Surfacing Differences in Practices When Building Fair Machine Learning Systems with Fairness Toolkits: an Empirical Study

Eva Noritsyna
Supervisor(s): Jie Yang, Ujwal Gadiraju, Agathe Balayn
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

The ability to identify and mitigate various risks and harms of using Machine Learning models in industry is an essential task. Specifically because these may produce harmful outcomes for stakeholders, including unfair or discriminatory results. Due to this there has been substantial research into the concepts of fairness and its metrics, bias and its mitigation, and algorithmic harms and their sources. Various toolkits have been created to guide practitioners to reflect on these topics and provide suggestions on algorithmic solutions to mitigate these risks. However, it is not yet known how widely these toolkits are used and how they are perceived in terms of usefulness. In this project, practitioners were interviewed in order to determine to what extend do envisioned practices of practitioners without experience with fairness toolkits differ from those with the experience. The two toolkits considered were the IBM AI Fairness360 and Microsoft FairLearn. The data collected from the interviews suggests that there could be fewer differences in practices of practitioners with experience and without experience with toolkits, than those with training or work roles in ethics and fairness in ML and those without. This suggests that experience the toolkit itself is not indicative of a more thorough approach to identifying and mitigating harms in fair Machine Learning.

1 Introduction

With the rising need of Responsible AI systems in various contexts, the ability to identify and mitigate various risks and harms of using Machine Learning models in industry is an essential skill. Specifically because these may produce harmful outcomes for stakeholders, including unfair or discriminatory results.[5] These could happen across all domains and without malicious intent, for example in the field of finances, machine learning models could discriminate against underprivileged groups in giving out loans or in the field of human resources, applicant tracking systems may benefit socially-privileged groups.

Due to this there has been substantial research into the concepts of fairness and its metrics [19], bias and its mitigation, [15] and algorithmic harms and their sources. Some of these harms include machine learning models propagating and strengthening "structural advantages and disadvantages"[3], and opening up the possibility of "homogenity of decision making" [3]. Both of these concepts could reinforce the unfair treatment of minority groups. These harms could be caused by a variety of issues ranging from data population to reproduction of historical data patterns.

Due to these harms and the need for their mitigation, toolkits have been created to guide practitioners to reflect on these topics and situations and provide suggestions on algorithmic solutions to mitigate these risks [9] These toolkits were designed for use by practitioners working with ML systems in order to assess the output's fairness and aid with a degree of mitigation. However, it is not yet known how widely used and useful these toolkits are percieved as. [17]. The two toolkits that this research project will involve are the IBM AIF360 [1] and FairLearn [2]. From analysis and close studies of the toolkits and metrics further issues were identified including the metrics not reflecting the gravity of the negative impact the unfair outcomes may have in real life contexts and the lack of clarity and direction on the use of toolkits; practitioners are not informed at which stage of data processing the toolkits are most beneficial to employ.

[1] https://aif360.readthedocs.io/en/stable/

[2] https://fairlearn.org/v0.7.0/user$_guide/index.html$

These issues create gaps in research surrounding the practical use and adoption of such toolkits in industry and whether they benefit the workflow of practitioners when building machine learning models. It raises the question whether experience and knowledge of a fairness-centered toolkit, increases the ability to identify and mitigate harms in building ML systems. Therefore, our research question is,

> To what extent do (envisioned) practices of practitioners without experience with fairness toolkits differ from those with the experience?

To answer this question, we compared the practices, in terms of steps taken, by practitioners who are experienced with fairness toolkits and those who are not. We contrasted which out of the steps that were taken were in common or different between the two groups, and compared which sources of harms were identified and mitigated in the process. We contrasted the reasoning behind the uses of specific functionalities of the toolkit in order to identify how the use of toolkits changes the practices in building ML models.

Due to the goals of the toolkits being educating about fairness in Machine Learning and facilitating the process of identifying and mitigating bias in the process of building ML systems [4] [6], we predicted that the experience with a fairness toolkit would change the practices of the practitioner in a way to include a larger focus on fairness and bias and therefore identify more sources of harm that can be present in the dataset and the model. Therefore, we expected significant differences in practices of practitioners with and without experience with toolkits, in terms of the stps they take and the reasoning behind them.

We first identify the sources of harms in building machine learning models and interview practitioners with varying degrees of experience, both in academia and industry, about their approach to building ML models while analysing their recognition, understanding and mitigation techniques of the researched sources of harms. Through this we investigate the typical practices of practitioners who work with creating and maintaining ML models with regards to making the systems fair and ,for those with toolkit experience, where the toolkits fit in their practices. Moreover, we analyse the reasoning behind the steps in the practitioners' practices. Consequently, it is possible to identify whether the toolkits indeed improve the awareness of fairness and bias in ML, and provide practitioners with functionalities that they may use in their typical practices.

We begin with defining and analysing algorithmic harms in machine learning and fairness toolkits as a viable solution to them in Section 2. In Section 3 we explore the experimental set up that allows to collect data from practitioners in industry and in academia. These results are presented in three categories in Section 4 and their implications are further discussed in Section 5. In section 6, the responsible research component of this paper is considered.

2 Background work

2.1 Algorithmic Harms in Machine Learning

Machine learning models are highly dependent on the data they are trained on, and with bias present in that data, this bias could potentially be propagated through the model and be reflected in the output of the model [15]. Thus, it is crucial to identify such sources of bias in the data. However, the data is not the only source of harm in ML models, as the algorithmic design choices can add additional bias to the model. An example of such could be a model that sorts certain results to be displayed first, getting the most attention and influencing the user the most [7]. Lastly, the way the models are evaluated could also be

a source of algorithmic harms in ML models, for example by not considering the broader environment in which the output of the system will be utilised [2].

