scenario introduced in the experiment where the participant is asked to analyse the dataset for the bank. This is the most important set of questions as it is directly related to the third sub-research question whether the application informs the user about bias located in the dataset. Analysing the dataset through FairData should result in making the user aware of bias as the dataset is not 100% fair when used to train the model for automatic loan attribution. Furthermore, it also focuses on how the bank and companies in general should responsibly address these type of situations and who should be involved in this process. This is to see and confirm which people should be involved in machine learning practices like this.

- What do you conclude from your exploration of the task/dataset with the tool?
- Do you think the bank could automate its system?
- What would be the difficulties of the bank?
- What would be your concerns?

- Do you feel responsible for the "fairness" of the application, of the data, of the ML algorithm? Or who should be responsible? Should it be shared between the data developer and the ML developer?

Their actions
The questions listed in this category are more focused on the usability and the understanding about this tool during the analysis. It is important to know how the user interacts with the tool in order to be able to make an even more successful tool. Also, it is good to know if there were features that the application is missing which could benefit their exploration of bias.

- What interrogations did you have when using the tool?
- What kind of challenges did you encounter?
- Were there things you would have liked to do but could not due to time, or due to the tool being limited? If there were, could you elaborate on that now, what would you have done and why?
- Why did you perform this task first? Why not this one?

Attributes
These questions and the ones in the following categories consist of a couple of follow-up questions in case the participant did not consider certain actions during the experiment. This category is more focused on how the participants think about certain attributes present in the dataset and how they think about certain attributes being used in training the model for loan attribution. It is important to know whether the participants have some ethical concerns not only about the attributes as a whole but also about the attributes its values. On top of that, it is also interesting to see which attributes they select as protected attribute and thus think that those attributes are sensitive to bias.

- Did they think about the relevance of each attribute and the primary attribute for a learning task? Which issues did they find? No ethical issue? Why or why not thinking about that? Actions to take: Would they want to collect more attributes, or remove some?
- Did they think of the values each attribute can take? What kind of issues could they find? (inclusive values or not?) Did they think of the limitations of the data, maybe they can't be applied to the "whole" population, but only to a subset? Why or why not thinking about it? Actions: Would they want to collect more data?
- Did they consider various protected attributes/groups? What were their questions? Did they think the formalization is too simple with two groups? Did they think about intersectionality? What was challenging for them?

Distributions & metrics
Another important part of the experiment with our system are the different distributions and metrics available to the participants. Here it's important to learn how they use the distributions to learn more about the attributes and to see how it helps the user in taking the right actions. Also, for the metrics it is important to know if the participants are able to correctly interpret those as they are crucial to understanding the bias.

- Did they think of how representative the data are of the real world population? Or of the desired decisions? (these two might be different) especially, with the fairness notions, did they look at distributions across protected groups?
- Did they look at distributions for protected groups and labels (credit)?
- Did they think that majority populations might not be easily trained on since we would have less data for them? Actions: what kind of actions would they imagine taking with regard to this issue of representativity? (does it fit the bias framing?
- Did they consider multiple fairness and accuracy metrics? How did they choose them? Did they think of their meaning in the context? Of their simplifications? What was difficult in using these metrics? Interpreting them? Who is responsible for choosing the metrics to optimize for?

Data transformation / metric optimization
This last category of questions is more focused on the second half of the third sub-research question whether the UI supports the participants to reduce the bias. This is all about the errors located in the different attributes, how they handle them and how it could impact fairness.

- Did they search for errors in the data? Did they search for missing values? Outliers? Maybe duplicates?
- How did they want to handle them? Simply remove? Transform quickly? Or would they want to collect more data?
- What kind of transformation did they do? Why? How did they or would they want to handle the data errors?
- Did they recompute the fairness metrics after each manipulation of the data? Why or why not? When did they recompute them?
- Did they understand any transformation might impact fairness, inclusivity, etc?

