Raport

Klaudia Gruszkowska, Bartosz Jamroży

16 kwietnia 2021

1 Opis zbioru danych

Zbiór dotyczy danych z dwóch portugalskich szkół średnich. Dane dotyczą cech demograficznych, społecznych i związanych ze szkołą. Zebrany był z kwestionariuszy i danych, które przekazała szkoła.

Opis cech:

	ď»ż"name	type	description"
0	"school	string	student's school (binary: "GP" Gabriel Pereira or "MS" Mousinho da Silveira)"
1	"sex	string	student's sex (binary: "F" female or "M" male)"
2	"age	integer	student's age (numeric: from 15 to 22)"
3	"address	string	student's home address type (binary: "U" urban or "R" rural)"
4	"famsize	string	family size (binary: "LE3" less or equal to 3 or "GT3" greater than 3)"
5	"Pstatus	string	parent's cohabitation status (binary: "T" living together or "A" apart)"
6	"Medu	integer	mother's education (numeric: 0: none, 1: primary education (4th grade), 2: 5th to 9th grade, 3 _ secondary education or 4 _ higher education)"
7	"Fedu	integer	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 _ 5th to 9th grade, 3 _ secondary education or 4 _ higher education)"
8	"Mjob	string	mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")"
9	"Fjob	string	father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")"
10	"reason	string	reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")"
11	"guardian	string	student's guardian (nominal: "mother", "father" or "other")"
12	"traveltime	integer	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)"
13	"studytime	integer	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)"
14	"failures	integer	number of past class failures (numeric: n if 1<=n<3, else 4)"
15	"schoolsup	string	extra educational support (binary: yes or no)"
16	"famsup	string	family educational support (binary: yes or no)"
17	"paid	string	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)"
18	"activities	string	extra-curricular activities (binary: yes or no)"
19	"nursery	string	attended nursery school (binary: yes or no)"
20	"higher	string	wants to take higher education (binary: yes or no)"
21	"internet	string	Internet access at home (binary: yes or no)"
22	"romantic	string	with a romantic relationship (binary: yes or no)"
23	"famrel	integer	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)"
24	"freetime	integer	free time after school (numeric: from 1 - very low to 5 - very high)"
25	"goout	integer	going out with friends (numeric: from 1 - very low to 5 - very high)"
26	"Dalc	integer	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)"
27	"Walc	integer	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)"
28	"health	integer	current health status (numeric: from 1 - very bad to 5 - very good)"
29	"absences	integer	number of school absences (numeric: from 0 to 93)"
30	"G1	integer	first period grade (numeric: from 0 to 20)"
31	"G2	integer	second period grade (numeric: from 0 to 20)"
32	"G3	integer	Predictor Class: final grade (numeric: from 0 to 20)"

