Metody Eksploracji Danych - Projekt - Analiza Eksploracyjna

Jakub Matłacz, Kamil Trysiński

3 stycznia 2022

Streszczenie

Termin – 3 stycznia 2022r. Ilość punktów do zdobycia – 20 pkt. Analiza eksploracyjna posiadanego zbioru/wycinka zbioru danych i postawienie tezy/zadania badawczego majacego na celu opracowanie modelu regresji dla opisanych danych. Celem analizy eksploracyjnej danych jest sprawdzenie zależności miedzy posiadanymi danymi (5 pkt.), zbadanie ich zakresów i stopnia zmienności (5 pkt.), analiza stopnia wypełnienia danych (5 pkt.), wizualizacja (5 pkt.). Wynikiem tych analiz, ma być raport zakończony postawieniem hipotezy badawczej majacej na celu znalezienie relacji miedzy zmiennymi objaśniajacymi, a zmienna objaśniana.

1 Tematyka Projektu

Celem projektu było znalezienie właściwego zbioru danych do jego analizy i postawienia hipotezy. Dalszym celem tych działań jest budowa efektywnego modelu regresji przewidujacego wartość wybranej zmiennej opisywanej na podstawie wartości pozostałych cech.

2 Omówienie wybranego zbioru danych

2.1 Źródło

Wybrano zbiór https://data.world/data-society/student-alcohol-consumption. Strona data.world udostepnia wiele ciekawych zbiorów danych i stanowi dobre ich źródło.

2.2 Omówienie cech

Poniżej przedstawiono omówienie znaczenia cech w zbiorze oryginalnym. Na dalszym etapie cześć z nich nie bedzie potrzebna. Jak widać zbiór oryginalny zawiera 34 cechy z czego wiele jest kategorycznych. Poza tym nawet jeśli liczbowe to sa to wartości dyskretne z jakiegoś zbioru możliwych wartości.

1. student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. traveltime - home to school travel time (numeric: 1 - 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - 1 hour)
14. studytime - weekly study time (numeric: 1 - 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - 10 hours)
15. failures - number of past class failures (numeric: n if 1=n3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)
31. These grades are related with the course subject, Math or Portuguese:
32. G1 - first period grade (numeric: from 0 to 20)
33. G2 - second period grade (numeric: from 0 to 20)
34. G3 - final grade (numeric: from 0 to 20, output target)

2.3 Wyzwania zbioru

Wybraliśmy ten zbiór, ponieważ posiada kilka ciekawych cech stanowiacych ciekawe wyzwanie.

1. Duża liczba cech.
2. Zawiera cechy kategoryczne o wielu możliwych wartościach.
3. Zawiera cechy binarne.
4. W ogólności wszystkie cechy (również po przekształceniach) sa dyskretne, a nie ciagłe (liczbowe).
5. Zbiór nie jest zbalansowany wzgledem interesujacej nas docelowej cechy opisywanej.

3 Wstepne przygotowanie danych

Dane sa tylko połaczone w zbiór, który jest nam potrzebny, ale należy traktować je jako dane oryginalne, czyli niepoddane normalizacji, czy innym inwazyjnym technikom. Jedyne co zrobiono to usunieto/połaczono pewne kolumny, zamieniono atrybuty kategoryczne na ilościowe w celu umożliwienia dalszej analizy zbioru i uczenia modelu, zmieniono nazwy kolumn i posortowano w celu wiekszej przejrzystości w dalszych wizualizacjach, zamieniono typy danych na int64, ponieważ wszystkie wartości były całkowite (dane składaja sie z wielu atrybutów dobieranych przez człowieka w wyniku rozmowy z uczniem, wiec sa to pewne wskaźniki). Ponadto połaczono picie w tygodniu i weekend we wspólny wskaźnik problemu alkoholowego z wyższa waga dla picia w tygodniu.

