Oceny szkolne - przewidywanie oceny uczniów na podstawie atrybutów stylu życia

Raport finalny

Adrianna Grudzień, Kinga Ułasik

20 kwietnia 2021 r.

Spis treści

1 Wprowadzenie

W ramach projektu z przedmiotu *Wstęp do uczenia maszynowego* zbudowałyśmy model przewidujący ocenę końcową z matematyki/języka portugalskiego w jednej z portugalskich szkół. Niniejszy raport dokumentuje naszą pracę z tym związaną.

Opis zbioru danych Dane dotyczą osiągnięć uczniów w szkołach średnich dwóch portugalskich szkół. Atrybuty danych obejmują oceny uczniów, cechy demograficzne, społeczne i związane ze szkołą i zostały zebrane przy użyciu raportów szkolnych i kwestionariuszy. Dostarczono dwa zbiory danych dotyczące wyników z matematyki lub języka portugalskiego.

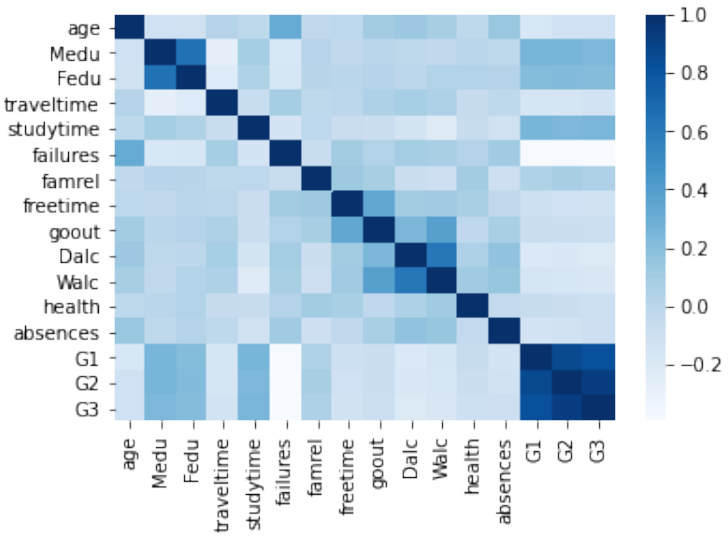
Zbiór danych został zamodelowany w ramach binarnych / pięciopoziomowych zadań klasyfikacji i regresji. Ważna uwaga: atrybut docelowy G3 ma silną korelację z atrybutami G2 i G1. Dzieje się tak, ponieważ G3 jest ostatnią klasą roku (wystawianą na III klasie), podczas gdy G1 i G2 odpowiadają klasom z I i II stopnia. Trudniej jest przewidzieć G3 bez G2 i G1, ale takie przewidywanie jest znacznie bardziej przydatne.

2 Eksploracja danych (EDA)

Statystyki danych Naszą analizę zaczęłyśmy od sprawdzenia statystyk naszych danych. Przykładowo, okazało się, że:

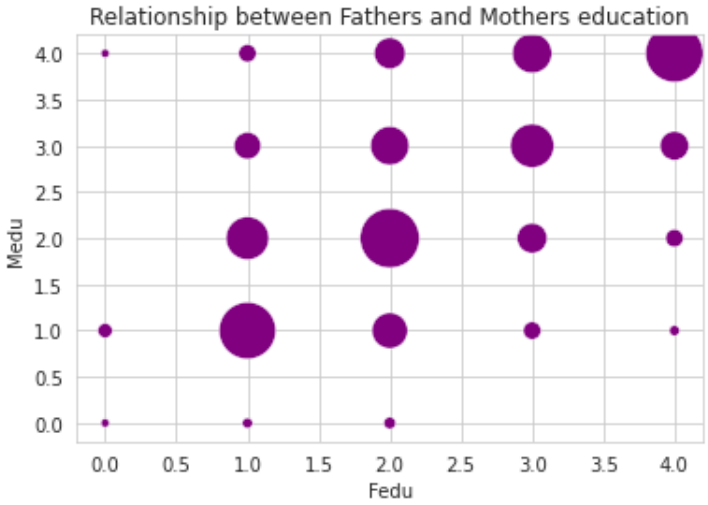
- średni czas dotarcia do szkoły wynosi ok. 22 min
- tygodniowo na naukę uczniowie poświęcają średnio ok. 3,5h