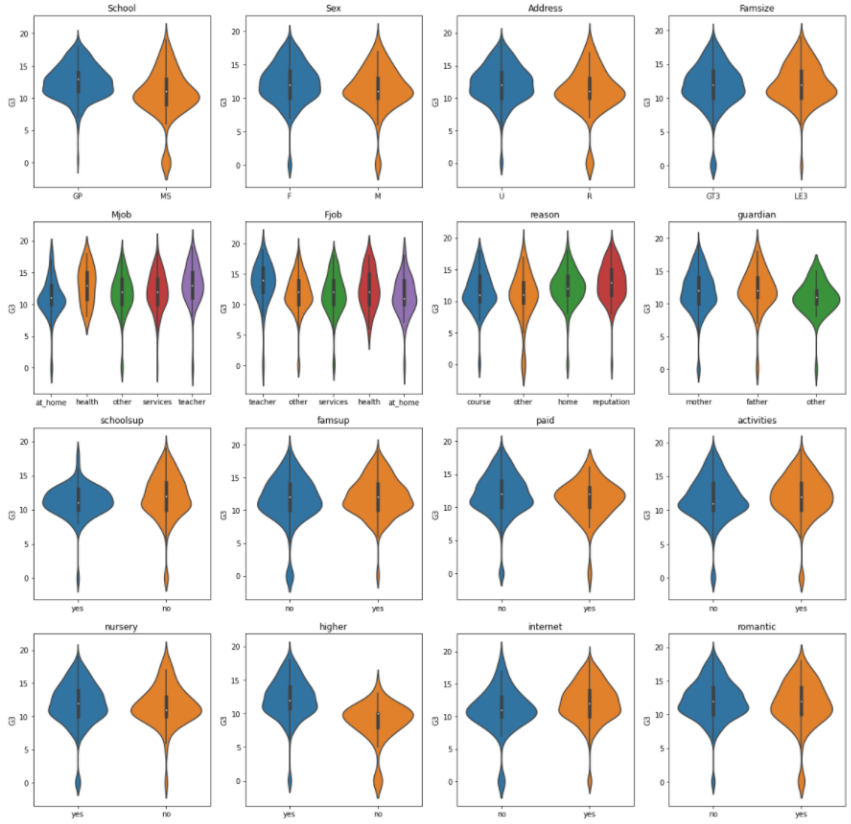
2 Opis problemu

Dane wykorzystujemy do problemu predykcji oceny końcowej ucznia. W tym celu wykorzystamy metody klasyfikacji.

3 EDA

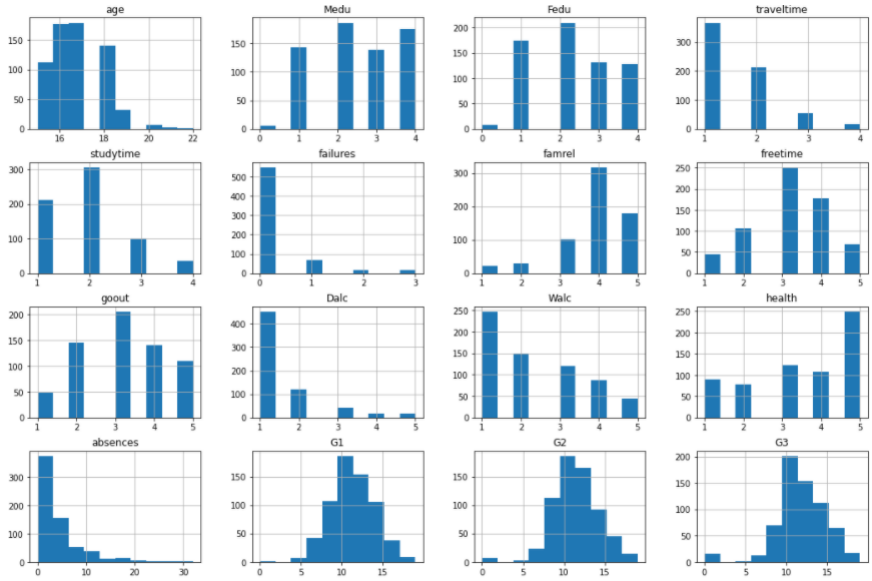
Na początku naszej pracy ze zbiorem wykorzystaliśmy metody EDA aby lepiej poznać sam zbiór, poszczególne cechy jak i zależności między nimi.

Po pierwsze przyjrzeliśmy się naszym zmiennym kategorycznym.



