Welcome and thank you for taking the time to participate in this survey.

This survey aims to explore how you, as an actual user, perceive explanations. By explanations we mean methods that explain how machine learning models make their decisions in a way that is understandable to humans. The specific methods that we will evaluate are called Counterfactual Explanations.

The next section will give you more background on machine learning models. After that, we will show you two different explanations for which we will ask you the same questions. The goal of this survey is to compare method A with method B.

Keep in mind that there aren't any right or wrong answers - we are purely interested in your opinion!

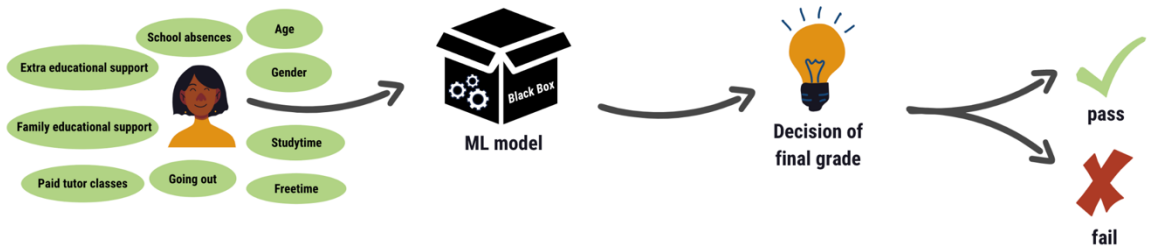
Q1 How familiar are you with machine learning models?

- ○ I am not that familiar with machine learning models
- ○ I am familiar with machine learning models through my studies and/or work
- ○ I am familiar with machine learning models through a different way (define below)

_____

Student ML **Machine Learning (ML) algorithms** are increasingly affecting our lives on a day-to-day basis. Currently the impact is rather small, such as the suggested route of our navigation system while driving. However, algorithms will increasingly be used for **critical decision making** to automate numerous processes. Examples include ML models deciding loan admissions for bank, school admissions for school, or insurance rates for insurance companies.

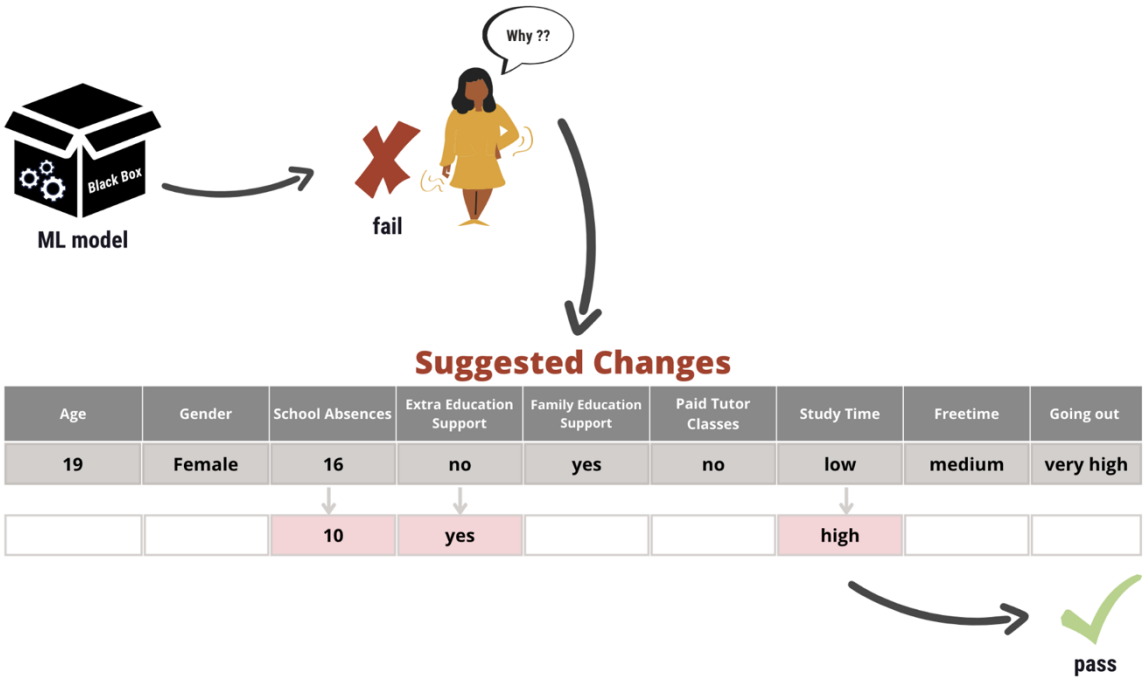
Throughout this study we will look at the following scenario: An ML model needs to decide if a student is **likely to pass or fail a course** based on attributes of the student.



