KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Prognosemodelle (Klassifikation und Regression): 02_05Daten _Datenbeschreibung

Datenbeschreibung

Erarbeitet von
Dr. Katja Theune

Lernziele

- Du kannst beurteilen, ob sich die Daten für den Anwendungsfall eignen
- Du kannst wichtige Aspekte der Datenqualität bezogen auf den Anwendungsfall erläutern

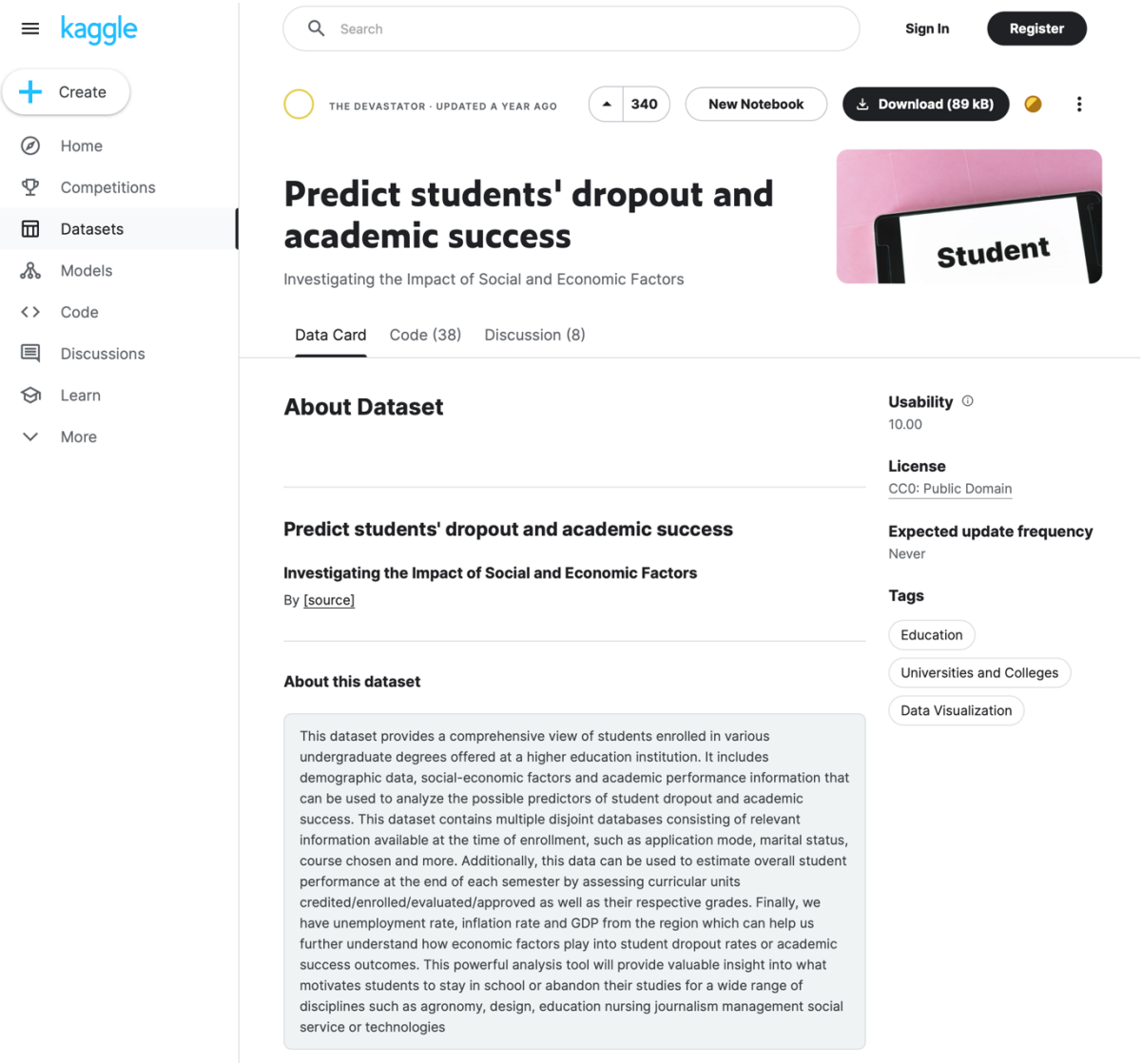
Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Bisher haben wir uns eher theoretisch mit dem Thema Prognosen beschäftigt. Um die Verfahren und Theorien ansprechend zu veranschaulichen, haben wir uns eines fiktiven Datensatzes bedient. Damit konnten wir uns Inputs und Outputs aussuchen und Werte wählen, die zu schönen Grafiken und nachvollziehbaren Ergebnissen führen. In der Praxis sieht das natürlich meistens etwas anders aus. Damit es nun etwas praktischer wird, wollen wir uns jetzt einmal einen realen Datensatz anschauen.