5.4 Results

In this section we report and discuss our observations made during the experiment and based on the interviews done afterwards. This will be done by discussing the two different groups of student/researchers as there are some differences between these groups when interacting with the system.

5.4.1 Computer science participants

After checking out the UI and seeing what is possible, both participants start with analyzing the dataset in a similar way. The participants look through the attributes and quickly start with doing either the modifications or providing the protected attributes.

CS1 for example decides to pick *Sex, Foreign_worker, Personal_status* as protected attributes where CS2 applies filters on *age, savings, credit_history*, substitutes null values on *credit_amount* and removes *personal status*. After that CS1 applies filters on

purpose and *age* where CS2 provides *sex, credit_history, other_debtors, foreign_worker* as its protected attributes.

In both cases, the protected attributes are chosen based on what they think is sensitive to bias looking at the distribution plots per attribute. This is similar for the filter they apply as they both decide to filter out everything above the age of 80. They both thought that persons above this age are outliers in this scenario of analysing for the bank.

Due to a error in the system, CS2 had to re-do all modifications which costed some time, but after this it was possible to go to the results page and to start analysing the results. After that, the participant went through the different metrics and modifications performed in order to "look for a positive trend in the metrics".

For participant CS1, it is not completely clear what the significance is of the bias shown in the metrics. After we provided some more info on the metrics it becomes clear to the participant that there is some bias shown in the metric view. This lack of information is something which could be easily added to the system and as CS1 suggested it would be nice to have formulas of the metrics.

The newly calculated correlations helps participant CS1 to select a new protected attribute (*Skill_level*) which indicates that the correlations can be very helpful. Besides that, a new filter is applied on *savings* which changes the bias, but also removes a decent amount of records. This is something on which the user of FairData should be informed about when making these kind of big modifications to the dataset as it negatively affects training a model when there are only a small amount of records.

Now that they tried to reduce the bias by applying different filters and checking out the results of their action, we wanted to know more about what they think the bank should do in such scenarios. Both participants think that the bank could automate the service, although they also think that it would require some additional analysis to make it even more fair for the bank its customers. They think that this should be mainly done by focusing more on attributes that are more crucial to the machine learning models. Participant CS2 also stated that it might be good to involve people from the bank into analysing the data as they have more experience with the dataset.

5.4.2 Industrial Design participants

The participants with the Industrial Design background, needed more time than the CS participants to discover the UI and also needed more time to see what they wanted to do with the dataset. The concept of protected attributes required a bit more explanation as it was not completely clear to the participants what they should do with that.

Struggling with this they both selected (a different) protected attribute in which they both select the wrong privileged group. Participant ID1 selected *sex* as their protected attribute and ID2 selected *age*. Picking the incorrect privileged group will result in metrics being outside their intended range, but still could help to see if certain groups are being treated unfairly.

After picking the first protected attribute, ID2 decides to first focus on seeing the dataset distributions in order to try and get a better representation of the protected attribute which doesn't satisfy their need to further explore that attribute. This is something which is not available in the current system. A similar idea was proposed by ID1 at a different time during the experiment where they opted for a distribution plot combining two protected attributes to learn more about the shared subgroups in those attributes.

Participant ID1 focused more on analysing the fairness metrics on the first protected attribute they selected. Both the fairness metric and the machine learning metrics need to be explained more to the participant in order to understand what's going on. This participant decides to pick another protected attribute (*foreign_worker*) and thinks "the dataset is problematic as there are a lot of foreign workers in the dataset". This is also shown in the bias metrics which motivates the participant to apply a filter.

It is interesting to see that both participants also said that the advice from experts could be very helpful in the analysis of the data. On top of that, participant ID2 said that in this tool "it might be useful to at least give some ethical guidelines on what is ethical to do and what is not". This is something which is hard to determine for datasets in general concerning that the attributes can be different, but it would be possible to provide some on those attributes which are seen more often like 'sex' and 'age'. This can be either provided by the system in some way or this is where the experts could step in and help the user with this issue.

