POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Wydział Informatyki

Inteligentne Systemy Przetwarzania Danych

Weronika Skoczylas
S14805

„Analiza i system predykcyjny osiągnięć edukacyjnych w kontekście społeczno-ekonomicznym ”

praca inżynierska
Dr inż. Adam Szmigielski

Warszawa, luty, 2020

STRESZCZENIE

Praca w pierwszej kolejności zajmuje się opisem danych i ich przekształceniem do odpowiedniej formy – mapowanie i skalowanie wartości. Dalsza część dotyczy analizy danych, składa się z takich elementów jak rozkład danych, przeanalizowanie poprawności rekordów, analiza korelacji atrybutów i analiza koszykowa. Na każdym etapie analizy zostały wysunięte wnioski i gdzie było to możliwe, postawione były kolejne kroki ku głębszej analizie. Po przeanalizowaniu danych zajęliśmy się predykcją przy użyciu drzewa decyzyjnego i kernel SVM. Następnie model został poddany ewaluacji pod względem dokładności i skuteczności. Na końcu przedstawiony został system predykcyjny, który na bazie zmian atrybutów i modelu predykcyjnego wystawia oceny oraz alternatywne podejście do problemu.

Słowa kluczowe: analiza, Drzewo decyzyjne, Kernel SVM, model, predykcja

Spis treści

WSTĘP

O pracy

Tematem pracy jest analiza danych i predykcja jednego z atrybutów, zwrócimy uwagę na metody przetwarzania danych, szukaniu powiązań w danych oraz wizualizacji danych. Na tej podstawie będziemy starać się wysunąć wnioski a tam gdzie się da, znaleźć potwierdzenie naszych hipotez. Następnie zajmiemy się utworzeniem modelu predykcyjnego oraz próbą wykorzystania go w codziennym życiu.

O danych

Dane określają wyniki edukacyjne uczniów szkół licealnych dwóch portugalskich liceów. Atrybuty zawierają oceny uczniów, ich styl życia oraz parametry o podłożu socjalnym i społecznym. Dane były zebrane przy użyciu szkolnych raportów i kwestionariuszy z 2008 roku. Zebrane dane składają się z dwóch zbiorów pierwszy dotyczący matematyki (mat) i drugi dotyczący języka portugalskiego (por) [3]. Ze względu na fakt, iż dane pochodzą z Portugalii, portugalski zostanie potraktowany jako język ojczysty. Jego znajomość będzie wyznacznikiem zdolności humanistycznych, a matematyka ścisłych .

Cel i zakres pracy

Głównym celem tej pracy będzie próba poprawy wyników edukacyjnych uczniów poprzez podniesienie ich motywacji oraz wyszukanie czynników mających szczególnie duży wpływ na edukację. Zajmiemy się więc w pierwszym kroku analizą danych, sprawdzimy ich rozkład oraz powiązania atrybutów między sobą, dzięki czemu w przyszłości będzie możliwość poprawy wyników poprzez skupienie się na aspektach, które mają największy wpływ na edukację. Zajmiemy się również porównaniem między sobą zbiorów danych z różnych przedmiotów, tak by zaobserwować atrybuty, które wpływają z taką samą siłą na wyniki z obu przedmiotów, jak i te, które są mocniej powiązane z naukami ścisłymi lub humanistycznymi. Następnie zajmiemy się oszacowaniem oceny z ostatniego roku na podstawie reszty atrybutów. Oprócz predykcji oceny końcowej stworzymy również symulator, który będzie przedstawiał uczniom predykcję ich ocen, gdyby zmienili któreś ze swoich zachowań (zmiana wartości atrybutów na które mamy wpływ) co miałoby szanse na podniesienie ich motywacji. System ten będzie musiał być zaopatrzony w interfejs, by był intuicyjny w użyciu dla każdego ucznia.