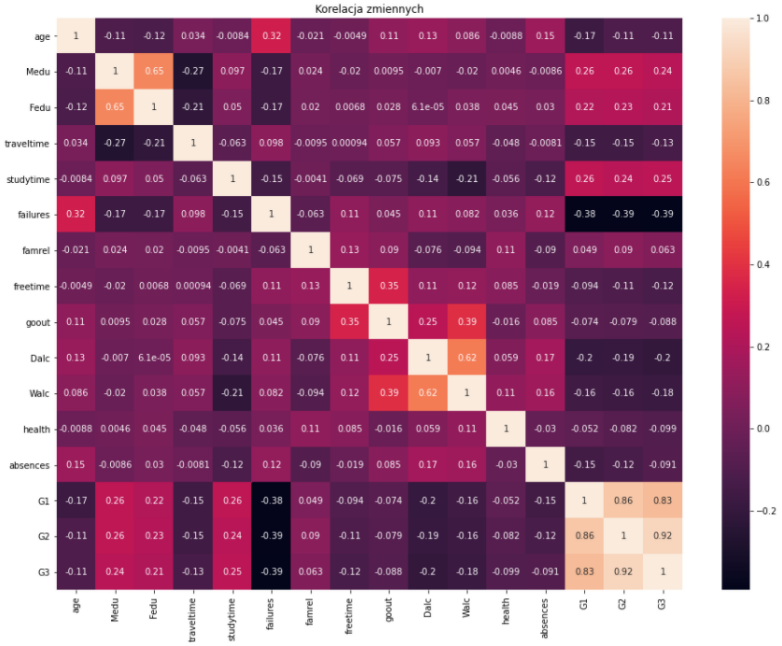
Z tych wykresów udało nam się odczytac kilka zalezności np:
- najwyższe oceny otrzymują uczniowie, którzy wybrali szkołę ze zwględu na jej reputację
- średnie oceny są wyższe dla szkoły Gabriel Pereira niż Mousinho da Silveira

Następnie zwróciliśmy uwagę na rozklady danych numeryczne:



Z tego wykresu odzytaliśmy, że zmienne G1, G2, G3 są nienaturalnie pochylone na prawo.

Jednak najwięcej dowiedzieliśmy się po stworzeniu macierzy korelacji.



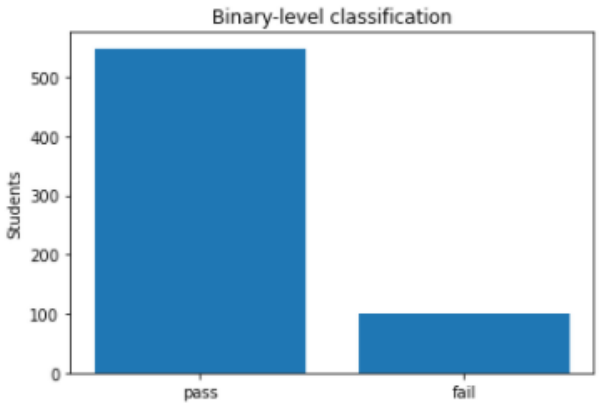
Dzięki niej zobaczyliśmy, że największe korelacje są pomiędzy zmiennymi ocen semestralnymi a zmienną celu. Zgodnie z tym oraz z pomysłem autorów artykułu, który bazuje na tych danych, podjęliśmy decyzję aby rozważać przypadki z użyciem i bez użycia tych zmiennych. Logiczne jest to, że ocena końcowa ucznia mocno zależy od ocen semetralnych jednak gdy pozbawimy model tych danych, nasze wnioski mogą być ciekawsze.

4 Inżynieria cech

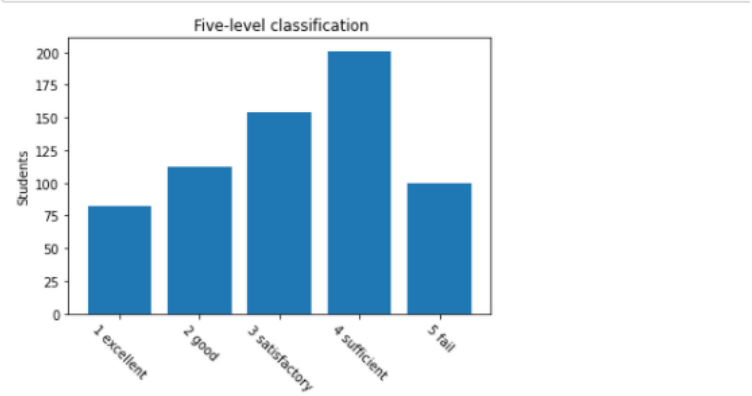
Po wykonaniu analizy danych stwierdzilismy, że rozbijemy je na dwa podproblemy :
- klasyfikacja binarna

- klasyfikacja 5-poziomowa

Przyjrzeliśmy się rozkładowi zmiennej celu. Zauważyliśmy, że przy podziale zbioru do klasyfikacji binarnej ilość obserwacji dla klas jest mocno niezbalansowana.