3.1 Edycja zbioru oryginalnego

Wczytano oba pliki tabelek dla 2 przedmiotów obserwowanego kursu. Nastepnie połaczono je ignorujac index. Usunieto columns=["G1","G2","G3","paid"], ponieważ sa to cechy różne w obu tabelkach dla tego samego ucznia, a chcemy w nastepnym kroku usunać duplikatowych uczniów (w dokumentacji zbioru jest opisane, że uczniowie w obu tabelach sie powtarzaja, ale różnia kursami). Nastepnie łacze picie alkoholu w tygodniu z piciem w weekend jako sume ważona, gdzie picie w tygodniu ma wage 2.5, a w weekend wage 1. Po operacjach połaczenia picia oraz usuniecia duplikatów usuwam reszte niepotrzebnych cech columns=[ścholl","Dalc","Walc"]. Gdzie school oznacza szkołe, do której chodzi uczeń, co dla nas nie ma znaczenia. Ponieważ chcemy, aby nasz model zwracał ryzyko alkoholowe dla każdego ucznia (nie tylko chodzacego do jednej z tych kilku szkół ze zbioru).

3.2 Binaryzacja cech kategorycznych

Cześć cech było łatwo zbinaryzować ['sex','address','famsize','Pstatus'], ponieważ należało jedynie zmapować nazwy typu tak/nie na 1/0, bo były one naturalnie binarne, a trzeba było jedynie zamienić je na numeryczne. Dla cech features_to_encode = ['Mjob','Fjob','reason','guardian'] należało użyć techniki one-hot-encode. Biblioteka sklearn posiada wiekszość narzedzi do budowy klasycznych modeli. Nic wiec dziwnego, że miała również to i bardzo dobrze opisane w dokumentacji. Technika ta zamienia każdy pojedynczy atrybut kategoryczny na zbiór atrybutów binarnych. Po jednym dla każdej mozliwej wartości cechy oryginalnej. Technika ta mocno poszerza dataframe, jednak jest o tyle lepsza od zwykłej zamiany kategorii w kolejne liczby, że minimalizuje zwiekszanie wpływu jednej wartości cechy nad inna.

3.3 Zmiany dla wygody

Nastepnie zamieniłem dla pewności wszystkie dane na int64 (po sprawdzeniu, że taki jest nasz zbiór). Zamieniłem również nazwy cech na z małej litery, aby nastepnie posortować kolumny dataframe alfabetycznie, bo ułatwia dalsza analize, np. w przypadku analizy macierzy korelacji łatwiej szybko znaleźć ceche, której sie szuka.

4 Braki w danych

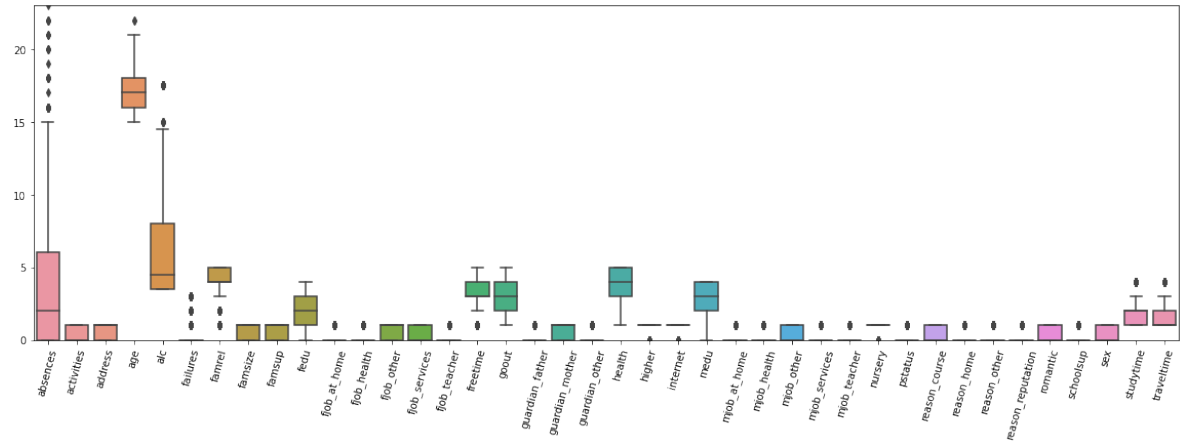
W zbiorze nigdy nie było i również po prostych edycjach nie ma braków danych.

