Technische Hochschule Augsburg

INVESTIGATION OF THE MOST IMPORTANT STUDENT PERFORMANCE FACTORS AND THE PREDICTION OF STUDENT EXAM SCORES

Michael Mertl
Data Science
Master Computer Science

Introduction

Student Tom

Tom started to study

But failed all his exams in the first semester…

Although he thought he studied enough and went to all lectures

How can we help him?

Research Question

Which factors influence student performance the most, and how can we predict a student's performance based on these factors?

(& How can I learn the most about data science in one semester with this topic?)

Dataset

Kaggle[1] provides a "Student Performance Factors" dataset[2]

Dataset Description:
This dataset provides a comprehensive overview of various factors affecting student performance in exams.

Column descriptions

Facts about the Dataset:
- 876 Upvotes
- Usability: 10.00
- 223k Views
- 57.1k Downloads

(as of 12/10/2024)

Attribute	Description
Hours_Studied	Number of hours spent studying per week.
Attendance	Percentage of classes attended.
Parental_Involvement	Level of parental involvement in the student's education (Low, Medium, High).
Access_to_Resources	Availability of educational resources (Low, Medium, High).
Extracurricular_Activities	Participation in extracurricular activities (Yes, No).
Sleep_Hours	Average number of hours of sleep per night.
…	…
Exam_Score	Final exam score.

[1]C. Staff, "What Is Kaggle and What Is It Used For?," *Coursera*, 17 Aug. 2023. https://www.coursera.org/articles/kaggle.
[2]Practice Data Analysis With Me, Student Performance Factors. [Online]. Available: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors.

Methodology

1. Data investigation

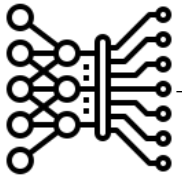
- ➔ Structure
- ➔ Column types
- ➔ Detect missing values

2. Data preprocessing

- ➔ Clean missing values
- ➔ Create encoded dataset (only numeric)

3. Statistics

- ➔ Heatmaps (normal + encoded)
- ➔ Correlation plots

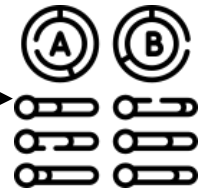


4. AutoML: AutoGluon[3]

- ➔ Split data
- ➔ Train data
- ➔ Tabular-Predictor
- ➔ Multi-Modal-Predictor

5. Determine best model

- ➔ Test data
- ➔ Predictions
- ➔ Comparison to determine best model
- ➔ Get feature importance



6. Comparison to others

- ➔ Get other prediction models from Kaggle[4]
- ➔ Compare to my best model
- ➔ Determine the best model

kedro

7. Kedro[5] pipeline for the whole process

[3]AutoGluon, AutoGluon. [Online]. Available: https://auto.gluon.ai/stable/index.html.
[4]Find Open Datasets and Machine Learning Projects | Kaggle. [Online]. Available: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/code?datasetId=5630996&sortBy=voteCount.
[5]Introduction to Kedro — kedro 0.19.10 documentation. [Online]. Available: https://docs.kedro.org/en/stable/introduction/index.html.

Results -> Pipeline Demo

1. Data investigation

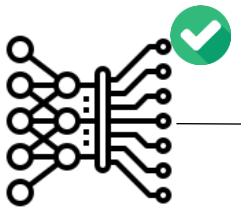
- ➔ Structure
- ➔ Column types
- ➔ Detect missing values

2. Data preprocessing

- ➔ Clean missing values
- ➔ Create encoded dataset (only numeric)

3. Statistics

- ➔ Heatmaps (normal + encoded)
- ➔ Correlation plots



4. AutoML: AutoGluon[3]

- ➔ Split data
- ➔ Train data
- ➔ Tabular-Predictor
- ➔ Multi-Modal-Predictor

5. Determine best model

- ➔ Test data
- ➔ Predictions
- ➔ Comparison to determine best model
- ➔ Get feature importance

6. Comparison to others

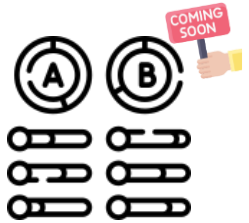
- ➔ Get other prediction models from Kaggle[4]
- ➔ Compare to my best model
- ➔ Determine the best model

kedro WORK in PROGRESS

7. Kedro[5] pipeline for the whole process

[3]AutoGluon, AutoGluon. [Online]. Available: https://auto.gluon.ai/stable/index.html.
[4]Find Open Datasets and Machine Learning Projects | Kaggle. [Online]. Available: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/code?datasetId=5630996&sortBy=voteCount.
[5]Introduction to Kedro — kedro 0.19.10 documentation. [Online]. Available: https://docs.kedro.org/en/stable/introduction/index.html.

What`s next



6. Comparison to others

➔ Get other prediction models from Kaggle[6]
➔ Compare to my best model
➔ Determine the best model

kedro WORK in PROGRESS

7. Kedro[7] pipeline for the whole process

➔ Finish kedro pipeline
➔ Make every useful information visible in the pipeline

8. Cleanup & finish research

➔ Once the research is finished, it needs to be made tracible and reproducible
➔ Moreover, sources need to be added that explain for e.g. my weighted ensemble model and the different level the TabularPredictor from AutoGluon uses
➔ GitHub repo needs to be cleaned to be made public
➔ Focus strives slowly but surely to the paper

But…

Tom already knows what factors influences his exam scores the most and is now able to only do the necessary work for each exam to archive 50% thanks to the prediction model.

[6]Find Open Datasets and Machine Learning Projects | Kaggle. [Online]. Available: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/code?datasetId=5630996&sortBy=voteCount.
[7]Introduction to Kedro — kedro 0.19.10 documentation. [Online]. Available: https://docs.kedro.org/en/stable/introduction/index.html.

Technische Hochschule Augsburg

MICHAEL MERTL

Fakultät Informatik
Masterstudent

Technical University of Applied Sciences Augsburg
An der Hochschule 1
86161 Augsburg
T +49 821 5586 3213
F +49 821 5586 3253
michael.mertl1@tha.de
www.tha.de

Questions? Ideas? Feedback?