Zauważyłyśmy również, że wszyscy uczniowie, którzy otrzymali 0 na egzaminie końcowym, na pewno podeszli do egzaminu w pierwszym semestrze, ale część z nich otrzymało wynik zero w egzaminie w drugim semestrze (możliwe, że zrezygnowali po pierwszym).
Korelacja zmiennych

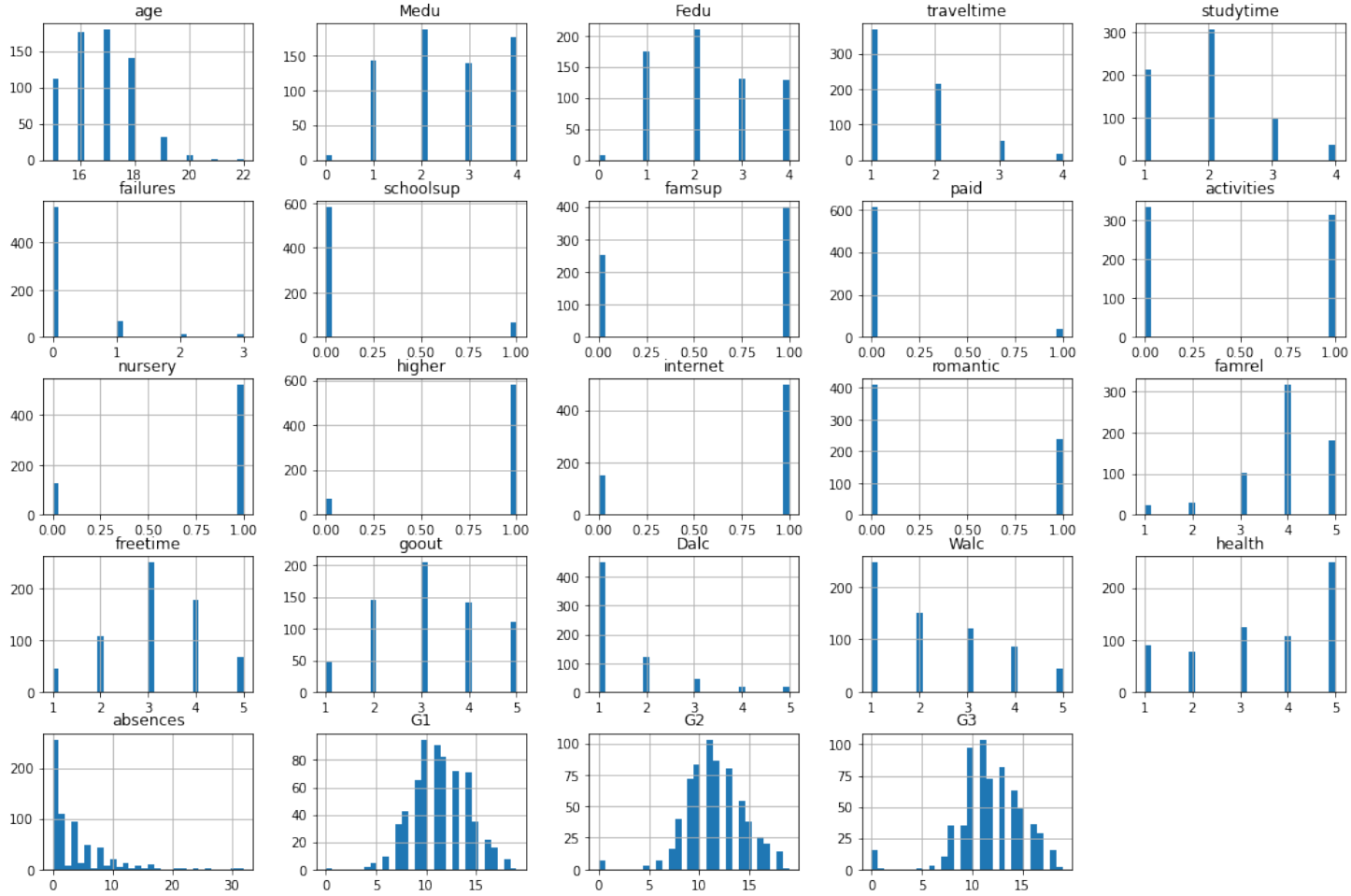


Można zauważyć silną korelację pomiędzy G1, G2, G3, co jest logiczne, biorąc pod uwagę, że są to kolejne wyniki z semestru (G1 i G2 silnie wpływają na G3).

Widać również korelację pomiędzy Fedu i Medu (czyli edukacją rodziców); między Walc i Dalc (spożycie alkoholu kolejno w weekend, dzień roboczy) oraz między goout i freetime oraz failures a wynikami końcowymi. Przyjrzyjmy się, przykładowo, zależności Fedu i Medu.



Widać, że największe „bąbelki" znajdują się na przekątnej. Można zaryzykować stwierdzenie, że ludzie najczęsciej dobierają się w małżeństwa z osobą z podobnym poziomem edukacji.
Mapowanie W celu wykorzystania w miarę możliwości wszystkich dostępnych danych, zmapowałyśmy te kolumny, które nie były numeryczne.
Histogramy Przyjrzałyśmy się rozkładom poszczególnych zmiennych.



Zgodnie z przewidywaniami, zmienne G1, G2, G3 mają rozkład w przybliżeniu normalny, podobnie zachowuje się goout i freetime. Z kolei zmienna absences oraz Dalc - rozkład wykładniczy.

3 Inżynieria cech, wstępne modelowanie