From these broader categories of sources of algorithmic harms in ML models, we have identified the following categorized list of sources of harm that can be present while building an ML model:

- Input dataset and its transformations
  - Data attributes
    - Irrelevant attributes
    - Incomplete set of relevant attributes
    - Oversimplified attributes
    - sensitive attributes
    - Proxies
    - Attributes transformations (feature engineering and protected attributes definition or removal)
  - Data population
    - Incorrect labelling
    - Over representation and under representation
    - Population transformations such as oversampling and undersampling
    - Concept drift
    - Covariate shift
  - Data errors
    - Missing data
    - Outliers
    - Duplicates
- Building of models
  - Algorithmic design choices
  - Model transformations
  - Environmental impact of model training
  - Invisible worker problem
- Evaluation of models
  - Incomplete or irrelevant choices of protected attributes
  - Incomplete or irrelevant choices of (fairness) metrics
  - Over-reliance on metrics in model evaluation
  - Task-related sources of harm
    - Undesired task
    - Task reproduces historical data patterns

For each of the categories, there are multiple approaches and mitigation techniques for dealing with the source of harms and the practitioner's chosen approach may depend on a variety of factors such as the domain. In this paper we will look further into whether this approach changes depending on the practitioners' experience and use of fairness toolkits. However, only a limited number of these sources of harms could be mitigated with the use of a toolkit.

2.2 Fairness Toolkits

Bias metrics and mitigation techniques are widely studied in academia and serve as the basis for fairness toolkits such as IBM AI Fairness 360 and Microsoft Fairlearn. However, recent research has delved deeper into assessing the usefulness and efficiency of such tools for practitioners in industry. Despite the research on fair ML and even the practitioners motivation to produce fair systems, some obstacles still exist in transferring this need to industry. Primarily, the focus and sole reliance on algorithmic solutions and metrics has been perceived as harmful for industry professionals and the focus has been suggested to shift on the quality of data and the datasets that are being used, being aware of the historical bias that may be present in it amongst others. [11]. Additionally, large gaps in understanding of these algorithmic solutions and metrics were identified between those working in industry and experts in fairness developing these [16]. The abundance of metrics relate to the numerous, conflicting fairness metrics, some requiring expertise and thorough understanding of situations suitable for their use, which may increase the implicit bias [19]. It has also been argued that the use of only metrics and mitigation algorithms may only turn the attention away from other sources of harm present in that data or model specifically [20]. This has been refered to as "ethics washing,"[13] defined by Madaio et. al as the notion of "reliance on only technical solutions" to achieve fairness.

Moreover, a more personal outlook and domain knowledge may be necessary for practitioners depending on the domain, particularly since the definition of fairness and its metrics may also be domain specific [10]. The use of toolkits such as IBM AIF 360 and Fairlearn has been criticised for not supporting such tasks. These toolkits pose a difficulty in adopting to already existing Machine Learning pipelines set up in industry and does not allow for comfortable customization [12]. This issues goes further in some cases, leaving practitioners unsure of whether fairness is even applicable in their domains [17].

Additional toolkit specific criticism and obstacle on the way to wider adoption amongst industry professionals has been the big knowledge gap that these toolkits imply and the generally low recorded *System Usability Survey Scores* that the toolkits received [12]. These were on average less than 70, which serves as a benchmark for systems that are "at least passable" [12] in terms of usability.

In terms of usefulness in the ML pipeline, practitioners have reported the toolkits to be hyperfocused on "model building and evaluation process" with limited to none support in the other phases, including examples such as checking the dataset for appropriate representation and proxies [12]. This increases the danger of practitioners solely relying on the metrics from the toolkits.

These findings suggest that although the motivation to produce fairer ML systems exists in industry, there are multiple obstacles on the way of adopting fairness toolkits in industry settings, causing a higher reliance on bias mitigation by the practitioner without reliance on such tools.

2.3 IBM AI Fairness 360 and Microsoft Fairlearn

The two toolkits examined in this paper offer a similar set of base functionalities available across a wide range of similar toolkits. For both toolkits, their focus is not only on providing Python (and R in the case of AIF 360) packages that practitioners in industry would be able to add to their pipelines, but also on the education in bias checking and guidance in selecting the correct approach [4][6]. These toolkits also have available code on GitHub and have the highest statistics on the site, as shown in Table 1.

Toolkit Name	Forks	Stars
IBM AIF 360	572	1.7k
Fairlearn	311	1.3k
Tensorflow Fairness Indicators	71	259
Pymetrics Audit-AI	43	284

Table 1: Fairness toolkits with open code bases and their GitHub statistics

The main difference between the two toolkits is the number of available algorithms. An overview of available metrics and bias mitigation techniques can be found in Figure 3. FairLearn only has four main algorithms (*Exponentiated Gradient, Grid Search, Threshhold Optimizer and Correlation Remover*. AIF 360 has a total of 14 algorithms covering Pre-, In- and Post-Processing. However, Fairlearn's post-processing algorithms have the ability to intake a trained model and transform it in order to fit the chosen fairness metrics [14]. Both toolkits have extnesive documentation and tutorials provided on their respective websites and have semi-active communities.

		Models covered				Group Fairness						Individual Fairness			
Tool	Open Source?	Regression	Classification (binaryoutcome)	Multi-class outcome	Handles multi-class protected feature?	Demographic parity (statistical parity)	Equal opportunity / True positive parity / false positive error rate balance	Equal odds (True positive and false positive parity)	Disparate impact	Discovery rate	Omission rate	Counterfactual fairness	Sample distortion metrics	Other fairness metrics	Bias mitigation
IBM Fairness 360	☑	☐	☑	☑	☑	☑	☑	☑	☑	☑	☑	☐	☑	Generalized Entropy Index Differential Fairness and Bias Amplification (full list here: https://aif360.readthedocs.io/en/latest/modules/generated/aif360.metrics.ClassificationMetric.html)	Optimized Preprocessing, Disparate Impact Remover, Equalized Odds Postprocessing, Reweighing, Reject Option Classification, Prejudice Remover Regularizer, Calibrated Equalized Odds Postprocessing, Learning Fair Representations, Adversarial Debiasing, Meta-Algorithm for Fair Classification, Rich Subgroup Fairness
Fairlearn	☑	☑	☑	☐	☑	☑	☑	☑	☐	☐	☐	☐	☐	Group max / min / summary	Exponentiated Gradient, GridSearch, Threshold Optimizer

Figure 1: Open source toolkit feature summary table. Adapted from [12]

3 Method

This research project is a combination of a literature review and an empirical study. The goal of the interviews is to understand how machine learning practitioners build fair machine learning models, and especially how these practices might differ with and without a fairness toolkit (specifically IBM AIF360, Microsoft FairLearn, etc.). It will contribute to science by identifying where practitioners might need more support.