5 Wizualizacja danych

Dane, które wizualizuje pochodza juz ze zbioru nieco przekształconego do naszych potrzeb jak opisano powyżej. Nie poddano ich jednak żadnym inwazyjnym metodom.

5.1 Wykresy pudełkowe

Najwiecej punktów oddalonych ma cecha nieobecności i trzeba było je przyciać, aby nie popsuć widoczności wykresu. Inna opcja była zamiana skali na log ale uważam, że w tym przypadku nie miałoby



Rysunek 1: Wykresy pudełkowe cech

to sensu, bo cieżko jest odczytać dokładne zakresy zmienności cech, które sa dość małe. Naturalnie wiele atrybutów binarnych (powstałych z kategorycznych) ma zakres zmienności 0,1 bo przyjmuje tylko te 2 wartości. Niezdania maja sporo punktów oddalonych, co ma sens, bo to jedna z tych rzeczy, które na poziomie liceum dotykaja tylko nielicznych uczniów. Tak samo relacje rodzinne. Co ciekawe czas wolny ma punkt oddalony. Prawdopodobnie uczeń z duża ilościa zajeć dodatkowych. Dodatkowo warto wspomnieć, że skoro dane były pobierane przez nauczyciela od ucznia na podstawie rozmowy to na pewno sa mocno zaszumione przez subiektywne odczucia ucznia odnośnie np. poziomu zdrowia czy relacji z rodzina. Ludzie maja tendencje do narzekania w tych sprawach co może niesłusznie pogarszać wyniki itp. Jak widać najwiecej punktów oddalonych maja cechy takie jak:

1. nieobecności - wiekszość uczniów prawie nie ma nieobecności,
2. niezadania do nastepnej klasy - wiekszości uczniów nigdy nie zdarzyło sie niezdać do nastepnej klasy,
3. relacje rodzinne - wiekszość uczniów ma dobre relace rodzinne.

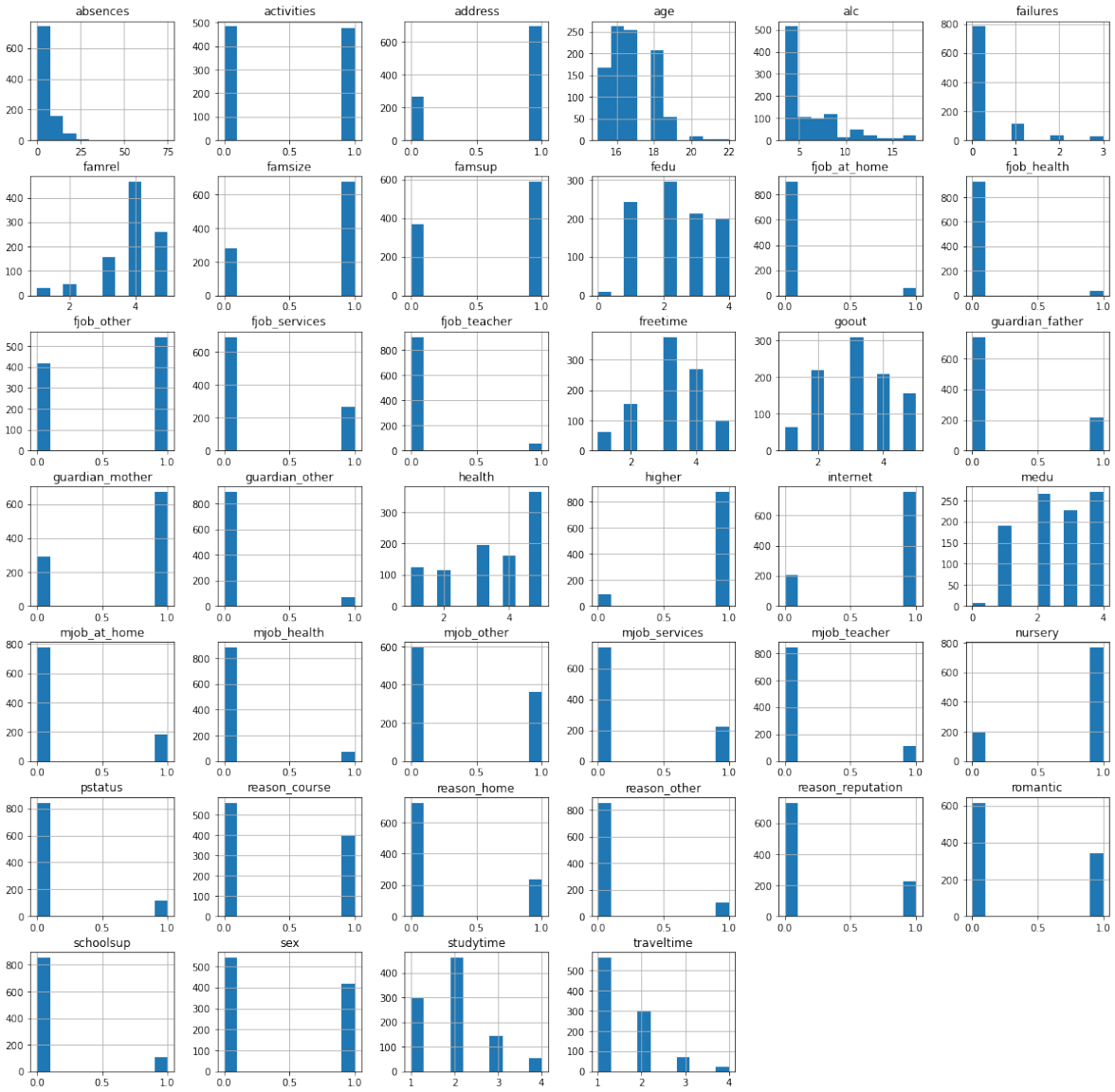
Możliwe, że odchylenia od normy w tak ważnych dla rozwoju dziecka metrykach beda mieć spory wpływ na ryzyko alkoholowe.