Datenbeschreibung: Allgemeiner Überblick



Wir verwenden den Datensatz “Predict students dropout and academic success”.

Quelle [1]

Die Daten stammen von kaggle und dort wiederum von hier:

Quelle [2]

Es geht also um die Vorhersage des Erfolgs- und Misserfolgs von Studierenden, und zwar einer Hochschule in Portugal. Wir betrachten hier 4424 Bachelorstudierende aus unterschiedlichen Kursen, die sich zwischen 2008 und 2019 eingeschrieben haben. Dieser Datensatz ist aus mehreren Datensätzen zusammenfügt, u. a. aus Daten aus dem Management-System der betrachteten Hochschule und auch aus einem Datensatz, der makroökonomische Informationen enthält.

Quelle [3]

Datenbeschreibung: Inputs und Output

Table 1. Attributes used grouped by class of attribute.

Class of Attribute	Attribute	Type
Demographic data	Marital status	Numeric/discrete
	Nationality	Numeric/discrete
	Displaced	Numeric/binary
	Gender	Numeric/binary
	Age at enrollment	Numeric/discrete
	International	Numeric/binary
Socioeconomic data	Mother's qualification	Numeric/discrete
	Father's qualification	Numeric/discrete
	Mother's occupation	Numeric/discrete
	Father's occupation	Numeric/discrete
	Educational special needs	Numeric/binary
	Debtor	Numeric/binary
	Tuition fees up to date	Numeric/binary
	Scholarship holder	Numeric/binary
Macroeconomic data	Unemployment rate	Numeric/continuous
	Inflation rate	Numeric/continuous
	GDP	Numeric/continuous
Academic data at enrollment	Application mode	Numeric/discrete
	Application order	Numeric/ordinal
	Course	Numeric/discrete
	Daytime/evening attendance	Numeric/binary
	Previous qualification	Numeric/discrete
Academic data at the end of 1st semester	Curricular units 1st sem (credited)	Numeric/discrete
	Curricular units 1st sem (enrolled)	Numeric/discrete
	Curricular units 1st sem (evaluations)	Numeric/discrete
	Curricular units 1st sem (approved)	Numeric/discrete
	Curricular units 1st sem (grade)	Numeric/continuous
	Curricular units 1st sem (without evaluations)	Numeric/discrete
Academic data at the end of 2nd semester	Curricular units 2nd sem (credited)	Numeric/discrete
	Curricular units 2nd sem (enrolled)	Numeric/discrete
	Curricular units 2nd sem (evaluations)	Numeric/discrete
	Curricular units 2nd sem (approved)	Numeric/discrete
	Curricular units 2nd sem (grade)	Numeric/continuous
	Curricular units 2nd sem (without evaluations)	Numeric/discrete
Target	Target	Categorical

Der Datensatz beinhaltet insgesamt 34 Inputs und einen Output.
Als Inputs finden wir einige Informationen zu Beginn der Einschreibung. Dazu gehören demographische und sozio-ökonomische Informationen über die Studierenden, wie z. B. Geschlecht, Alter oder auch der Bildungsstand und der Berufsstatus der Eltern. Akademische

Informationen beinhalten z. B. den Kurs, in den sich die Studierenden eingeschrieben haben. Darüber hinaus finden wir auch Auskünfte über ihre akademischen Leistungen am Ende des ersten und zweiten Semesters, wie z. B. die Noten. Interessant sind auch ein paar makroökonomische Inputs, wie die Arbeitslosenquote, die Inflationsrate und das Bruttoinlandsprodukt der Region.