Mapowanie i kodowanie W celu doprowadzenia danych do lepszej używalności, oprócz mapowania na poziomie EDA, przeprowadziłyśmy również kodowanie zmiennych kategorycznych.

Wstępne modelowanie W naszym modelu korzystamy z regresji liniowej. W celu lepszego wglądu w efektywność modelu policzyłyśmy błąd RMSE oraz r2. Najniższy RMSE, jaki byłyśmy w stanie otrzymać wynosił około **2.178**.

Normalizacja Po przyjrzeniu się wartościom zmiennej age oraz jej rozkładowi, doszłyśmy do wniosku, że warto by było ją znormalizować - przedział wiekowy był wystarczająco szeroki. Wybrałyśmy normalizację logarytmiczną i rzeczywiście - wyniki modelu nieco się poprawiły. Oto wynik w jednej z prób dla modelu uwzględniającego G1 i G2:

PRZED normalizacją:
Root Mean Squared Error: 0.9732349517944905
R-squared: 0.8666668222615632

PO normalizacji:
Root Mean Squared Error: 0.9727333142228294
R-squared: 0.8668042355224069

Zadanie klasyfikacji Dodatkowo przygotowałyśmy model, który zadanie przewidywania wyniku końcowego uczniów traktuje jako zadanie z klasyfikacji, zamiast z regresji. Mamy więc 20 klas (ilość punktów na egzaminie) i sprawdzamy, czy taki model daje sobie radę lepiej niż model regresji. Obliczony RMSE wynosił:

Zadanie klasyfikacji:
Root Mean Squared Error: 1.7715703594442809

Jak widać, RMSE, w porównaniu do rozwiązywania problemu za pomocą regresji, wychodzi dużo większy.