3.1 Literature review

In order to define and categorise the algorithmic harms listed in Section 2 and their sources that the participants of the interviews will be aiming to identify, a thorough literature study was conducted. Through various sources, a comprehensive for the task, list of algorithmic harms' sources was identified. These sources of harms were then compared and contrasted with the functionalities of the two toolkits that were studied in order to gain insight to what harms can be mitigated by the use of the toolkits, concluding that some of the harms stemming from the algorithmic design choices, model transformations, proxies, attribute transformations and data population can be addressed with the use of the mitigation algorithms provided by the toolkit.

3.2 Interviews

3.2.1 Participant recruitment

The target participants for the interviews were Machine Learning practitioners with varying degrees of experience. In total the interviews required at least 30 participants; 10 of which have had previous experience with Microsoft Fairlearn toolkit, 10 of which have had previous experience with the IBM Artificial Intelligence 360 toolkit and 10 who were unfamiliar with the toolkits but had to experience with the general notion of fairness in the design of Machine Learning systems. From each of the 3 groups, 4 participants were aimed to be Senior Data Scientists, 2 were aimed to be Medior Data Scientists and 4 were aimed to be Junior Data Scientists or MSc students. This would allow a comprehensive understanding of different approaches depending on the level of familiarity and knowledge of Machine Learning and fairness concepts. The final distribution of participants can be found in Figure 2, with further identification whether a participant works in academia as this is valuable for evaluation. The participants were recruited through personal network as well as online scouting. For the latter, the IBM AIF 360 official Slack channel [3] and Microsoft Fairlearn official Discord channel [4] and individuals on LinkedIn were messaged with the statement available in A.

The recruited participants gave their informed consent for the interviews to be recorded and transcribed for analysis purposes, after which to be discarded.

	None	**Microsoft Fairlearn**	**IBM AIF360**
Senior	3	3	3
Medior	3	3	3
Junior/MSc	4	4	4

Table 2: Number of participants per level of experience with ML and experience with toolkits

[3] https://aif360.slack.com/
[4] https://discord.gg/R22yCfgsRn

After the participants were recruited, two pilot studies were conducted; one with a student with no experience with fairness toolkits and one with a PhD candidate with experience with IBM AI Fairness 360. This was done in order to confirm that the prepared interview questions cover relevant topics, identify missing questions or code blocks and quality assure that the Google Collab notebooks with use-cases function as expected. The results from the pilot studies were not used in this study.

3.2.2 Interview protocol

The interviews were designed to be semi-structured, following the "Think Aloud" method, where participants were first asked some general background questions and then presented with a use case. After the initial introduction to the dataset, the interviewees were encouraged to walk through what their approach would typically be to training a machine learning model. The goal is to note their approach to fairness and identifying algorithmic harms and biases present in the data.

The interviewee will be prompted to expand on each identified source of risk, including how they understand it, how it is detected and what would their approach to the mitigation of this source be.

The interview included pre-defined questions prepared in two levels of abstractness; the first level helps guide the participant towards an area of sources of harms that they may have forgotten. Examples of these type of questions are "What would you say about the quality of the dataset in this case(s)? Are there any things you would improve/solve before building a model?". These are predefined per each category of sources of harms. If the participant still does not pick up on any further harms, the level two questions are asked which are more direct. These include questions such as "Do you think the data is representative to the real population in this case(s)? Is this something you would usually consider? And if yes, what would you think about? If not, why is it not important?". As the participant has a choice of whether or not to code, some pre-prepared code snippets were ready to be transferred to the participants if they wished to execute them in practice.

Participants with experience with a fairness toolkit were asked to explore one use-case, with the task of *predicting hospital re-admissions in diabetic patients.*. Participants with no prior experience with the toolkits, were first asked to explore the same use-case as those with experience. However, afterwards, they were faced with a notebook with a short tutorial and demo of the respective fairness toolkit. An additional use case notebook was prepared with the task of *predicting whether a person would have 'high' healthcare utilization*, to which they could apply the knew toolkit knowledge they have gained from the demo.

Depending on if the participant had or did not have experience with a toolkit, toolkit centred questions were asked, aiming to delve deeper into the impact the toolkits may have had on the perception of algorithmic harms of the participant.

3.3 Datasets

The use cases used for the interviews belong to the medical fields. The first one utilizes Medical Expenditure data [5] with the model classification task to predict whether a person would have 'high' healthcare utilization. This dataset was used to interview people who have previously not had experience with toolkits, as the part of the interview that they do without toolkits. The second one utilizes the Diabetes Hospital Readmission dataset [6], with

[5] https://raw.githubusercontent.com/pablobiedma/datasets/main/final_diabetes.csv
[6] https://raw.githubusercontent.com/pablobiedma/datasets/main/dataset.csv

the classification task being whether the patient will readmit within 30 days. This dataset was used for interviews with people who are familiar with toolkits as it is a dataset in the same domain and is not very popular amongst ML tutorials, hopefully being a dataset that participants are initially not familiar with.

Each dataset was analysed and edited in order to include all of the easily identifiable sources of harms identified in 2 and how they were added to the datasets.

3.4 Coding and analysis

In order to analyse the results of the interviews, an open coding technique described by D. Thomas [18] was employed. Through watching 30 recorded interviews and analyzing some of their corresponding the transcripts, the interviews were open coded by defining categories of discussed information that may in any way be relevant to the project. The higher-level categories that have been identified are (1) identifying harm source; (2) understanding harm source; (3) mitigating harm source; (4) identifying impacts of technique; (5) identifying alternate approaches; (6) business factors; (7) domain factors; and (8) task factors. The achieved findings are part of the participants' personal choice of approach rather than "best practices." These were aligned with the goals of further result evaluation. Afterwards, the categories were further analysed and grouped to establish the overarching motivations for practitioners to introduce toolkits in their approach.