Curricular units 2nd sem (evaluations)	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	Curricular units 2nd sem (without evaluations)	Unemployment rate	Inflation rate	GDP	Target
0	0	0.0	0	10.8	1.4	1.74	Dropout
6	6	13.666666666666666	0	13.9	-0.3	0.79	Graduate
0	0	0.0	0	10.8	1.4	1.74	Dropout
10	5	12.4	0	9.4	-0.8	-3.12	Graduate
6	6	13.0	0	13.9	-0.3	0.79	Graduate
17	5	11.5	5	16.2	0.3	-0.92	Graduate
8	8	14.345	0	15.5	2.8	-4.06	Graduate
5	0	0.0	0	15.5	2.8	-4.06	Dropout
7	6	14.142857142857142	0	16.2	0.3	-0.92	Graduate
14	2	13.5	0	8.9	1.4	3.51	Dropout
7	5	14.2	0	13.9	-0.3	0.79	Graduate
8	7	13.214285714285714	0	12.7	3.7	-1.7	Graduate
0	0	0.0	0	12.7	3.7	-1.7	Dropout
8	5	11.0	0	8.9	1.4	3.51	Graduate
5	5	12.0	0	10.8	1.4	1.74	Graduate
7	0	0.0	0	15.5	2.8	-4.06	Dropout
14	2	11.0	0	10.8	1.4	1.74	Enrolled
8	8	14.545	0	15.5	2.8	-4.06	Graduate
8	4	12.25	2	10.8	1.4	1.74	Graduate
8	6	13.5	0	16.2	0.3	-0.92	Enrolled
0	0	0.0	0	11.1	0.6	2.02	Graduate
9	8	11.425	0	12.7	3.7	-1.7	Enrolled
12	7	12.857142857142858	0	12.7	3.7	-1.7	Graduate
7	6	12.285714285714286	0	11.1	0.6	2.02	Graduate
9	7	14.114285714285714	0	11.1	0.6	2.02	Graduate
12	4	11.0	0	7.6	2.6	0.32	Enrolled
9	6	13.285714285714286	0	16.2	0.3	-0.92	Graduate
7	4	13.0	0	9.4	-0.8	-3.12	Enrolled
6	5	12.333333333333334	0	16.2	0.3	-0.92	Graduate
7	6	13.716666666666669	0	16.2	0.3	-0.92	Enrolled
17	5	10.571428571428571	0	16.2	0.3	-0.92	Enrolled
9	4	13.4	0	8.9	1.4	3.51	Graduate
8	2	13.5	0	8.9	1.4	3.51	Enrolled
8	6	14.375	0	12.4	0.5	1.79	Graduate
9	5	13.428571428571429	0	13.9	-0.3	0.79	Graduate
7	1	10.0	0	8.9	1.4	3.51	Dropout
0	0	0.0	0	7.6	2.6	0.32	Dropout
8	2	12.0	2	16.2	0.3	-0.92	Dropout

Hier sehen wir einen Ausschnitt aus den Daten, die wir nun weiterverwenden wollen.
Der Output, hier als Target bezeichnet, ist als Drei-Klassen-Problem formuliert. Es gibt die Klasse „Dropout", also „Studienabbruch", „Graduate", also „Studienerfolg" und „Enrolled", was „noch eingeschrieben" bedeutet. Der Status wird am Ende der üblichen Kursdauer festgehalten. Das sind hier bei den meisten Kursen drei Jahre. Für unser Prognosemodell wollen wir aber nur die beiden Klassen „Dropout" und „Graduate" betrachten.
Weitere Informationen über die einzelnen Inputs, den Output und ihre Ausprägungen findet man z. B. hier:

Quelle [3]

Insgesamt sehen wir, dass sich der Datensatz sehr gut für unseren Anwendungsfall eignet. Er beinhaltet viele Studierende, sehr viele relevante Inputs, die verschiedene Aspekte des Studienverlaufs abdecken, und einen passenden Output.