Student CE ML models are often a **black-box** for humans, meaning that we cannot trace back **how and why** the decision is made. This is a critical weakness of using ML models, since it is harder to trust systems that we do not understand. Making ML models more transparent can be achieved by **explaining the reasons behind their decisions**.

Counterfactual explanations (CE) are one approach to provide such an explanation. It explains by showing what attributes **need to change to get a different (preferred) outcome**.

To look at our scenario: A student was classified as likely to "fail the course" and wants to understand how the ML model came to this conclusion. A Counterfactual Explanation suggests what the attributes would need to be to get the preferred outcome, which is likely to "pass the course".



Student1_0 In this survey, we will focus on **two different methods** to explain the outcome to the affected person. In the following, you will be presented with a specific scenario, followed by two different methods to explain the outcome. **The goal of this survey is to compare these methods**.

The scenario is the following:
An **ML model** predicts whether Charlie is likely to pass or fail a course. This decision is made based on personal attributes.
In our scenario, the model concludes that Charlie is likely to “fail the course”. Charlie has the right to an explanation of why the model came to this outcome based on the

attributes. As an explanation, **Counterfactual Explanations** are shown that suggest what attribute change would result in the model outcome “passing the course”.

	Age	Gender	School Absences	Extra Education Support	Family Education Support	Paid Tutor Classes	Study Time	Freetime	Going out	
Original	18	Male	0	no	yes	no	very low	high	low	Fail

How the attributes can change:

Age	Gender	School absences
Number	Female / Male	Number

Extra educational support	Family educational support	Paid tutor classes
Yes / No	Yes / No	Yes / No

Study Time	Freetime	Going out
• Very low • Low • Medium • High • Very high	• Very low • Low • Medium • High • Very high	• Very low • Low • Medium • High • Very high

0 What attribute(s) would you expect to change for Charlie to instead get the outcome of “passing the course”?

- [ ] Age
- [ ] Gender
- [ ] School Absences
- [ ] Extra Educational Support
- [ ] Family Educational Support

- [ ] Paid Tutor Classes
- [ ] Study Time
- [ ] Freetime
- [ ] Going out

We will now look at some explanations, starting with Explanation Method A. We will ask you some question about this, remember once again, there aren't any right or wrong answers - we are purely interested in your opinion!

Explanation Method A provides the following suggestions on how the attributes need to change to get the outcome of "passing the course":

Method A

	Age	Gender	School Absences	Extra Education Support	Family Education Support	Paid Tutor Classes	Study Time	Freetime	Going out	
Original	18	Male	0	no	yes	no	very low	high	low	Fail
Suggestion 1					no					Pass
Suggestion 2							medium			Pass
Suggestion 3							low		high	Pass
Suggestion 4					no		low			Pass
Suggestion 5			2				low	medium		Pass

A1 How surprised are you with the suggested changes in attributes to get the outcome of "passing the course"?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Not at all	○	○	○	○	○	○	○	Very Surprised

A2 How well does the method explain to you what Charlie needs to change to get the model outcome "passing the course"?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Not at all	○	○	○	○	○	○	○	Very Well

A3 Based on the explanation, what attribute(s) would you consider as most important to change the model outcome?

- [ ] Age
- [ ] Gender
- [ ] School Absences
- [ ] Extra Educational Support
- [ ] Family Educational Support
- [ ] Paid Tutor Classes
- [ ] Study Time
- [ ] Freetime
- [ ] Going out

A4 In your opinion, the amount of five different suggestions is _____ to explain the model outcome.

- ○ Enough
- ○ Too little
- ○ Too many

A5 In your opinion, the variation of attributes in the suggestions is _____ to explain the model outcome.

- ○ Enough
- ○ Too little
- ○ Too much

A6 Do you think Charlie could realistically act upon the suggestions to change the model outcome to "passing the course"?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Not at all	○	○	○	○	○	○	○	Fully

A7 Do you think the suggestions make sense in order to retrieve the outcome to "passing the course"?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Not at all	○	○	○	○	○	○	○	A lot

Now, we look at Explanation Method B and answer the same questions. Your answers below should only reflect the explanation of method B.