Chapter 6

Discussion

In this chapter, we discuss the results of the experiment

6.1 Discussion

There have been significant differences between the two groups of participants when it comes to start using the system. Where the CS participants were able to analyse the dataset on their own, the ID participants needed some time to figure out what they were supposed to do which was expected as they have less experience working with datasets in this setting. Besides that, the concept of protected attributes and the metrics related to fairness were relatively difficult new concepts for the ID participants and it was clear that these aspects need a better explanation.

From the interviews it became evident that the distributions made it clear to the participants what they wanted to filter from certain attributes. On the other hand, both groups also expressed a need for more visualizations/distributions to better understand the relation between multiple attributes as the dataset distribution plots only focused on the relation between the primary attribute and a protected attribute. Compared to the tools we have seen in section 2.3, it is definitely possible to have some more visualizations/distributions to better visualize the treatments of different subgroups per attribute like the FairVis and DiscriLens tools [2, 21].

Furthermore, we saw that the CS participants made use of the different metrics available in the metric view. These metrics were not clear for all participants and as one of the participants said in section 5.4 that it would be helpful to improve the information about these metrics by adding formulas or examples. Another option would be to have something similar to the AIF360 toolkit where you make multiple metrics on bias and say how much of those could be seen as unfair rather than doing this per bias metric [4].

It was also surprising to see that the toolbar for selecting different dataset versions based on the modifications was not used that extensively by all of the participants. Only for participant CS2 it was used to look for positive trends in the metrics and actually made a comparison between different versions. This was different with the ID participants as they needed most of their time on just discovering unfairness in the dataset and did not do an extensive comparison between dataset modifications. Another reason is that the benefits of using this toolbar were probably not clear enough to the participants which is different than in CHAMELEON [12] where this toolbar is the main focus of the experiment.

The questions about the design brief were all answered similarly, as both type of participants thought that the bank should do more specific analyses on certain attributes before they can automate their service. In order to do this they also think that employees of the bank which have more experience with the data could improve this analysis, as they probably know more about certain cases of individuals

or groups which appear more frequently. As proposed by CS2 it also might be nice to have some ethical guidelines in this tool when working with these type of datasets. This is similar to [8] where they opt for the collaboration of ethnographer, ML experts and designer

The results of the experiment show that our system is able to inform participants about bias present in the dataset that is used for the experiments, but that there is still a lack of informational features which could improve the analysis on the dataset. Besides that, there is also room for improving the representation of bias as it seems that the understanding of bias is not optimal.

Chapter 7

Conclusions

In this chapter, we will provide conclusions on the three sub-research questions and the main research question. This is done by providing answers per question and with that we will answer the main research question.

RSQ1: Why should the practitioners of data-centric research be informed about potential bias and how can bias be found in datasets obtained by data-centric research?

In order to answer the first sub research question, we conducted a literature study. With the help of this study we found out that there is a decent amount of researchers that would like to do data-centric research, but do not have the knowledge and tools to do this by themselves [20, 8]. These practitioners of data-centric research would like to do analysis similar to those of data scientists that use a lot of Machine Learning techniques.

However, significant amount of research has pointed out that using datasets based on personal data could contain or could introduce bias in systems/applications which use that data for ML applications. Several different practices pointed out that combining ML techniques with data that contains personal data could do a lot of harm to certain groups of people as what we have seen with the COMPAS situation [3].

With the help of this we acknowledged that there was a need for some tool which is able to make people aware of the impact bias could have, especially for those which aren't data-scientists. So we went through all the tools related to finding bias in machine learning models such that we were able to make a better specification of what kind of system we needed. Subsequently, we saw that the AIF360 tool was the best tool suitable for a system where we will let the practitioners of data-centric research explore a dataset with the help of this tool.

RSQ2: What is required to create an application that helps exploring datasets that is able to find potential cases of bias located in datasets?

