AI, fairness and bias- Worksheet

Learning outcomes

After completing this worksheet, you will be able to:

- Identify common ethical issues in data science
- Formulate an application as an AI problem, with a specific focus to avoid vagueness and hidden assumptions
- Outline the information needed to solve the AI problem, and plan for data collection
- Assess sources of bias in a dataset
- Evaluate a classification algorithm's accuracy, recall and precision, as well as other fairness metrics
- Consider the risks associated with developing an AI algorithm on a limited dataset, and with applying it to a population different from what it was trained on

Scenario 1

In the country of Dataland, the police department uses an algorithm to assess the risk level of people reporting cases of domestic abuse and violence. Thanks to this algorithm, they can identify the most serious threats and intervene accordingly. The algorithm has had a positive impact, assessing cases with more accuracy than other prior strategies and allowing the police force to make an efficient use of their resources. However, it occasionally fails to correctly identify people at high risk of violence (false negatives), leaving them without the protection they need. It is also affected by other issues. For each issue outlined in this table, check whether it is a Fairness, Accountability or Transparency problem.

Issue	Fairness	Accountability	Transparency
When the algorithm fails to identify a high-risk case and violence occurs, it is unclear if the police department should shoulder any responsibility.			
An analysis of the algorithm's results suggests that false negatives occur more frequently among victims with physical disabilities.			
The majority of people reporting domestic abuse are not aware that their cases are being evaluated by an algorithm, or do not know the score they received.			
The police department receives a recommendation for each case, but does not know which characteristic(s) of the case have resulted in the final evaluation.			
The algorithm was trained using past cases filed by the police department, but the people			

involved were not informed that their information was being used for this purpose.			

In completing your answers, it may help to remember the following definitions:

- **Fairness:** The idea that every group or population that is affected by a technological application is being treated equally and not receiving a different outcome solely because they belong to their group.
- **Accountability:** Clear definition of who should be held responsible for the outcome of the technological application and under what circumstances.
- **Transparency:** The technical definition of transparency in Data Science refers to being able to understand why a technological application produced a specific outcome. This is also called explainability. But transparency can also refer to the demand of making the use of algorithms more transparent to the public, including informing the users about when they are used, where the data used was sourced from, and making algorithms available for auditing.

Note: this case is fictional but inspired by a real algorithm, called VioGén, used in Spain to determine the risk level of victims of gender-based violence and assign protection measures. The algorithm has been recently going under severe scrutiny (Read more).

Scenario 2

You are working with a family physician, Dr. C. Lever, who contacted you because he wants your help to solve a medical problem. Doctor Lever is worried about the rising incidence of type 2 diabetes in his community. He has some familiarity with AI, so he thought about creating an algorithm to help him identify which of his patients, based on available information, are at higher risk of diabetes, so that he could follow up with them and suggest a preventive plan.

Problem definition

This is an example of a **classification** problem. We want the algorithm to take patient data and return whether or not the patient is considered at high risk of diabetes or not (assuming that we are happy with only these two labels, and a third category is not needed, then it is a **binary classification** problem).

The problem seems simple enough but, before we go any further, we need to define *how* we are going to separate the patients in our database between those who have diabetes and those who do not. The table below shows a list of possible information that we could use for this purpose. For each item, write down if you believe it to be good enough to answer our question or not and why. Additionally, even for those items that you believe to be potentially good discriminant, write

down if you think they may be imperfect, carry assumptions, or fail under some circumstances. The first row is filled as an example.

Information	Good discriminant? (yes/no)	Why?	Blind spots/ assumptions
Family history (diabetes in parents or grandparents)	No	The patient may not have diabetes even if someone in their immediate family does	Information may be missing or unavailable (e.g. patients with adoptive parents)
Positive result of glycated hemoglobin (A1C) test (test specifically for type 2 diabetes)			
Patient has an active prescription for insulin			
Patient is obese			
Glycated hemoglobin (A1C) test falling in pre-diabetes range			

You may want to consult these pages to get a basic understanding of type 2 diabetes causes and diagnoses

- https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193
- https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/diagnosis-treatment/drc-20351199

We can count ourself lucky, because type 2 diabetes is a known disease with reliable diagnostic tools. With enough information available (e.g. results of a blood test) we could know exactly if any person has or does not have type 2 diabetes. In other scenarios, creating labels is a much grayer area! Think, for example, how you would separate "people who deserve a loan", or "people who deserve to be released on parole" from people who do not. There is no blood test for that!