3.5 Limitations

The interviews have been conducted over Microsoft Teams, in a pre-determined time frame, often not allowing for flexibility. Additionally, due to the nature of recruitment, some further limitations were identified. Multiple participants with experience with toolkits have had some direct employment from the developers of the toolkit. Furthermore, the majority of practitioners from academia and significant number of practitioners from industry, directly work with the concepts of fairness and bias in Machine Learning and therefore could be expected to have more knowledge on the topic than an average practitioner. Lastly, the practitioners with experience who were recruited were aware that they were identified due to their toolkit experience and therefore could often deduce that the focus of the study was on fairness and bias.

4 Results

The results of the conducted interviews explore the practices in terms of the steps that practitioners took and which sources of harms they affected. The results are presented in two main sections; those practices that the two groups had in common and those that differed between the two groups, specifically focusing on the additional functionalities presented by the toolkits.

4.1 Common Activities

4.1.1 Task Analysis

Practitioners without experience with toolkits as well as those with experience attempted a form of task analysis. The majority of the participants noted early on in the interviews the need to consult a domain professional to be able to draw educated decisions and conclusions

about the data and/or relevancy of some of the attributes. Later on, this reflected in multiple participants mentioning that they are unable to make decisions on whether some attributes are irrelevant for the task. Other domain factors included practitioners relying on their medical domain knowledge mentioning that they are *"aware that there exist diseases which certain genders and races are more predisposed to"*, therefore possibly straying away from some practices such as removing those sensitive features. More experienced practitioners, specifically in industry, also questioned the source of the data. With this question, a variety of implications was considered; correctness of the data, possibility to add more features, goal of the output of the task, and what stakeholders could this data concern. Practitioners in industry pointed out that *"some of these attributes are subjective, [and if given the possiblity], I'd collect some more objective metrics."*. It was also pointed out by practitoners that *"across industries [there may be] different practices."*

4.1.2 Data exploration

All of the interviewed practitioners took several steps to explore the data without any further prompts by the interviewer. The initial step was shared amongst all the participants; looking over the data manually to attempt to understand the features and the task at hand; *"understanding the data is the most important step."*. Multiple practitioners referred to this being the step in exploratory data analysis, that requires the longest time, with the goal of *"[looking] at each variable to understand what influence each variable is going to make."*.

The majority of practitioners, highlighted missing values and uneven distributions as sources of harm to look out for. Those who have not identified those explicitly, referred to normally considering those in their usual workflow and having forgotten during the interview, or have referred to them implicitly. However, out of the **80%** of practitioners without experience with toolkits that brought them up, a minority viewed them as sources of harm in terms of fairness, and mostly viewed them as *"noise reduction"* and ways to improve the performance of the model.

When exploring the data, the participants that had experience both in industry and in research pointed out the differences in approach that they would take depending on whether this data was purely for research or will be used in industry, *"there's two ways in my head, [the more academic] like when I was writing my thesis, which was more data crunching and [what I do at work] which is just dig in into data"*. Similarly, another practitioner with similar experience pointed out that *"if this [were] for research, maybe you'd go deeper."*

Within the practitioners without experience with toolkits, around **50%** explored the correlation of features within the dataset, however not specifically mentioning the harms that could source from highly coupled features, but more for the *"[reduction] of space complexity"*. The people with experience with the toolkits who are in the field of fairness in ML, identified that correlation tables and *heatmaps* could help identify underrepresented groups.

Sensitive attributes were paid greater attention two by the large majority of both groups, however sometimes implicitly, for example by stating that they would like to look at the distributions of *gender* and *age* attributes. The participants with more experience in the field of fairness, also analysed sensitive attributes beyond the typical ones, often implying the issue of possible *Proxies* but with only **20%** naming it explicitly.

4.2 Different Activities

4.2.1 Processing

Metrics: Those practitioners with experience with toolkits often identified which fairness metrics they would like to use, if any, in order to evaluate the fairness of the model. A widely chosen option was the *Disparate Impact Ratio*, for reasons including the it *"can be used before applying the learning algorithm and after."* This reasoning came up often, as several practitioners have identified the ability to *"trial and error"* different practices with simple models as very useful in building a ML model. Some of the practitioners with formal education or training in ethics and fairness in ML brought up the concept of *"group fairness versus individual fairness"*

Pre-processing: The widest use of the toolkits appeared to be in the pre-processing phase, where practitioners showed preference to the algorithms of *Reweighing*. A highlighted reason for that was the *"pre-processing algorithms [being] model agnostic"* and therefore providing significant flexibility in choosing ML algorithms later on in the pipeline. *Reweighing* in particular was highlighted due to its more thorough approach to balancing datasets than is offered by *under or over sampling*. An identified disadvantage of using pre-processing algorithms was pinpointed as being *"difficult to decode for [...] explainability"*

In-processing: From the participants, very few have discussed their use of in-processing algorithms. However, those who have, have described them to be crucial in their pipelines, specifically the *Adversarial Debiasing* algorithm, despite not being model agnostic.

Post-processing: Participants did not highlight the use of post-processing algorithms as part of their usual workflow, with post-processing algorithms only being mentioned in 3 interviews, disclaiming that *"I know of it, but I do not have experience using it."*, and alike. A practitioner noted that an abundance of algorithms in a toolkit is more useful in the context of research rather than industry, *"It suggests you too much for what you need for production. [...] Maybe for research you can go deeper."*. This again highlights the issue of

4.2.2 Business factors

Out of the practitioners with experience with toolkits who work in industry, and are not part of the toolkit development team, several have pointed out the addition of the toolkit to their practices in industry, specifically for its business factors. The main one that has been stated as the *"explainability to stakeholders."* Senior Data Scientists have stated that the ability to rely on a third party to confirm that their model is fair and showcase to stakeholders in a transparent way how fairness was considered in the particular scenario is important to their respective industries. The *"easy compatibility with Scikit-learn"* allows for