To answer the second sub research question, we defined what was needed to create such a system that help the users to explore a dataset. For that we designed FairData which lets users discover the bias that a particular dataset will introduce when used for training a machine learning model. Over time this system its UI changed to give the user the full experience of exploring a dataset and letting them see the change in bias by making simple modifications.

RSQ3: How does the application inform the user about potential bias located in the dataset and does it support the practitioners of data-centric research to reduce

this bias?

The third sub research question is answered by a experiment with two different groups of participants of possible data-practitioners to see whether FairData makes them aware of bias located in datasets. The results of the experiment show that our system is able to inform participants about bias present in the dataset that is used for the experiments, but that there is still a lack of informational features which could improve the analysis on the dataset.

Furthermore, it became clear that in order to support those practitioners in reducing the bias there should be a better representation of bias and a better visualization of subgroups in attributes as seen in toolkits made for finding bias. This could help those users to get a better idea of what modifications they have to make in order to reduce bias. Other ways to support these practioners would be to inform them better about bias in general and how it could be addressed/avoided.

How can practitioners of data-centric research (data-scientist or not) be supported to become aware of bias with the help of an application that 'explores' datasets?

Eventually, we can conclude that practitioners of data-centric research can be made aware of bias with our application FairData, by having this iterative process where the user can see the results of their actions in terms of bias metrics. The different dataset actions in combination with the information on the attributes and the information on what data is needed to calculate the bias, support the users to become aware of bias. On top of that, we have also seen that the system is also able to support the users in reducing the bias by trying out different dataset modifications.

7.1 Future work

To improve the experimentation part of this research, there are a couple of things we could have done to obtain better insights on how the practitioners of data-centric research get informed about bias in a dataset. The first thing to do would be to try and get more participants to get even more insights on how our tool informs the users. Besides that, it would be also nice to have a more diverse group of participants in terms of background like policymakers as this is a group of people which could be involved when creating/accepting tools which include machine learning algorithms.

Before having more participants it would first be good to improve the visualizations/representations of both the attributes in the dataset and the metrics on bias. It could both help to support the practitioners of data-centric research to get better insights into the treatment of different subgroups in the dataset. This could be added through implementing distributions that show outcomes of different subgroups related to the outcome of machine learning models.

Another thing to do would be to add some kind of Machine Learning functionality to the system where the participant would instruct and train their own models. With this it would be possible to see both bias present in the dataset itself and the bias introduced by machine learning algorithm which are the crucial issues that practitioners of data-centric research should become aware of. It will be an extension of the iterative process where it would also be possible to see the different outcomes between trained models and to see the two types of bias visualizations.

Bibliography

[1] Yongsu Ahn and Yu-Ru Lin. "FairSight: Visual Analytics for Fairness in Decision Making". In: *IEEE Transactions on Visualization and Computer Graphics* (2019), 1–1. ISSN: 2160-9306. DOI: 10.1109/tvcg.2019.2934262. URL: http://dx.doi.org/10.1109/TVCG.2019.2934262.

[2] Ángel Alexander Cabrera et al. *FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning*. 2019. arXiv: 1904.05419 [cs.LG].

[3] Julia Angwin et al. "Machine bias. ProPublica". In: *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* (2016).

[4] Rachel K. E. Bellamy et al. "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias". In: *IBM Journal of Research and Development* (2019), pp. 1–1. ISSN: 0018-8646. DOI: 10.1147/jrd.2019.2942287. arXiv: 1810.01943.

[5] Sarah Bird et al. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Tech. rep. Technical Report MSR-TR-2020-32, Microsoft, May 2020. URL https://www . . ., 2020.

[6] Sander Bogers et al. "Connected Baby Bottle". In: *Proceedings of the 2016 ACM Conference on Designing Interactive Systems - DIS '16*. DIS '16. New York, NY, USA: ACM, 2016, pp. 301–311. ISBN: 9781450340311. DOI: 10.1145/2901790.2901855. URL: http://dl.acm.org/citation.cfm?doid=2901790.2901855.