Opis atrybutów

Atrybuty binarne

school – szkoła, do której uczęszcza uczeń ("GP" - Gabriel Pereira lub "MS" - Mousinho da Silveira)
sex – płeć ucznia ("F" – kobieta lub "M" - mężczyzna)
address – miejsce zamieszkania ucznia ("U" – miasto lub "R" - wieś)
famsize – rozmiar rodziny ("LE3" – mniejsza lub równa 3 lub "GT3" – większa niż 3)
Pstatus – Jak mieszkają rodzice ucznia(„T" – razem lub "A" - osobno)
schoolsup – dodatkowe wsparcie naukowe (tak lub nie)
famsup – dodatkowe wsparcie dla rodziny (tak lub nie)
paid – dodatkowe płatne zajęcia z (Matematyki lub Portugalskiego) (tak lub nie)
activities – dodatkowe zajęcia (tak lub nie)
nursery – czy uczeń uczęszczał do przedszkola (tak lub nie)
higher – czy uczeń chce mieć wyższe wykształcenie (tak lub nie)
internet – dostęp do Internetu w szkole (tak lub nie)
romantic – czy uczeń jest w związku (tak lub nie)

Atrybuty numeryczne

age – wiek ucznia (od 15 do 22)
Medu – edukacja matki (0 - brak, 1 – edukacja do 4 klasy, 2 – edukacja między 5 a 9 klasą, 3 – wykształcenie średnie lub 4 – wykształcenie wyższe)
Fedu - edukacja ojca (0 - brak, 1 – edukacja do 4 klasy, 2 – edukacja między 5 a 9 klasą, 3 – wykształcenie średnie lub 4 – wykształcenie wyższe)
traveltime – czas podróży z domu do szkoły (1 - <15 min, 2 - 15 do 30 min, 3 - 30 min do godziny, lub 4 – więcej niż godzina)
studytime – tygodniowy czas poświęcony na naukę (1 – mniej niż 2 godziny, 2 - 2 do 5 godzin, 3 - 5 do 10 godzin, lub 4 – ponad 10 godzin)
failures – ilość wcześniej oblanych klas (1-1,2-2,3-3, 4-więcej niż 3)
famrel – jakość relacji w rodzinie (od 1 – bardzo zła do 5 – bardzo dobra)
freetime – czas wolny po szkole (od 1 – bardzo mało do 5 – bardzo dużo)
goout – wychodzenie z kolegami / koleżankami (od 1 – bardzo rzadko do 5 – bardzo często)
Dalc – spożycie alkoholu w dniu pracującym (od 1 – bardzo mało do 5 – bardzo dużo)
Walc – spożycie alkoholu w weekend (od 1 – bardzo mało do 5 – bardzo dużo)
health – obecny stan zdrowia (od 1 – bardzo zły do 5 – bardzo dobry)
absences – ilość nieobecności (od 0 do 93)
G1 – ocena za pierwszy rok (od 0 do 20)
G2 - ocena za drugi rok (od 0 do 20)
G3 - ocena końcowa(od 0 do 20, wartość do predykcji)

Atrybuty nominalne

Mjob – praca matki ("teacher"-nauczyciel, "health" – praca związana ze służbą zdrowia, "services" – cywilne stanowisko np. Policja, "at_home" – niepracująca lub "other"-inne)
Fjob - praca ojca ("teacher"-nauczyciel, "health" – praca związana ze służbą zdrowia, "services" – cywilne stanowisko np. Policja, "at_home" – niepracująca lub "other"-inne)
reason – przyczyna wyboru szkoły ("home" – blisko domu, "reputation" – reputacja szkoły , "course" – preferencyjny kurs lub "other"-inne)
guardian – opiekun ucznia ("mother"-matka, "father" –ojciec lub "other" - inny)

ANALIZA DANYCH

Dane dostarczane są w dwóch zbiorach danych student-mat.csv (zawierający dane dotyczące matematyki) oraz student-por.csv (zawierający dane dotyczące portugalskiego) studenci portugalskiego i matematyki pokrywają się tylko częściowo. Dlatego też będziemy posługiwać się czterema zbiorami danych:

1. danePor - wszystkie rekordy ze zbioru tudent-por.csv
2. daneMat- wszystkie rekordy ze zbioru tudent-mat.csv
3. daneSuma - rekordy występujące w danePor, daneMat lub obu
4. danePrzeciecie- rekordy występujące w danePor i w daneMat