Jeżeli chodzi o klasyfikację 5-poziomową tutaj balans klas jest już lepszy.



Do encodingu zmiennych kategorycznych użyliśmy trzech metod encodingu:
- one-hot encoding
- target encoding
- mapowanie słownika

Wybralismy takie metody dopasowując je do naszych zmiennych. Dla tych, które zawierały tylko dwie klasy 'yes' i 'no' użyliśmy mapowanie ('yes':1, 'no':0). Dla zmiennych, którym nie chcieliśmy nadawać sztucznego porządku np dla rodzaju szkoły lub płci wykorzystaliśmy target encoding. Dla pozostałych zmiennych aby nie zwiększać za bardzo wymiarowości zbioru danych wykorzystaliśmy target encoding.

Po tych zamianach uzyskaliśmy zbiór danych bez zmiennych kategorycznych, gotowy do pracy.

5 Wstępne modele

Na początku sprawdziliśmy jak wiele modeli radzi sobie na naszych danych z parametrami domyślnymi. Najlepsze okazały się :

-dla klasyfikacji binarnej

Z powodu niezbalansowania klass przy ocenie modeli binarnych stosowaliśmy miary recall i precision. W tabeli przedstawiamy wyniki dla kombinacji powyższych miarę f1.

Model	LR	DT	SVM	GNB	RF	AdaBoost	GradientBoosting	XGB
Z G1 i G2	0.94	0.97	0.92	0.95	0.97	0.97	0.96	0.97
Bez G1 i G2	0.93	0.88	0.92	0.92	0.9	0.95	0.93	0.93

Miara f1 dla różnych modeli w dwóch wariantach. Z ocenami semestralnymi (G1 i G2) oraz przy pominięciu tych cech

- dla klasyfikacji 5-poziomowej

W klasyfikacji 5-poziomowej podział kolumny celu jest w miarę zrównoważony. Dlatego do porównywania wyników używamy domyślnego accuracy.

Model	RandomForest	GaussianNB
Z G1 i G2	0.77	0.54
Bez G1 i G2	0.44	0.35

Miara accuracy dla różnych modeli w dwóch wariantach. Z ocenami semestralnymi (G1 i G2) oraz przy pominięciu tych cech

6 Końcowe modele

Następnym krokiem było strojenie wybranych modeli. Kryterium wyboru modelów do strojenia były zadawalające wyniki podczas wstępnego modelownia. Wybraliśmy cztery różne modele. Strojenie realizowaliśmy poprzez GridSearch na wybranych parametrach .

	Klasyfikacjia binarna, f1	Klasyfikacjia 5-poziomowa, accuracy
Z G1 i G2	LogisticRegression f1=0.97	RandomForest ACC=0.74
Bez G1 i G2	AdaBoostClassifier f1= 0.94	GaussianNB ACC=0.39

Wyniki wytrenowanych modeli

7 Podsumowanie

Wyniki zadania są zadowalająco wysokie. Modele, które korzystały ze zmiennych określających oceny semestralne uzyskały wyższe wyniki jednak taką zależność przewidywaliśmy już wcześniej. Dodatkowo modele klasyfikacji binarnej uzyskiwały lepsze wyniki od klasyfikacji 5-poziomowej. Co ciekawe wynik klasyfikacji binarnej nie korzystającej ze zmiennych G1 i G2 jest niewiele gorszy a w niektórych modelach bardzo podobny do tych korzystających z G1 i G2.

8 Źródła

Dane: `https://www.apispreadsheets.com/datasets/110`