However, practitioners working in industry, but not specifically in the area of fairness have also pointed out their reasons for not using the toolkit, such as *"[it is] difficult to add to an existing pipeline", "it is not mandated to look at this", "we already had good insights into the data"*. These practitioners' practices were not influenced by their experience with the toolkits. Another obstacle highlight by a Senior Data Scientist has been that the inclusion of these toolkits and metrics into already existing pipelines in industry, have to be motivated by someone and there have been instances where no one in the team wanted to take the additional responsibility, considering it was not enforced by the higher level employees. A Senior Data Engineer has stated their precautions with the use of such toolkits, *"Fairness for some companies is just a small checkbox [and when using the toolkit] they put the check without any questions."*

5 Discussion

5.1 Impact of toolkit experience on the practices

The present study was designed to determine the effect of experience with fairness toolkits on the practices of practitioners when building Machine Learning models. Contrary to expectations stated in Section 1, this study did not find a significant difference between those two groups of practitioners. The results of the interviews show that the main difference between the practices of practitioners with experience with fairness toolkits and those without, was in the consideration of fairness metrics provided by the toolkits. However, this could be attributed to the fact that the interviewed practitioners were aware of the interview involving the toolkits, and therefore being motivated to include the fairness metrics which appear to be the most used functionalities of the toolkits. Moreover, despite this being the biggest difference between the results of the two groups, not all of the interviewed practitioners brought up fairness metrics without further prompts of the interviewer, suggesting that they are not typically used in their workflows.

Surprisingly, the main differences in approach and in the types of harms that were picked up were related more to the level of experience with concepts of fairness in Machine Learning, rather than direct experience with the toolkit. Participants with formal education in ethics and fairness in Machine Learning were aware of more sources of harms without further prompts than those who came from different educational backgrounds and work in fields not directly related to fairness. It can also be highlighted that experience with Machine Learning on its own, without the formal education or training in ethics and fairness, does not yield more awareness of fairness. This can be seen through the Senior Data Scientists also failing to pick up on many of the sources of harms in the data, focusing more on the sources that may also influence the accuracy of the model, such as missing values and outliers. The sole introduction of metrics to their practices, can be seen as "ethics washing" which can lead to incorrect techniques of mitigation [8].

Practitioners with experience with toolkits did show a more thorough analysis of sources of harms explicitly in the presented dataset, however, using mostly methods that did not require a toolkit. This could, again, be attributed to the way they have been recruited or to fact that a large part of them come from a job that is in some degree related to fairness and ethics. Similar results came from practitioners who has a formal education with Machine Learning, especially that involved conversations of fairness and responsibility in ML.

5.2 Reasons for differences in practices for practitioners with toolkit experience

During the interviews the participants who work in industry also responded on why they choose to include certain functionalities of the toolkits in their practices and why not. Three overarching factors were identified in the practitioners' willingness to use toolkits: usefulness of the toolkits and its effect on the practitioners' practices, usability of the toolkit and its effect on the practitioners' practices, and the use of toolkits in industry.

Across most interviews it was found that practitioners with experience with toolkits who used some functionalities of the toolkit, did so due to their usefulness to the task, such as having a reliable fairness metric that they can optimize for.

Otherwise, practitioners use the toolkit to introduce a metric or algorithm to their codebase efficiently, implying they would have introduced this functionality regardless of whether the toolkit was used. Lastly, some practitioners use the toolkits solely for business oriented

goals, such as explainability to other stakeholders and reliance on a third party to show their impartiality.

A strong identified reason for practitioners to use the toolkits is the wide variety of available functionalities, pre-, in- and post-processing techniques, that may be useful on a case-by-case basis.

One of the most picked up on reasons to introduce toolkits into their approach and thus changing their practices was the usability of the toolkits. With an open-source library that is easy to import into the project, it becomes much simpler to introduce some functionality into the code through that over coding it from scratch. This sentiment was shared by the majority of the interviewees who use the toolkit in industry or intend to. In the cases where practitioners with experience with toolkits and without want to arrive to the same functionality such as *Reweighting*, those with experience with toolkits can do it in a fewer number of steps and more efficiently.

The majority of practitioners who work in industry and have experience with toolkits have identified an issue in introducing them to their typical work practices and approach. These obstacles include a steep learning curve for to familiarise themselves and the team with the functionality and required syntax of the toolkits. Other obstacles are the rigidness of the datatypes that IBM AIF 360 operates with, often cannot be extended to the non-tabular real world data that is being dealt with in industry. obstacles to adoption in industry - learning curve, rigid datatypes. Moreover, some interviewees identified a difficulty in introducing the toolkit into the pre-existing pipeline in their respective companies, however, other interviewwes praised the ease of this integration as on of the biggest benefits.

Another widely identified benefit of changing typical practices to include the toolkit is the explainability it provides for stakeholders including the business-side employees of the companies and the clients in some cases.

6 Responsible research

In order to assure that the research was conducted responsibly, the five principles from the Netherlands Code of Conduct for Research Integrity (2018) [1] were considered.

Honesty: The results communicated were made sure to be fully derived from the interviews directly, or supplemented with background knowledge from the literature review to assure that the results are the basis of all the founded claims and the process is reported accurately.

Scrupulousness: The methods taken in this research were scientific or scholarly in terms of a semi-structured interview, often based on methods described in literature, such as think-aloud interviews, pilot studies and open coding.

Transparency: The method of conducting the interviews as well as analysing and collecting the data were explained thoroughly in the study in order to make the process possible to replicate as much as possible. However, the **reproducibility** of the exact results this study has come to, may be difficult as there are several confounding variables such as; whether the person works in the field of fairness, whether the person has had formal training in ethics and fairness of ML and whether they were recruited with the reason for their recruitment being their experience with a toolkit. Regardless of this, all the material used in the study, including the prompt for recruitment, semi-structured interview topics and several screenshots of the Google Collab use-cases are provided to increase reproducibility .

Independence: The research is guided solely by the interest in the topic, without consideration of the desired outcomes of any third parties for this research.