Pojawiają się jednak pewne problemy przy zbiorze daneSuma, gdyż będzie nam brakowało części danych, stąd pomocniczo wprowadzimy nowe parametry binarne „mat" i „por", które mówią nam, czy dany student uczy się danego przedmiotu, czy nie (1-tak,0-nie). Połączenie obu zbiorów prowadzi również do potrzeby lepszego opisania atrybutów unikalnych dla danego przedmiotu i tu dla portugalskiego został dodany suffix 'por' a dla matematyki 'mat' przez co otrzymaliśmy parametry: failures_por , absences_por , G1_por , G2_por , G3_por , failures_mat , absences_mat , G1_mat , G2_mat , G3_mat. W pierwszym kroku sprawdzimy średnie wartości, rozkład wartości dla poszczególnych atrybutów i czy występują braki w danych. Po utworzeniu naszych czterech zbiorów przeprowadzona została wstępną analiza zbioru daneSuma przy użyciu funkcji pandas. DataFrame.describe, z której możemy zorientować się co do rozkładu atrybutów poznać ich minimalne i maksymalne wartości oraz kwartyle. Możemy też zliczyć liczbę wartości dla każdego atrybutu, dzięki czemu wiemy, że nie brakuje żadnych wartości w danych źródłowych, oraz wywnioskować wielkość zbiorów.
674 – daneSuma
395- daneMat
649 – danePor
369 - danePrzeciecie

Mapowanie danych

Mapowanie danych jest niezbędnym krokiem, by móc dokonywać obliczeń czy eksperymentować z różnymi modelami predykcyjnymi

Atrybuty binarne

Z racji tylko dwóch możliwości możemy łatwo je ujednolicić i zamienić na numeryczne (1 , 0) . Postaramy się, by mapowanie to było możliwie jak najbardziej intuicyjne. Mamy mniejszą szansę na popełnienie błędów później, jeśli mapujemy 1 na tak a 0 na nie niż odwrotnie.

Atrybuty numeryczne

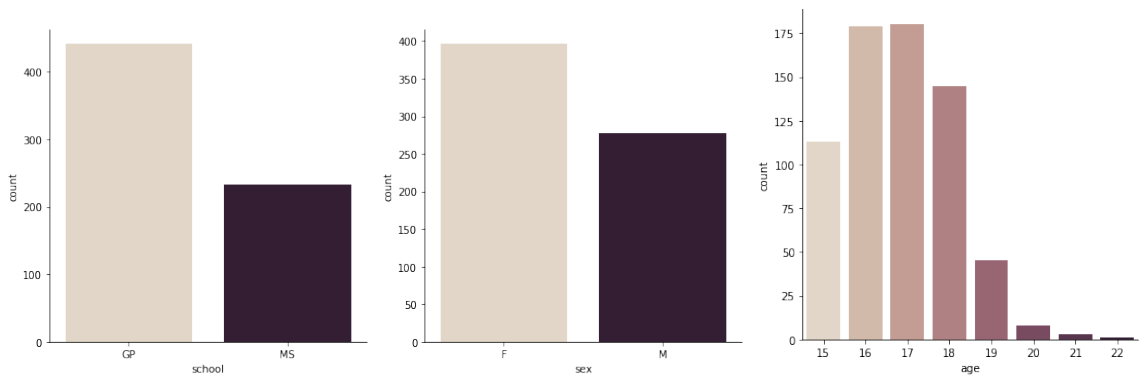
W naszym przypadku nie ma potrzeby przekształcania atrybutów numerycznych, aczkolwiek mogłaby się ona pojawić przy ciągłych danych o dużej rozbieżności. W takich wypadkach warto rozpatrzeć utworzenie kubełków dla różnych zakresów wartości