5.2 Histogramy cech

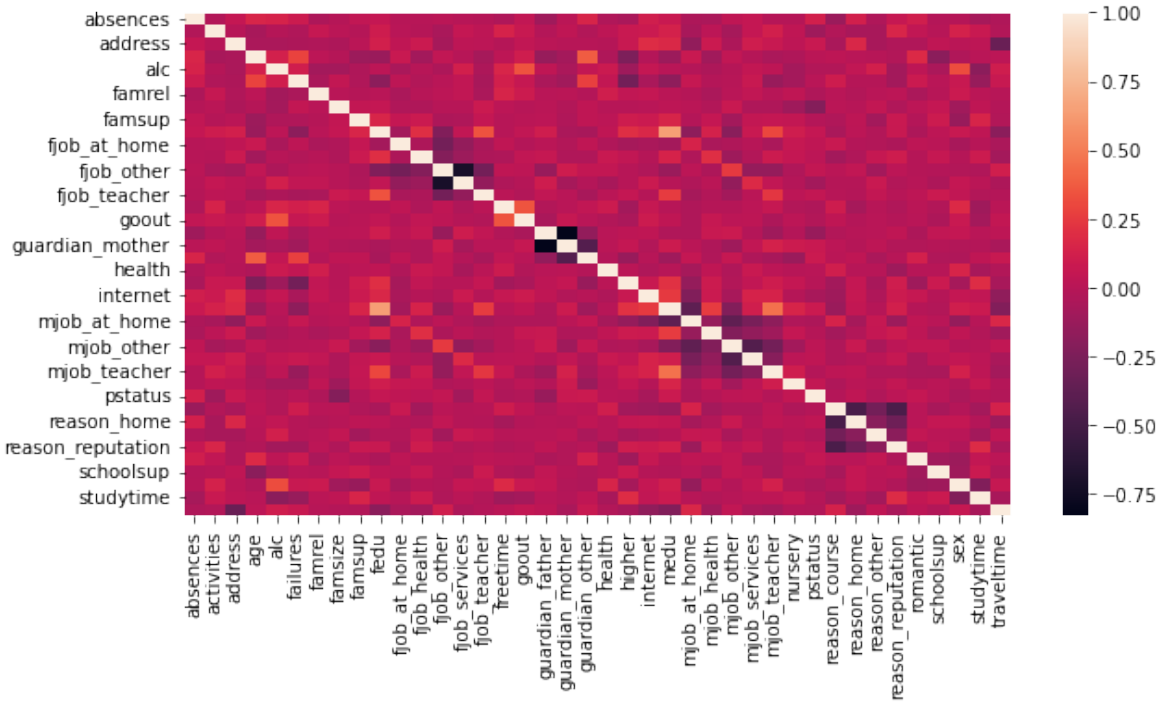
Poniżej widać rozkłady wszystkich cech. Jak widać wiele z nich jest binarnych lub ogólniej dyskretnych. Cechy nie sa ciagłe, co może stanowić wyzwanie trenowania modelu, co nas zacheciło do wyboru tego zbioru.