Responsibility: The participants of the study were directly approached with the goal of the study and asked for their consent of being filmed and transcribed, including by requesting verbal informed consent prior to starting the interviews. The recording were stored securely in a private cloud-hosted storage with access only for the research team. No identifiable information were ever mentioned past the recordings and the recordings are to be disposed off after the termination of analysis.

7 Conclusion and Future Work

This study set out to identify the differences in practices of practitioners with and without experience with fairness toolkits as a way to determine whether such toolkits raise the practitioners' awareness to fairness and educates the practitioner of the importance of considering fairness and bias when building machine learning systems. While, the study confirmed that there exists some difference in the practices, it is not possible to asses with certainty whether this difference comes from the experience or from certain confounding factors. Moreover, the study found that these differences also exist within the two groups of participants suggesting that generally the experience and formal education in ethics and fairness in Machine Learning also may play a big role in the steps taken during the approach in order to identify and mitigate sources of harms while building Machine Learning models. Lastly, the study found that some of the differences in practices between practitioners with and without experience with fairness toolkits, may be correlated with factors only relevant in industry and not in academia. The contribution of this study has been to analyse the effect of experience with fairness toolkits on practitioners' typical practices and their reasoning. The new understanding of its effect can help determine the most effective way to increase the consideration of fairness in practitioners' usual practices.

7.1 Future Work

A limitation of this study identified in Section 3.5, related to how the practitioners with toolkit knowledge were recruited, as they knew prior to the interviews that they were recruited primarily because of their experience with the toolkits. Due to this, the results that came from them may not be reflective of their usual practices as they might have been more focused on fairness. For the future work, the study could be repeated but recruiting participants in a way that does not state that they have been recruited for their knowledge with the toolkit. That could possibly yield more certain results, as the possibility of interference of recruitment cannot be ruled out. Additionally, the study could be repeated with practitioners in industry and in academia as that variable also could potentially affect the typical practices of practitioners. Lastly, the study could be repeated with people who do not work directly in the field of fairness in ML, as their insights differ strongly from other ML engineers and data scientists.

References

[1] Keimpe Algra et al. “Nederlandse gedragscode wetenschappelijke integriteit”. In: (2018).

[2] Agathe Balayn and Seda Gurses. *Beyond Debiasing: Regulating AI and its inequalities*.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. “Fairness and Machine Learning Limitations and Opportunities”. In: 2018.

[4] R. K. E. Bellamy et al. “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias”. In: *IBM Journal of Research and Development* 63.4/5 (2019), 4:1–4:15. DOI: 10.1147/JRD.2019.2942287.

[5] Reuben Binns. “Fairness in Machine Learning: Lessons from Political Philosophy”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 23–24 Feb 2018, pp. 149–159. URL: https://proceedings.mlr.press/v81/binns18a.html.

[6] Sarah Bird et al. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Tech. rep. MSR-TR-2020-32. Microsoft, May 2020. URL: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[7] Robert Epstein and Ronald E. Robertson. “The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections”. In: *Proceedings of the National Academy of Sciences* 112.33 (2015), E4512–E4521. DOI: 10.1073/pnas.1419828112. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1419828112. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1419828112.

[8] Sina Fazelpour and Zachary C. Lipton. “Algorithmic Fairness from a Non-Ideal Perspective”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’20. New York, NY, USA: Association for Computing Machinery, 2020, 57â63. ISBN: 9781450371100. DOI: 10.1145/3375627.3375828. URL: https://doi.org/10.1145/3375627.3375828.

[9] Sorelle A. Friedler et al. *A comparative study of fairness-enhancing interventions in machine learning*. 2018. DOI: 10.48550/ARXIV.1802.04422. URL: https://arxiv.org/abs/1802.04422.

[10] *The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning*. Stockholm, Sweden, 2018.

[11] Kenneth Holstein et al. “Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?” In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1â16. ISBN: 9781450359702. DOI: 10.1145/3290605.3300830. URL: https://doi.org/10.1145/3290605.3300830.

[12] Michelle Seng Ah Lee and Jat Singh. “The Landscape and Gaps in Open Source Fairness Toolkits”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445261. URL: https://doi.org/10.1145/3411764.3445261.

[13] Michael A. Madaio et al. “Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, 1â14. ISBN: 9781450367080. URL: https://doi.org/10.1145/3313831.3376445.

[14] Afra Mashhadi, Annuska Zolyomi, and Jay Quedado. “A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education”. In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. CHI EA ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391566. DOI: 10.1145/3491101.3503568. URL: https://doi.org/10.1145/3491101.3503568.

[15] Ninareh Mehrabi et al. “A Survey on Bias and Fairness in Machine Learning”. In: *CoRR* abs/1908.09635 (2019). arXiv: 1908.09635. URL: http://arxiv.org/abs/1908.09635.

[16] Brianna Richardson and Juan Gilbert. “A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions”. In: Dec. 2021.

[17] Brianna Richardson et al. “Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits”. In: May 2021, pp. 1–13. DOI: 10.1145/3411764.3445604.

[18] David R. Thomas. “A General Inductive Approach for Analyzing Qualitative Evaluation Data”. In: *American Journal of Evaluation* 27.2 (2006), pp. 237–246. DOI: 10.1177/1098214005283748. eprint: https://doi.org/10.1177/1098214005283748. URL: https://doi.org/10.1177/1098214005283748.