4 Finalny model

Końcowo zdecydowałyśmy się na model zbudowany za pomocą narzędzia *LinearRegressior* należącego do pakietu *sklearn*. Najlepsze wyniki byliśmy w stanie uzyskać, wybierając jako parametry zmienne:

- G2 - 2nd period grade
- failures - failures - number of past class failures
- higher - wants to take higher education (binary: 1 or 0)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- school - student's school (binary: 1 - Gabriel Pereira or 0 - Mousinho da Silveira)
- age - student's age (numeric: from 15 to 22) [normalized]
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

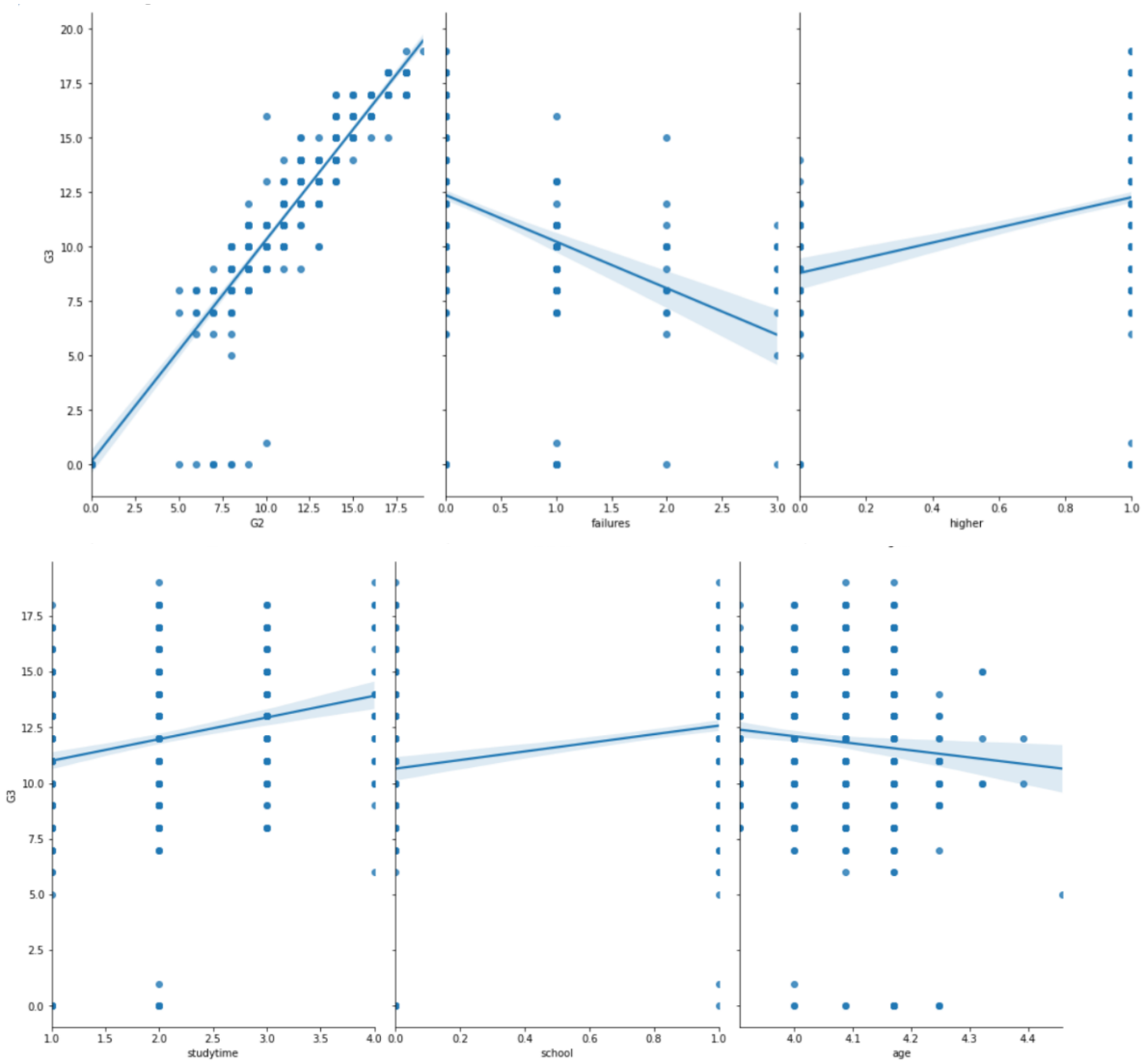
Dla porównania, automat (Feature_Selection) uznał za ważne następnujące zmienne: ['school', 'Medu', 'Fedu', 'failures', 'Dalc', 'Walc', 'absences', 'G1', 'G2', 'Mjob_teacher', 'Fjob_teacher', 'other']

