IN4080 exercise session, week 13: Ethics (with answers)

You have developed an NLP model for automated essay scoring in Norwegian, and you wish to ensure your model is *fair*, in particular when it comes to whether the student is ethnically Norwegian or not.
To this end, you compare the essay scores with scores assigned by experienced teachers. To simplify our problem we will rely on binary pass/fail scores. In addition, we will assume that the human teachers themselves are free from social biases regarding the ethnicity of the students.
Here are the scores produced by your model and by the human teachers for a group of 21 students:

ID	Ethnical Norwegian?	Score from model: Pass (✓) or Fail (**F**)	Score from teachers: Pass (✓) or Fail (**F**)
1	✓	✓	✓
2	No	✓	✓
3	✓	✓	✓
4	✓	**F**	**F**
5	✓	✓	✓
6	✓	**F**	**F**
7	✓	✓	✓
8	No	**F**	**F**
9	No	✓	✓
10	✓	✓	✓
11	✓	✓	✓
12	No	✓	✓
13	✓	✓	✓
14	No	**F**	**F**
15	✓	✓	✓
16	✓	✓	**F**
17	No	✓	✓
18	✓	**F**	✓
19	No	✓	**F**
20	No	**F**	✓
21	✓	✓	**F**

Based on this data, determine which fairness criteria[a] covered during the course (demographic parity, predictive parity and equalised odds) are satisfied or not satisfied by your essay scoring model.

[a] We assume the essay scoring model does not have direct access to the ethnicity of the student, and the "unawareness" criteria is thus irrelevant here.

Possible answer:

First, some notations:

- The two demographic groups will be written eno (ethnically Norwegian) and $\neg eno$ (non-ethnically Norwegian).
- $\hat{Y}$ corresponds to the predictions of the model
- Y corresponds to the scores from the human teachers (which we assume in this exercise to be bias-free, and thus corresponds to some "true" value)

We can then look at various fairness criteria:

Demographic fairness We need to look whether the probabilities of getting a pass or fail are the same across the two groups:

$$P_{eno}(\hat{Y}) \stackrel{?}{=} P_{\neg eno}(\hat{Y}) \tag{1}$$

For the eno group, 3 out of 13 students get a fail from the scoring model, while this proportion rises to 3 out of 8 students for the $\neg eno$ group. The demographic fairness criteria is therefore *not* satisfied.

Predictive parity We need to look at the *precision* of our model predictions (compared to the scores provided by the human teachers):

$$P_{eno}(Y = y|\hat{Y} = y) \stackrel{?}{=} P_{\neg eno}(Y = y|\hat{Y} = y) \tag{2}$$

We can start with the value $y = \checkmark$. For the eno group, we have 10 students that get a pass from the model. 8 of those students also get a pass from the human teachers, which means that the precision $P_{eno}(Y = \checkmark|\hat{Y} = \checkmark) = 0.8$.

For the $\neg eno$ group, we have 5 students that get a pass from the model, and 4 of them also get a pass from the human teachers. The precision $P_{\neg eno}(Y = \checkmark|\hat{Y} = \checkmark)$ is thus also equal to 0.8.

Now, for the value $y = \mathbf{F}$, we can do the same calculations: for the eno group, 3 students failed, and 2 of them were also marked as failed by the human teachers, giving a precision of $2/3$. For the $\neg eno$ group, 3 students failed as well, as 2 were marked as failed by human teachers, which also gives a precision of $2/3$.

In other words, the predictive parity criteria is satisfied.

Equalised odds We need to look at the *recall* of our model predictions compared to the scores provided by the human teachers:

$$P_{eno}(\hat{Y} = y|Y = y) \stackrel{?}{=} P_{\neg eno}\hat{Y} = y|Y = y) \tag{3}$$

We can start with the value $y = \checkmark$. For the eno group, we have 9 students that get a pass from the human teachers. 8 of those students also get a pass from the model, which means that the recall is $^8/_9$.

For the $\neg eno$ group, we have 5 students that get a pass from the human teachers, and 4 of them also get a pass from the human model, giving a recall of $^8/_{10}$.

In other words, the non-ethnical Norwegians will have a higher risk of being a false positive (receiving a fail mark when one should have gotten a pass). An ethnical Norwegian that should receive a pass will have a 11 % change of being mistakenly scored as failed, while this risk increases to 20 % for students that are non ethnical Norwegians.

We do the same calculations for the value $y = \mathbf{F}$: for the eno group, 4 students failed according to the teachers, and 2 of them were also marked as failed by the model, giving a recall of $^1/_2$. For the $\neg eno$ group, 3 students were failed by the human teachers, and 2 of them were marked as failed by the model, which also gives a recall of $^2/_3$.

The criteria of equalised odds is thus *not* satisfied.

> Would you consider your model as being fair to the students that are not ethnical Norwegian? Explain your answer.

Several answers are possible here.

Personally, I would say that it is fine if the demographic fairness criteria is not satisfied: whether a student is ethnically Norwegian or not is presumably correlated with their fluency in Norwegian. And the fluency in Norwegian should be allowed to influence the likelihood of getting a pass/fail score to evaluate the quality of an essay.

However, the fact that the equalised odds criteria is not satisfied is much more problematic. As mentioned above, it means that a "good" student (that should receive a pass) will have a higher chance of being mistakenly attributed a failing score if they are not ethnical Norwegian. And the difference is fairly large, since the non-ethnical Norwegians will have a 20 % risk, compared to an 11 % risk for the ethnical Norwegians. In this light, I would not consider the scoring model to be fair.