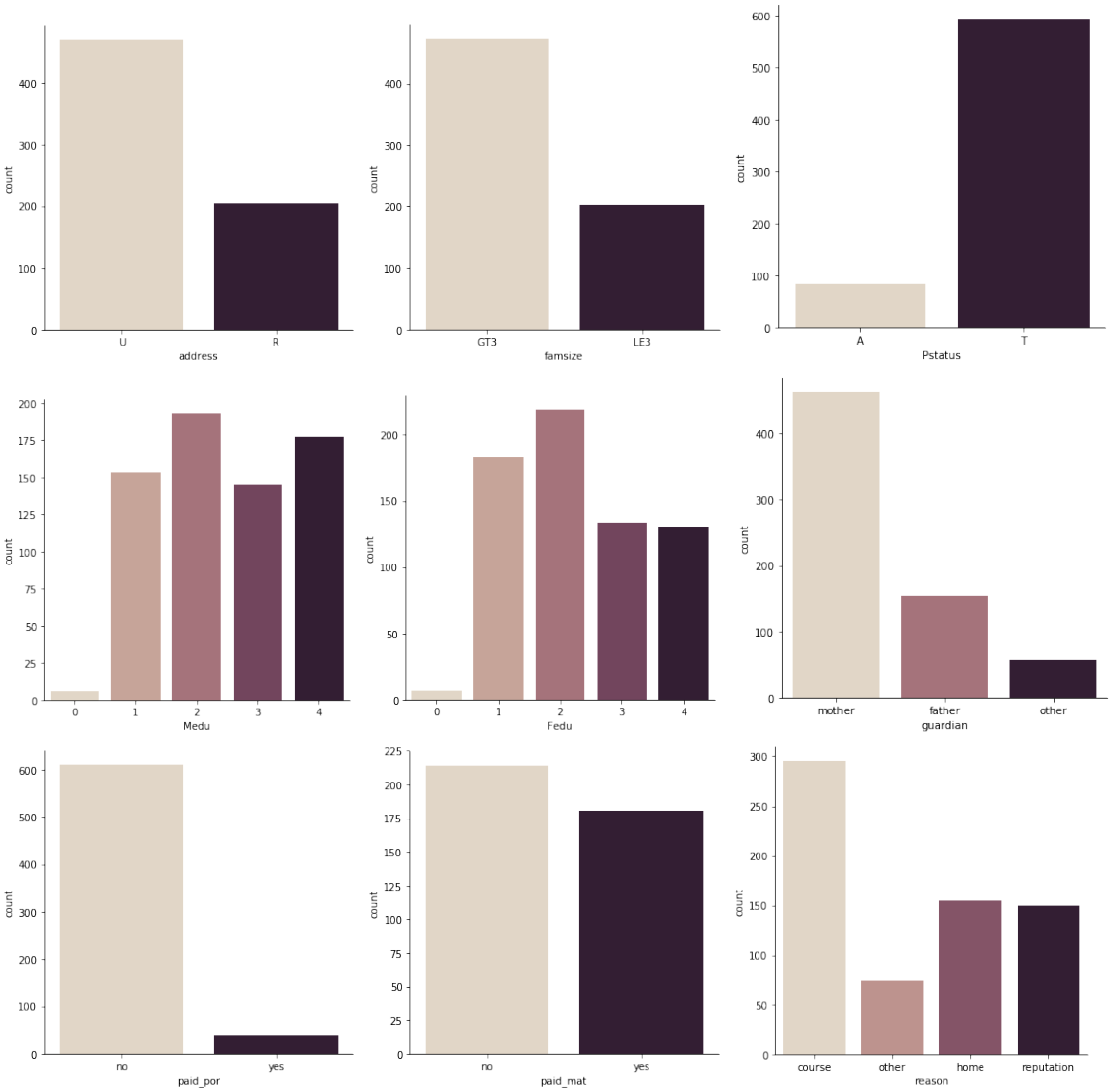
Atrybuty nominalne

Problem pojawia się przy atrybutach nominalnych, gdyż nie jesteśmy w stanie ich stopniować i tym bardziej opisać zależności między nimi [2]. Na przykład dla parametru ‘Mjob’ nie jesteśmy w stanie ułożyć w kolejności istotności bycia nauczycielem, pracy w sektorze zdrowia czy tym bardziej pracy określanej jako inna. Ponieważ nie ma bardzo dużych zakresów wartości dla atrybutów nominalnych, każdy z nich zamienimy na tyle atrybutów, ile miał on możliwych wartości. Nowe atrybuty są atrybutami binarnymi dla każdej wartości określają „1”, jeśli występowała i „0”, jeśli nie i tak na przykład z atrybutu guardian = „mother” powstaną guardian_mother=1 , guardian_father = 0 , guardian_other = 0