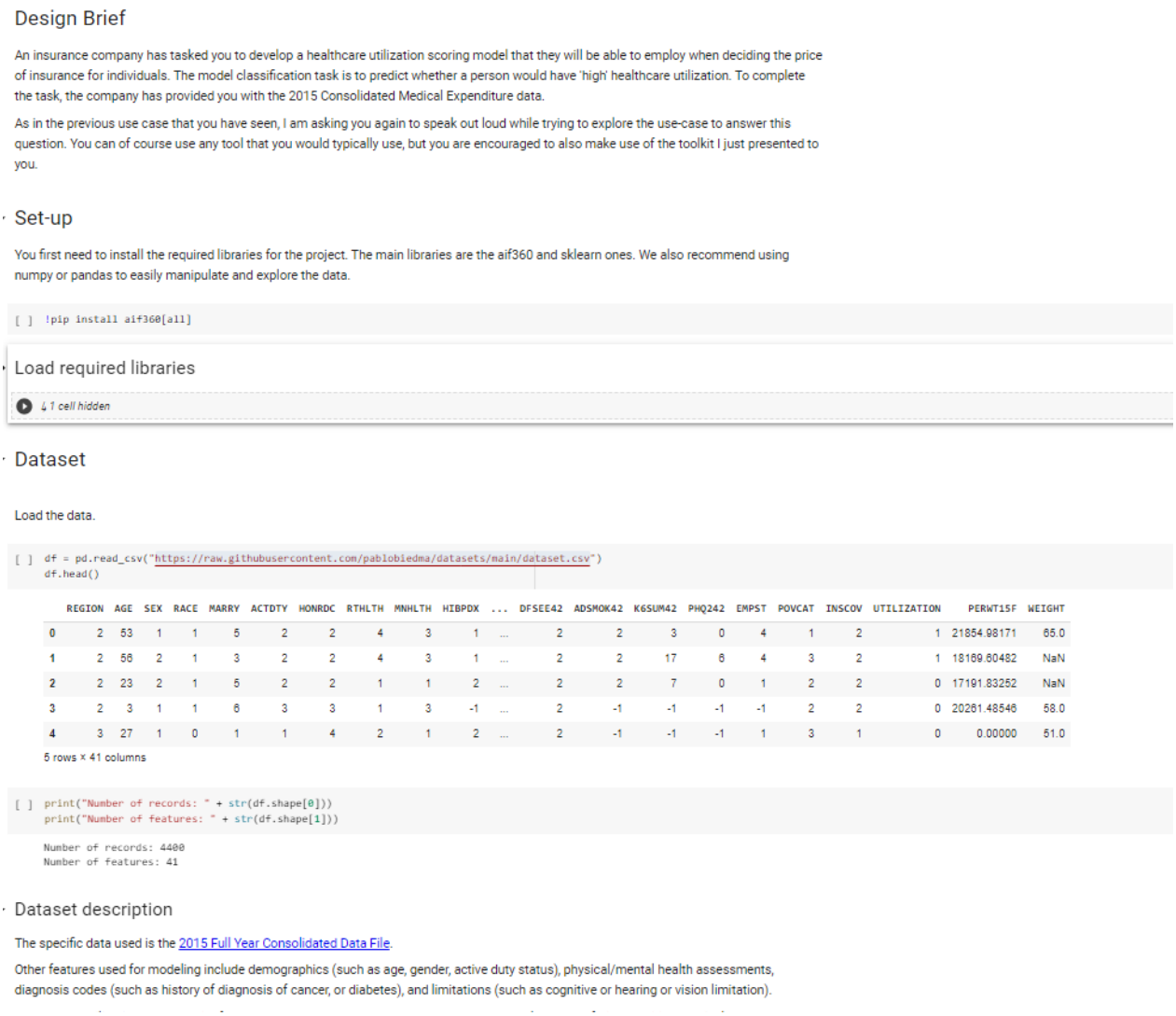
[19] Sahil Verma and Julia Rubin. “Fairness Definitions Explained”. In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. 2018, pp. 1–7. DOI: 10.23919/FAIRWARE.2018.8452913.

[20] Kiri Wagstaff. “Machine Learning that Matters”. In: *CoRR* abs/1206.4656 (2012). arXiv: 1206.4656. URL: http://arxiv.org/abs/1206.4656.

A Recruitment Statement

I am a Computer Science Engineering student at TU Delft in the Netherlands. I am currently writing my thesis on the subject of machine learning practices. I'm trying to understand how machine learning practitioners build fair machine learning models, and especially how these practices might differ with and without a fairness toolkit (e.g. IBM AIF360, Microsoft FairLearn, etc.). It will contribute to science by identifying where practitioners might need more support. In order to get insight into the industry, I am searching for data and machine learning practitioners who have experience with using IBM AIF360. I would love to be able to hold an interview with you to gain some more knowledge about this toolkit and your experiences working on it. The interview should last maximum of 1 hour. It will be online and will be recorded for the sole purpose of statistical analysis (no personal information will be used in the final report), but deleted at the end of the research. Thank you in advance. I am really looking forward to hopefully meeting you! Please let me know if you have any questions and if we can schedule a call!

B Google Collab screenshots



Design Brief

An insurance company has tasked you to develop a healthcare utilization scoring model that they will be able to employ when deciding the price of insurance for individuals. The model classification task is to predict whether a person would have 'high' healthcare utilization. To complete the task, the company has provided you with the 2015 Consolidated Medical Expenditure data.

As in the previous use case that you have seen, I am asking you again to speak out loud while trying to explore the use-case to answer this question. You can of course use any tool that you would typically use, but you are encouraged to also make use of the toolkit I just presented to you.

Set-up

You first need to install the required libraries for the project. The main libraries are the aif360 and sklearn ones. We also recommend using numpy or pandas to easily manipulate and explore the data.

```
[ ] !pip install aif360[all]
```

Load required libraries

4 1 cell hidden

Dataset

Load the data.

```
[ ] df = pd.read_csv("https://raw.githubusercontent.com/pablobiedma/datasets/main/dataset.csv")
    df.head()
```

	REGION	AGE	SEX	RACE	MARRY	ACTDTY	HONRDC	RTHLTH	MNHLTH	HIBPDX	...	DFSEE42	ADSMOK42	K6SUM42	PHQ242	EMPST	POVCAT	INSCOV	UTILIZATION	PERWT15F	WEIGHT
0	2	53	1	1	5	2	2	4	3	1	...	2	2	3	0	4	1	2	1	21854.98171	65.0
1	2	56	2	1	3	2	2	4	3	1	...	2	2	17	6	4	3	2	1	18169.60482	NaN
2	2	23	2	1	5	2	2	1	1	2	...	2	2	7	0	1	2	2	0	17191.83252	NaN
3	2	3	1	1	6	3	3	1	3	-1	...	2	-1	-1	-1	-1	2	2	0	20261.48546	58.0
4	3	27	1	0	1	1	4	2	1	2	...	2	-1	-1	-1	1	3	1	0	0.00000	51.0