Creating the dataset

Now that we have found one or more ways to identify patients at the clinic who have type 2 diabetes (and assuming we would have this information for a good number of patients, if not all), we can move on to determine what information will help us answer the question: "is this patient at high risk of type 2 diabetes"?

A good place to start would be the medical records already available at the clinic. Look at this list of possible information (which we will call **features**), and decide whether or not it should be included in the dataset. Also, write down reasons that may make this information difficult to use. As before, the first row is filled as an example.

Notes:

- This is a toy example, and we do not expect students to have a medical background, so we are going to limit our analysis to basic information, such as the risk factors included at this link: https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193
- A big issue with using personal data (especially medical data) to train algorithms is the issue of **privacy**, meaning the risk that the information could be accidentally leaked and end up revealing sensitive information about a person, information that they would have wished would remain private. The issue of privacy, although very important and prominent, is beyond the scope of this worksheet, and you will not have to list it as a reason not to include some information.

Information	Good feature? (yes/no)	Why?	Blind spots/ assumptions
Glycated hemoglobin (A1C) test falling in pre-diabetes range	Yes	A test result in pre-diabetes range would be a strong signal that the patient is at high risk	Information will not be available for many patients; likely, only those already at medium-high risk have received a test
Patient age			
Family history (diabetes in parents or grandparents)			

Patient weight			
Patient COVID-19 vaccination status			

Dr. Lever is also considering asking patients coming to the clinic to fill a questionnaire about their eating and exercising habits, because he knows that this information will help determine who is at higher risk of diabetes. This is not a bad idea, but it could introduce bias in the dataset.

1. Who is most likely to be excluded from the data collection if the questionnaire were to be administered only to people coming to the clinic?

_____
_____
_____
_____
_____
_____
_____
_____

2. What is this type of bias called?

_____

3. Imagine we changed the strategy and, instead of having patients fill the questionnaire in person at the clinic, we sent it to them by email and asked them to complete it. Do you think this is a better collection system? Could we still accidentally exclude a specific set of patients?

_____
_____
_____
_____
_____
_____
_____
_____

Evaluating models

Now that you have enough data, you can train and evaluate a classification model. There are many to pick from:

- Random Forests
- Support Vector Machines
- Artificial Neural Networks
- …

The details regarding the implementation of these models are not within the scope of this exercise (but we encourage you to look into courses about Machine Learning, if you are interested!). One important thing to understand is that all the models mentioned above behave as **black boxes**: given a certain input (a set of features describing a patient's health status), we will get a classification (high or low risk of diabetes), but not an explanation of *why* the model picked that class (for example, that the patient is obese and has a family history of diabetes). This is important for a few reasons, including:

a) The doctor will not be able to recommend a treatment based on the model's recommendation (such as encouraging the patient to change diet)
b) It implies a certain level of *trust* in the model; of course, some trust is necessary – or the model is rendered useless – but, especially in delicate applications such as this one, the classification should be considered a suggestion and used accordingly by an expert (the doctor)

Let's say that we have narrowed the choice between **Model A** and **Model B**. Below, you can see their **confusion matrices**. Remember that here positive means that the model believes that the patient is at high risk of diabetes.

Model A

		Classification	
		Positive	Negative
True label	Positive	90	30
	Negative	50	830

Model B

		Classification	
		Positive	Negative
True label	Positive	105	15
	Negative	100	780

Complete the following table based on the information available in the confusion matrices:

Question	**Answer**
How many patients are included in the dataset?	
How many of them are at high risk of diabetes?	

How many of them are not at high risk of diabetes?	
What is the accuracy of Model A?	
What is the accuracy of Model B?	
What is the recall of Model A?	
What is the recall of Model B?	
What is the precision of Model A?	
What is the precision of Model B?	

Based on your answers in the table above, which model would you recommend to use for this application, and why?