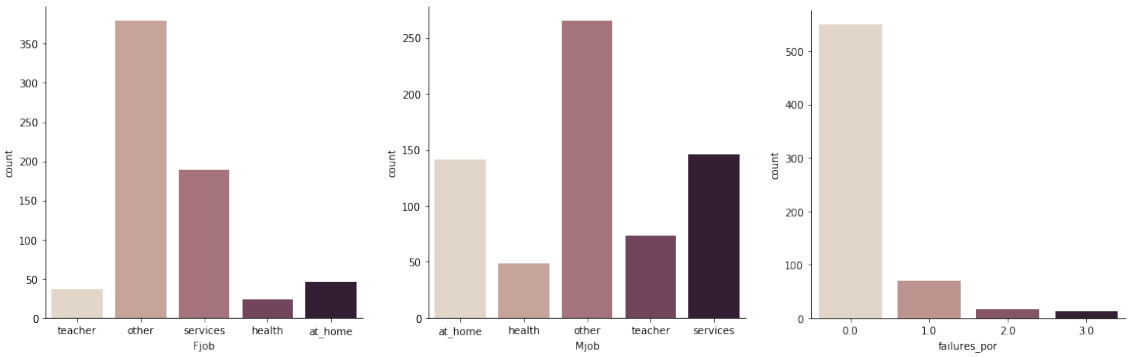
Statystyczna analiza danych

Przed przystąpieniem do jakichkolwiek działań warto przyjrzeć się danym. Aby nie pracować na ślepo, zaczniemy od rozkładu danych, który został przedstawiony poniżej na zbiorze daneSuma.