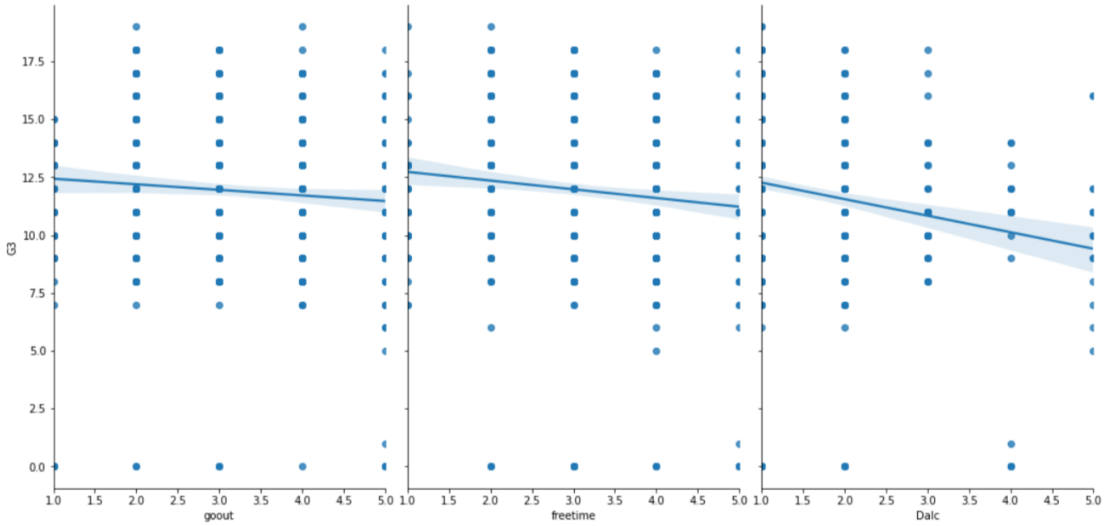
Jak widać, nie różnią się one zbytnio od tych wybranych przez nas.
Wyniki, jakie otrzymałyśmy to:

Dla zbioru testowego:
Root Mean Squared Error: 0.9809479465281395
R-squared: 0.8645450875659495

Dla kroswalidacji:
Root Mean Squared Error: 0.9135234415045292
R-squared: 0.7884095196388906

Dodatkowo, aby łatwiej było wyobrazić sobie zależności występujące w naszych danych, przyjrzałyśmy się bliżej relacjom między wybranymi przez nas zmiennymi a zmienną targetową:





Na koniec porównałyśmy stworzony przez nas model do modelu wygenerowanego automatycznie (za pomocą pakietu TPOT). Można zaobserwować, że wartość R2 squared jest praktycznie identyczna, jednak w modelu stworzonym przez nas RMSE jest mniejsze. Taki wynik jednocześnie nas cieszy, ale też trochę martwi (czy już jesteśmy zastąpieni?)

Nasz model:
Root Mean Squared Error: 0.9622588737977708
R-squared: 0.8645450875659498

Model TPOT:
Root Mean Squared Error: 0.9822666024095801
R-squared: 0.8641806676739481