_____
_____
_____
_____
_____
_____
_____

Because we are aware of the risk of bias in the dataset, we also want to check that our models treat our female and male patients equally (for the purpose of this exercise, we assume that both sexes* are equally at risk of diabetes). Here are the confusion matrices of the two models divided by sex of the patient.

Model A - males

		Classification	
		Positive	Negative
True label	Positive	55	5
	Negative	30	410

Model A - females

		Classification	
		Positive	Negative
True label	Positive	30	30
	Negative	20	420

Model B - males

		Classification	
		Positive	Negative
True label	Positive	40	20
	Negative	20	420

Model B - females

		Classification	
		Positive	Negative
True label	Positive	45	15
	Negative	30	410

What is each model's ratio of positive predictions across the two groups? In other words, do the models achieve **statistical parity?** It may help you write the formula in terms of True Positives, True Negatives etc. before attempting to calculate the result. As you compute the ratios, place the group of males as denominators.

Ratio of positive predictions formula =

Ratio of positive predictions - Model A =

Ratio of positive predictions - Model B =

What is each model's ratio of false positives to predicted positives across the two groups? In other words, do the models achieve **equal opportunity?**

Ratio of false positives to predicted positives formula =

Ratio of false positives to predicted positives - Model A =

Ratio of false positives to predicted positives - Model B =

What is each model's ratio of false negatives to predicted negatives across the two groups? In other words, do the models achieve **predictive equality?**

Ratio of false negatives to predicted negatives formula =

Ratio of false negatives to predicted negatives - Model A =

Ratio of false negatives to predicted negatives - Model B =

What do the ratios of positive predictions say about how the models classify patients? What is the fairest model according to this metric, and why?

What do the ratios of false positives to predicted positives say about how the models classify patients? What is the fairest model according to this metric, and why?

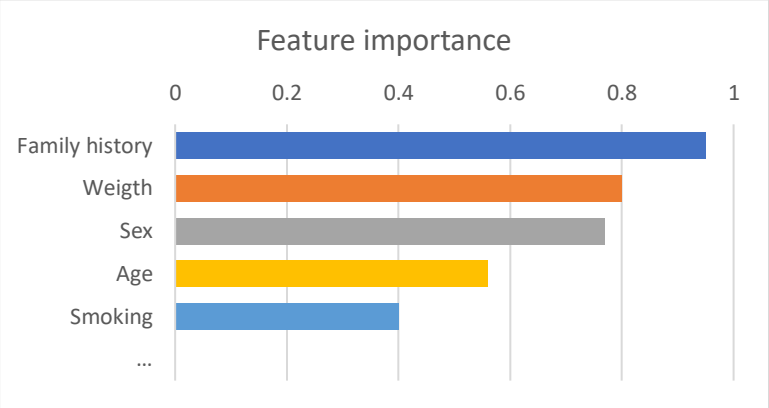
What do the ratios of false negatives to predicted negatives say about how the models classify patients? What is the fairest model according to this metric, and why?

Considering the fairness metrics, as well as accuracy, precision and recall of the two models, would you still recommend the same model you chose earlier for this application? Why or why not?

Even with black box models, it is possible to use techniques to evaluate **feature importance**, that is, which features the model relies more heavily or more frequently on to base its classification. We said earlier that, for this exercise, we assumed that both sexes are equally at risk of diabetes. If sex of the patient was used as a feature in these models, under this assumption, do you think it should have (pick one):

a) high importance
b) low importance
c) can't say/not enough information

When plotting the feature importance, you can see that sex of the patient is quite high. If sex is supposed to not have an impact on the chances of diabetes, what do you think could be the reason behind the model giving it such high importance? Can you think of a way to fix or at least improve this behavior?



A note on Sex and Gender

Sex and Gender have different meanings, despite often being used interchangeably. In humans, sex refers to a set of biological features such as chromosomes and gene expression. It is usually characterized as male or females, although intersex attributes are also possible. Gender refers to socially constructed roles, behaviours and identities, such as man, woman, or gender diverse. Sex can be important to consider in medical and biological applications, while gender can be a source of bias and differential treatment.

"Brave Men" and "Emotional Women": A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5845507/