IN4080 exercise session, week 13: Ethics

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[1] covered during the course (demographic parity, predictive parity and equalised odds) are satisfied or not satisfied by your essay scoring model.

[1] 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.