Explanation Method B provides the following suggestion on how the attributes need to change to get the outcome of "passing the course":

Method B

	Age	Gender	School Absences	Extra Education Support	Family Education Support	Paid Tutor Classes	Study Time	Freetime	Going out	
Original	18.0	Male	0	no	yes	no	very low	high	low	Fail
Suggestion	17									Pass

B1 How surprised are you with the suggested changes in attributes to get the outcome of “passing the course”?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Not at all	○	○	○	○	○	○	○	Very Surprised

B2 How well does the method explain to you what Charlie needs to change to get the model outcome "passing the course"?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Not at all	○	○	○	○	○	○	○	Very Well

B3 Based on the explanation, what attribute(s) would you consider as most important to change the model outcome?

- [ ] Age
- [ ] Gender
- [ ] School Absences
- [ ] Extra Educational Support
- [ ] Family Educational Support
- [ ] Paid Tutor Classes
- [ ] Study Time
- [ ] Freetime
- [ ] Going out

B4 In your opinion, the amount of one suggestion is _____ to explain the model outcome.

- ○ Enough (1)
- ○ Too little (2)
- ○ Too many (3)

B5 In your opinion, the variation of attributes in the suggestion is _____ to explain the outcome.

- ○ Enough
- ○ Too little
- ○ Too much

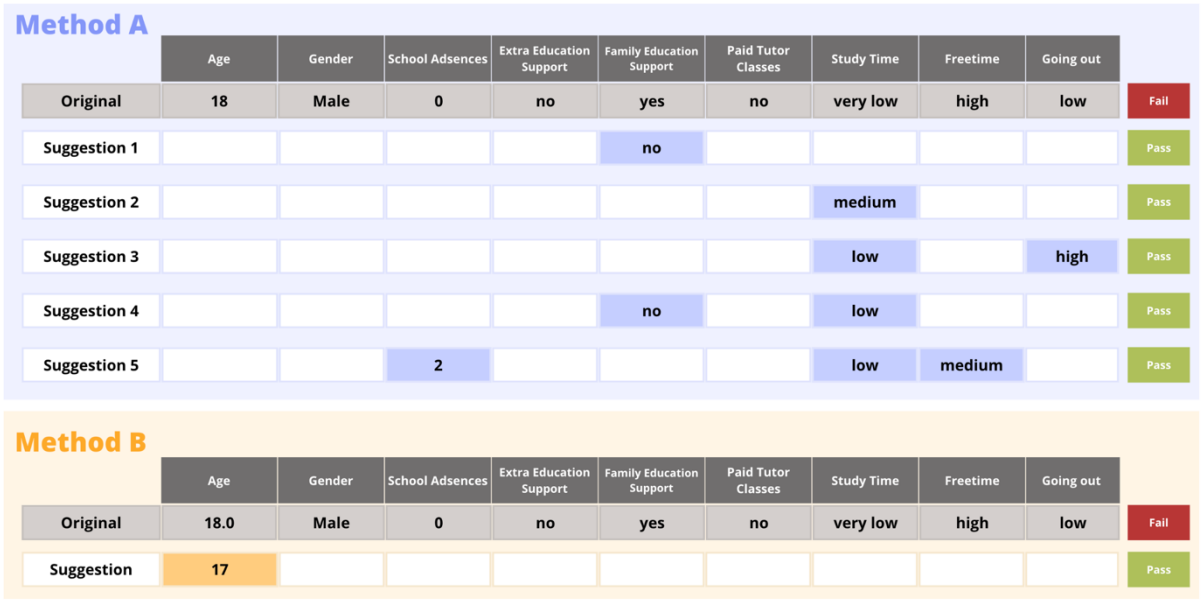
B6 Do you think Charlie could realistically act upon the suggestion to change the model outcome to "passing the course"?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Not at all	○	○	○	○	○	○	○	Fully

B7 Do you think the suggestion makes sense in order to retrieve the outcome to "passing the course"?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Not at all	○	○	○	○	○	○	○	A lot

Comparison of Explanation A and B



Which explanation method would you prefer as an explanation for the outcome of the ML model?

- ○ Method A
- ○ Method B