Raspberry Pi Foundation

Avoiding bias

Your school has tried using a machine learning (ML) model to predict which subject you will study. During the trial, it has appeared that the predictions are inaccurate.

Task 1

Your school has asked the software company that made the ML model to tell them what data was used to train the model.

Look at the sets of data below and evaluate which ones would be suitable and which ones would be unsuitable.

Sets of data	Suitable or unsuitable?
Data from this school (last 10 years)	
Regional data (last 3 years)	
National data (1960–2000)	
International data	

For the set(s) of data that you have identified as **unsuitable**, write a short justification as to why they would be unsuitable.

Explorer task

The following fields of data have been included in the sets of data. In an attempt to reduce the bias, which would you remove from the data used to train the model?

Fields	Remove?
Achievement in each subject (ages 11–16)	
Name	
Extracurricular activities	
Gender	
Career ambitions	
Teacher subject recommendations	
Behaviour record	

Explain how bias could be introduced into the model if all of the fields of data above were used for training.