Datenqualität

Ein wichtiger Punkt bei der Datenqualität ist die Repräsentativität der Daten bzw. das Problem der Selektion. Meistens sind in unseren Daten bestimmte Beobachtungen unter- oder überrepräsentiert. Über dieses Thema und seine Auswirkungen kann man ganze Bücher füllen. Ist die Repräsentativität nicht gewährleistet, dann ist eine Generalisierbarkeit der Ergebnisse auf die Grundgesamtheit nicht möglich. Gerade wenn es um Umfragen zu Studienerfolgen geht, hat man häufig das Problem, dass Studierende mit Misserfolgen unterrepräsentiert sind. Denn üblicherweise machen lieber Personen bei Umfragen mit, die engagiert und erfolgreich sind. Man nennt das auch „Erfolgsbias".

Quelle [4]

So entspricht dann der Anteil der Studienabbrüche im Datensatz meistens nicht dem Anteil der Studienabbrüche in der jeweiligen Grundgesamtheit. Bei unserem vorliegenden Datensatz können wir keine endgültige Aussage darüber treffen, wie repräsentativ dieser z. B. bezüglich der Charakteristika der Studierenden und insbesondere auch bezüglich der Klassen ist. Wir sollten das Problem der Repräsentativität, was vermutlich auch hier besteht, aber immer im Hinterkopf behalten.

Auch fehlende Werte können einem Selektionsproblem unterliegen. Hier können wir uns vorstellen, dass insbesondere nicht so erfolgreiche Studierende keine Angaben z. B. über ihre Noten machen wollen. Da der Datensatz dahingehend schon bereinigt ist, also keine Lücken mehr enthält, können wir hier auch dazu keine konkreten Aussagen machen. Aber auch das kann bei der Interpretation der Ergebnisse von Bedeutung sein.

Das waren jetzt nur zwei Punkte der Datenqualität, die diskutiert werden müssen, wenn wir mit realen Daten arbeiten möchten. Es gibt aber natürlich noch einige weitere Aspekte zu berücksichtigen, die uns dabei helfen, die Daten zu verstehen.

Abschluss

Wir kennen jetzt einen konkreten Datensatz, seinen Aufbau und die dort enthaltenen Inputs und Outputs. Zudem haben wir Aspekte bezüglich seiner Qualität diskutiert. Insgesamt eignet er sich aufgrund seiner vielfältigen Inputs und dem Output gut, um unseren Anwendungsfall der Studienabbruchsprognose bearbeiten zu können.

Quellen

Quelle [1] https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data

Quelle [2] Realinho, V., Machado, J., Baptista, L., & Martins, M.V. (2021). Predict students' dropout and academic success (1.0)[Data set]. Zenodo. https://doi.org/10.5281/zenodo.5777340

Quelle [3] Realinho, V., Machado, J., Baptista, L., Martins, M.V. (2022). Predicting Student Dropout and Academic Success. *Data*, 7(11), 146. https://doi.org/10.3390/data7110146

Quelle [4] Pannier, S., Rendtel, U., & Gerks, H. (2020). Die Prognose von Studienerfolg und Studienabbruch auf Basis von Umfrage- und administrativen Prüfungsdaten. *AStA Wirtschafts- und Sozialstatistisches Archiv,* 14, 225–266. https://doi.org/10.1007/s11943-020-00278-5

Disclaimer

Transkript zu dem Video „Prognosemodelle (Klassifikation und Regression): Datenbeschreibung“, Dr. Katja Theune.