

Data Set for evaluating an in-session dropout prediction model for fairness

Files: preprocessed_fairness_data.pkl

Survey method: This is an anonymized sample from the original dataset. Data were collected on the Orthografietrainer.net learning platform from between March and April 2020. This is clickstream data that is automatically logged when a user is on the platform. Demographic data was collected as part of a survey that was displayed to each user 3 months after registration on the platform. Answering the survey was voluntary. Not all features listed are part of the model training.

Variable	Variable name	Measurement Unit	Allowed values	Description
ID	ID	Numeric	0-1000	ID of the users' task processing
Session ID	UebungsID	Numeric	0-1000	ID of the session
Sentence ID	satzID	Numeric	0-1000	ID of the sentence
First solution	Erstloesung	Numeric	0-100	Count of sentences that are displayed to the user for the first time
Distracted	Schussel	Numeric	0-100	Count of sentences where the user submitted a task while missing a field
Date	Datum	Date	dd-mm-yyyy	Date when the task was processed
Success	Erfolg	Numeric	0-100	Count of correct answers
Number of the solution	Loesungsnr	Numeric	0-100	Number of the attempt this task was processed by this user
Difficulty	Schwierigkeit	Numeric	-10 - 10	Mean difficulty of the sentences
Parental Education	AbiEltern	Numeric	[0,1,2]	Indicates whether one [1], both [2], or neither [0] parent of the user a high school diploma
Books	Buecher	Numeric	[10,50,100,200]	Estimated number of books in the users' household. The following options were possible: max ten books [10], max 50 books [50], max 100 [100], more than 100 [200].
First Language	eigSprache	Numeric	[0.0,1.0]	Indicates whether the users' first spoken language is

				German [1.0] or not [0.0]
User Attribute	UserAttribut	Numeric	[0,1]	Group of users is student [1] or other [0]
Mistakes	Fehler	Numeric	0-100	Count of mistakes
Class level	Klassenstufe	Numeric	5-12	Class level of the user
Years registered	Jahredabei	Numeric	0-10	Count of years the user is registered on the platform
Count of pending tasks	AnzahlAufgaben	Numeric	0-100	Count of pending tasks
School hours	Ist_Schulzeit	Numeric	[0,1]	Session was conducted during school hours (8 am to 2 pm)
Multiple false	MehrfachFalsch	Numeric	0-100	Count of sentences that were answered incorrectly several times
Matrix	OrderNumber	Numeric	15	Number of the matrix indicating the current sentence number
Previous Break	Vorher_abgebrochen	Numeric	[0,1]	Describes if the assignment was interrupted earlier
Drop out	y	Numeric	[0,1]	Describes if the user is predicted to drop out of the session or not.
Type Capitalization	Art__GK	Numeric	[0,1]	Type of task was capitalization
Type Grammar	Art__GR	Numeric	[0,1]	Type of task was grammar
Type Hyphenation	Art__GZ	Numeric	[0,1]	Type of task was hyphenation
Type Comma Formation	Art__K	Numeric	[0,1]	Type of task was comma formation
Type Sounds and Letters	Art__LB	Numeric	[0,1]	Type of task was sounds and letters
Homework	HA__HA	Numeric	[0,1]	Type of session: homework
Voluntary work	HA__Self	Numeric	[0,1]	Type of session: voluntary work (not assigned by a teacher)
Post-test	HA__nt	Numeric	[0,1]	Type of session: post-test
Pre-test	HA__vt	Numeric	[0,1]	Type of session: pre-test

Interim test	HA__zt	Numeric	[0,1]	Type of session: interim test
Gender male	Sex__m	Numeric	[0,1]	Gender male
Gender female	Sex__w	Numeric	[0,1]	Gender female
Testposition test	Testposition__pruefung	Numeric	[0,1]	Sentences was displayed in test mode
Testposition training	Testposition_training	Numeric	[0,1]	Sentence was displayed in training mode
Testposition version	Testposition_version	Numeric	[0,1]	Sentence was displayed in version mode
Steps	steps	Numeric	0	Describes the difference between the next possibility to finish an assignment and the current sentence number