5 rows × 41 columns

```
[ ] print("Number of records: " + str(df.shape[0]))
    print("Number of features: " + str(df.shape[1]))
```

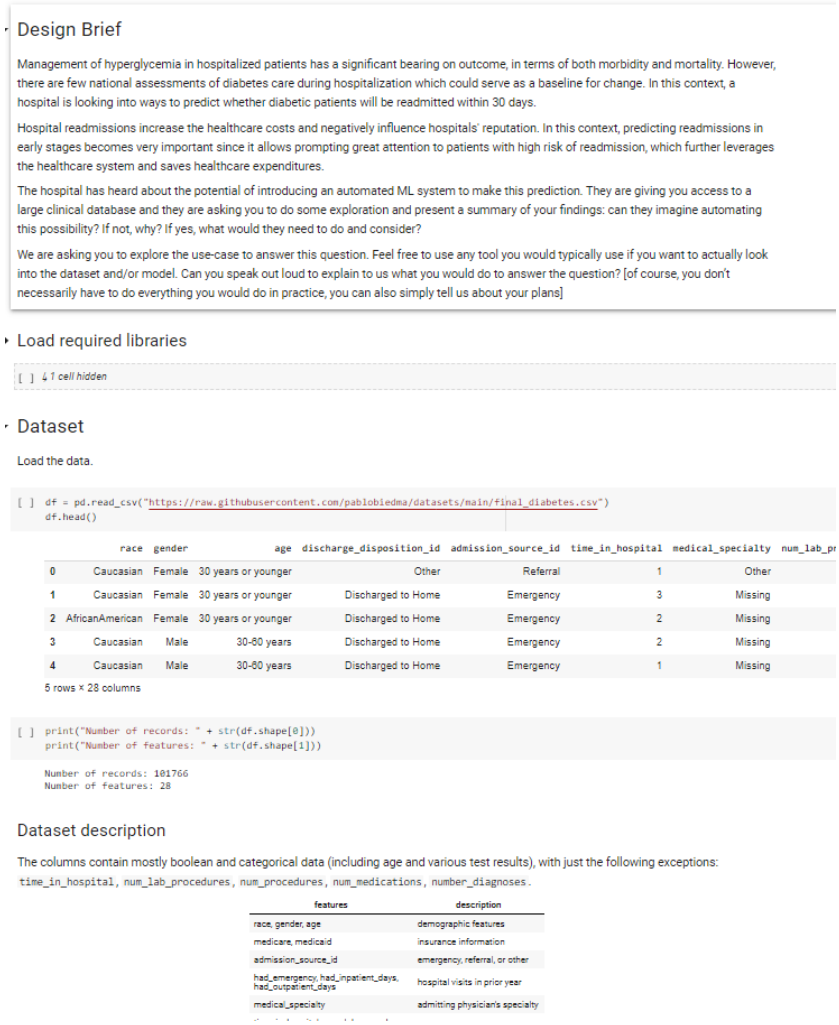
```
Number of records: 4400
Number of features: 41
```

Dataset description

The specific data used is the 2015 Full Year Consolidated Data File.

Other features used for modeling include demographics (such as age, gender, active duty status), physical/mental health assessments, diagnosis codes (such as history of diagnosis of cancer, or diabetes), and limitations (such as cognitive or hearing or vision limitation).

Figure 2: Screenshot of the Google Collab Notebook with Diabetes Usecase



Design Brief

Management of hyperglycemia in hospitalized patients has a significant bearing on outcome, in terms of both morbidity and mortality. However, there are few national assessments of diabetes care during hospitalization which could serve as a baseline for change. In this context, a hospital is looking into ways to predict whether diabetic patients will be readmitted within 30 days.

Hospital readmissions increase the healthcare costs and negatively influence hospitals' reputation. In this context, predicting readmissions in early stages becomes very important since it allows prompting great attention to patients with high risk of readmission, which further leverages the healthcare system and saves healthcare expenditures.

The hospital has heard about the potential of introducing an automated ML system to make this prediction. They are giving you access to a large clinical database and they are asking you to do some exploration and present a summary of your findings: can they imagine automating this possibility? If not, why? If yes, what would they need to do and consider?

We are asking you to explore the use-case to answer this question. Feel free to use any tool you would typically use if you want to actually look into the dataset and/or model. Can you speak out loud to explain to us what you would do to answer the question? [of course, you don't necessarily have to do everything you would do in practice, you can also simply tell us about your plans]

Load required libraries

[ ] *4 1 cell hidden*

Dataset

Load the data.

```
[ ] df = pd.read_csv("https://raw.githubusercontent.com/pablobiedma/datasets/main/final_diabetes.csv")
    df.head()
```

	race	gender	age	discharge_disposition_id	admission_source_id	time_in_hospital	medical_specialty	num_lab_pr
0	Caucasian	Female	30 years or younger	Other	Referral	1	Other	
1	Caucasian	Female	30 years or younger	Discharged to Home	Emergency	3	Missing	
2	AfricanAmerican	Female	30 years or younger	Discharged to Home	Emergency	2	Missing	
3	Caucasian	Male	30-60 years	Discharged to Home	Emergency	2	Missing	
4	Caucasian	Male	30-60 years	Discharged to Home	Emergency	1	Missing	

5 rows × 28 columns

```
[ ] print("Number of records: " + str(df.shape[0]))
    print("Number of features: " + str(df.shape[1]))
```

```
Number of records: 101766
Number of features: 28
```

Dataset description

The columns contain mostly boolean and categorical data (including age and various test results), with just the following exceptions:
`time_in_hospital`, `num_lab_procedures`, `num_procedures`, `num_medications`, `number_diagnoses`.

features	description
race, gender, age	demographic features
medicare, medicaid	insurance information
admission_source_id	emergency, referral, or other
had_emergency, had_inpatient_days, had_outpatient_days	hospital visits in prior year
medical_specialty	admitting physician's specialty

Figure 3: Screenshot of the Google Collab Notebook with Medical Usecase