6 Zbalansowanie zbioru

Można powiedzieć, że zbiór jest niezbalansowany wzgledem naszej zmiennej opisywanej (odpowiednik klasy w klasyfikacji). Wynik maksymalny to 17.5. Połowa osób ma wynik mniejszy od 4.5. Ćwierć osób ma wynik 4.5,8 . Tylko pozostałe ćwierć osób ma wynik wyższy niż 8. Jest to jednak zrozumiałe i na całe szczeście tylko nieliczni maja problem w badanej próbie. Jak widać mamy około tysiaca obiektów w zbiorze. Zbiór naturlanie nie jest zbalsowany pod katem zmiennej opisywanej ąlc". Przyjmuje ona wartości z zakresu 3.5,17.5 . Wartości rosna wraz ze stopniem alkoholizmu. Tylko 25 procent uczniów pije wiecej niż około połowa tego zakresu, czyli średni stopień problemu. Jeden uczeń (punkt oddalony) wyznacza górny limit skali stanowiac dla nas naturalny punkt odniesienia (skala empiryczna). Poniżej widać opis zmiennej opisywanej.



Rysunek 2: Histogramy cech



Rysunek 3: Macierz korelacji

Cecha opisywana	
mean	6.094369
std	3.308938
min	3.500000
25 procent	3.500000
50 procent	4.500000
75 procent	8.000000
max	17.500000

7 Stopień korelacji cech

7.1 Opis najbardziej znaczacych korelacji

1. opieka ojca albo matki w rozwiedzionych parach
2. edukacja matki i ojca
3. praca nauczyciela i edukacja dla matek i ojców ale w mniejszym stopniu
4. praca matki w domu, a wykształcenie
5. tak zwany inny opiekun dla uczniów, których wiek jest nietypowo wysoki w szkole średniej (dom dziecka, etc.)
6. wyjścia z domu, a alkohol i czas wolny
7. czas podróży tym wyższy jeśli mieszkaja poza miastem
8. płeć meska i alkohol
9. wybór szkoły ze wzgledu na reputacje kontra ze wzgledu na wygode czyli bliskość od domu

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Rysunek 4: Interpretacja stopnia korelacji

7.2 Dokładne wartości

1. guardian_mother guardian_father -0.830369
2. fjob_services fjob_other -0.711511
3. medu fedu 0.637790
4. reason_home reason_course -0.477166
5. reason_reputation reason_course -0.467695
6. mjob_teacher medu 0.450380
7. mjob_services mjob_other -0.430532
8. guardian_other guardian_mother -0.425006
9. mjob_at_home medu -0.380988
10. mjob_other mjob_at_home -0.379828
11. guardian_other age 0.376413
12. goout freetime 0.342537
13. fjob_teacher fedu 0.341598
14. goout alc 0.336565
15. traveltime address -0.335628
16. sex alc 0.320500
17. reason_reputation reason_home -0.314566

8 Teza badawcza

Teza naturalnie pojawiała sie już w powyższej cześci jednak powtórze ja raz jeszcze. Celem bedzie przewidywanie cechy ąlcśtanowiacej sume ważona picia w tygodniu i weekedny na podstawie cech ucznia. Cechy to można traktować jako zmienne środowiskowe i psychologiczne. Mamy nadzieje, że model bedzie na tyle dokładny, że posłuży jako narzedzie dla psychologów do automatycznego obliczenia stopnia ryzyka alkoholowego ucznia na podstawie badanych czynników pobranych w formie rozmowy/wywiadu z psychologiem szkolnym.