[7] Jacky Bourgeois, Gerd Kortuem, and Fahim Kawsar. "Trusted and GDPR-compliant research with the internet of things". In: *ACM International Conference Proceeding Series* (2018). DOI: 10.1145/3277593.3277604.

[8] N. Cila et al. "Listening To an Everyday Kettle : How Can the Data Objects Collect Be Useful for Design Research?" In: *Proceedings of the 4th Participatory Innovation Conference* January (2015), pp. 500–506.

[9] Antonin Danalet et al. "Location choice with longitudinal WiFi data". In: *Journal of Choice Modelling* 18 (2016), pp. 1–17. ISSN: 1755-5345. DOI: 10.1016/J.JOCM.2016.04.003. URL: https://www.sciencedirect.com/science/article/pii/S1755534515301020.

[10] Elisa Giaccardi, Chris Speed, and N. Rubens. "Things Making Things: An Ethnography of the Impossible". In: *Proceedings of the 1st International Research Network for Design Anthropology (online proceedings)* (2014), pp. 10–11.

[11] Hans Hofmann. *Statlog (German Credit Data)*. UCI Machine Learning Repository. 1994.

[12] Fred Hohman et al. "Understanding and Visualizing Data Iteration in Machine Learning". In: *Conference on Human Factors in Computing Systems - Proceedings* (2020). DOI: 10.1145/3313831.3376177.

[13] Imre Horváth. "Differences between 'research in design context' and 'design inclusive research' in the domain of industrial design engineering". In: *Journal of Design Research* 7.1 (2008), pp. 61–83. ISSN: 15691551. DOI: 10.1504/JDR.2008.018777.

[14] Fariba Karimi et al. "Homophily influences ranking of minorities in social networks". In: *Scientific reports* 8.1 (2018), pp. 1–12.

[15] Eirini Ntoutsi et al. "Bias in data-driven artificial intelligence systems—An introductory survey". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* September 2019 (2020), pp. 1–14. ISSN: 19424795. DOI: 10.1002/widm.1356.

[16] Pedro Saleiro et al. "Aequitas: A Bias and Fairness Audit Toolkit". In: *arXiv preprint arXiv:1811.05577* (2018).

[17] Shreya Shankar et al. "No classification without representation: Assessing geodiversity issues in open data sets for the developing world". In: *arXiv preprint arXiv:1711.08536* (2017).

[18] Martin W. Traunmueller et al. "Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities". In: *Computers, Environment and Urban Systems* 72 (2018), pp. 4–12. ISSN: 0198-9715. DOI: 10.1016/J.COMPENVURBSYS.2018.07.006. URL: https://www.sciencedirect.com/science/article/pii/S0198971517305914.

[19] Janne Van Kollenburg et al. "How design-inclusive UXR influenced the integration of project activities: Three design cases from industry". In: *Conference on Human Factors in Computing Systems - Proceedings* 2017-May (2017), pp. 1408–1418. DOI: 10.1145/3025453.3025541.

[20] Arnold P.O.S. Vermeeren, Virpi Roto, and Kaisa Vï¿ï¿nï¿nen. "Design-inclusive UX research: Design as a part of doing user experience research". In: *Behaviour and Information Technology* 35.1 (2016), pp. 21–37. ISSN: 13623001. DOI: 10.1080/0144929X.2015.1081292.

[21] Qianwen Wang et al. *Visual Analysis of Discrimination in Machine Learning*. 2020. arXiv: 2007.15182 [cs.HC].

[22] James Wexler et al. "The What-If Tool: Interactive Probing of Machine Learning Models". In: *IEEE Transactions on Visualization and Computer Graphics* (2019), 1–1. ISSN: 2160-9306. DOI: 10.1109/tvcg.2019.2934619. URL: http://dx.doi.org/10.1109/TVCG.2019.2934619.

[23] Ilia Zaitsev. *The best format to save Pandas data*. 2019. URL: https://towardsdatascience.com/the-best-format-to-save-pandas-data-414dca023e0d.