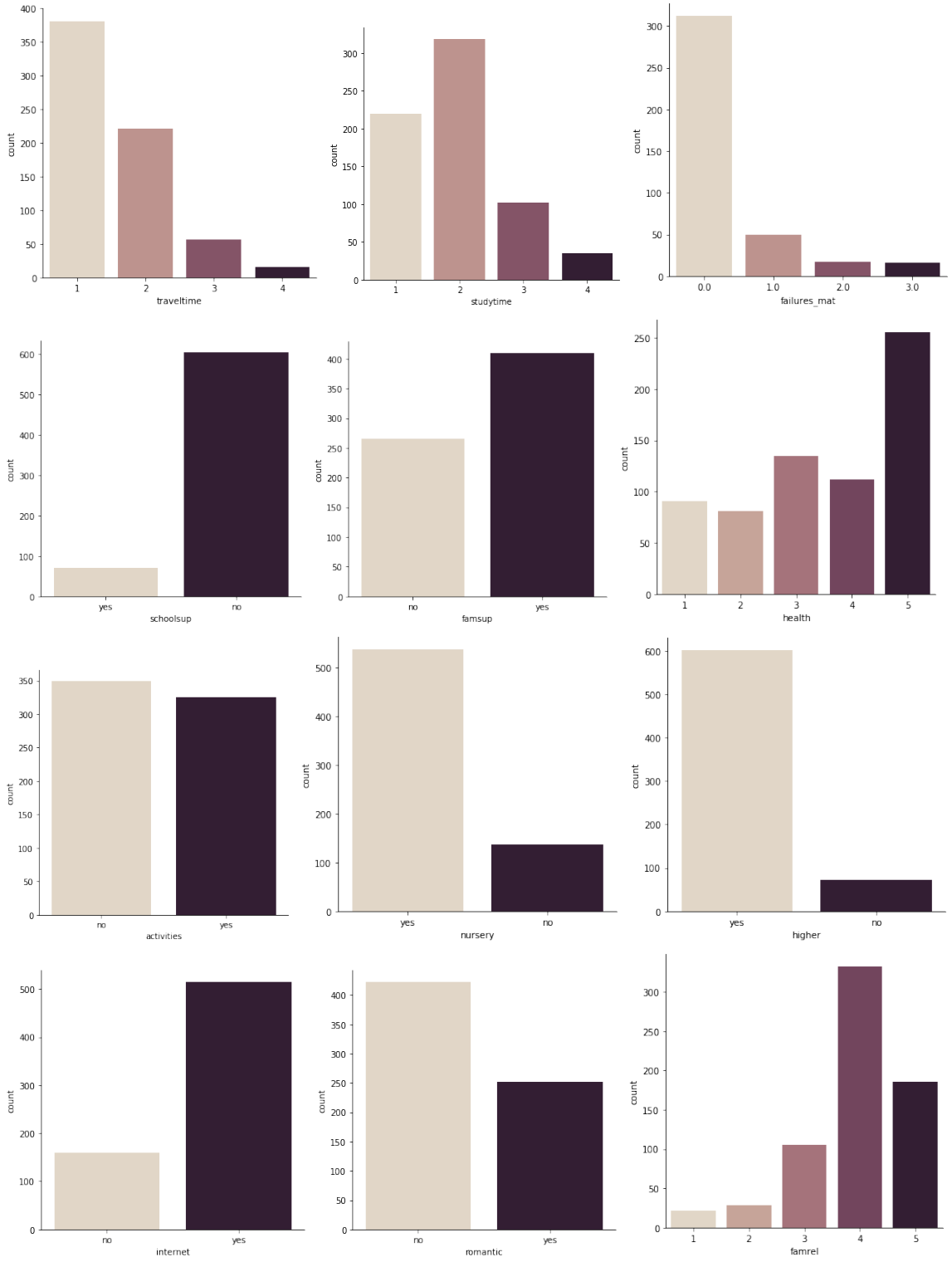


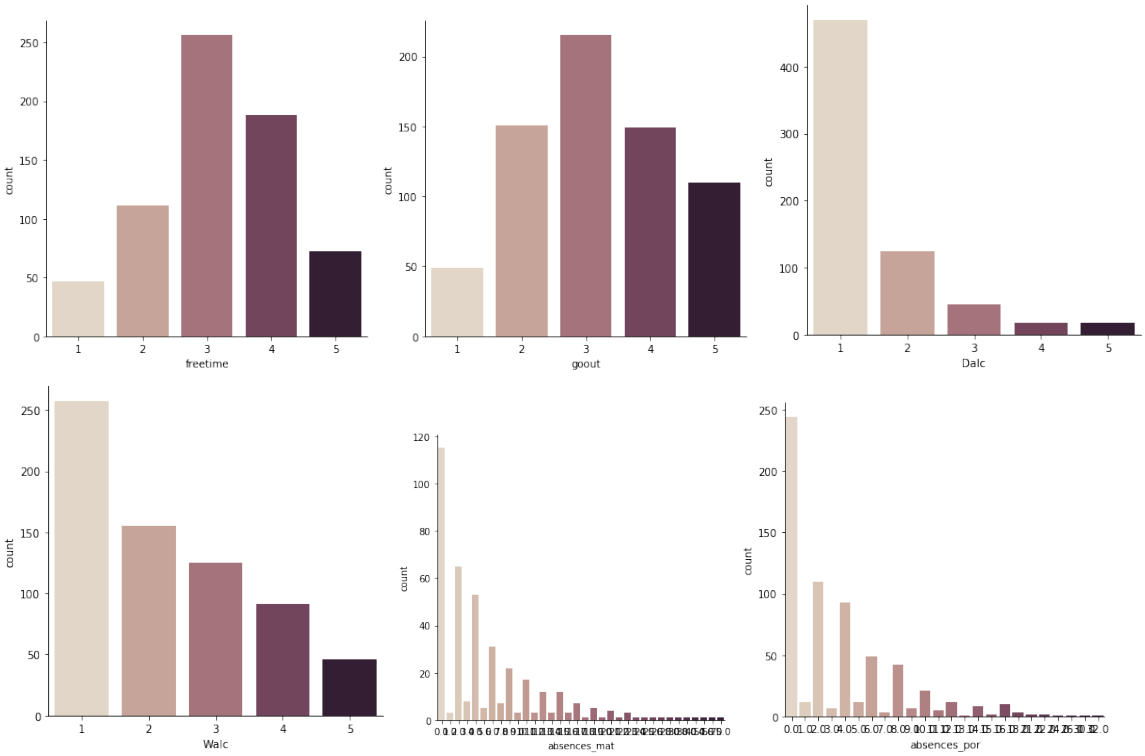


Z ciekawszych obserwacji widzimy na przykład, że dla zajęć dodatkowych z matematyki rozkład jest niemal równy, czyli prawie połowa uczniów potrzebuje zajęć dodatkowych, natomiast przy rozkładzie zajęć dodatkowych z portugalskiego widzimy, że zdecydowana większość nie uczęszcza na takie zajęcia. Dla szkół ciekawą obserwacją może być, iż przy doborze szkoły większość osób zwraca uwagę głównie na oferowane przez nią kursy, a nie reputację szkoły.



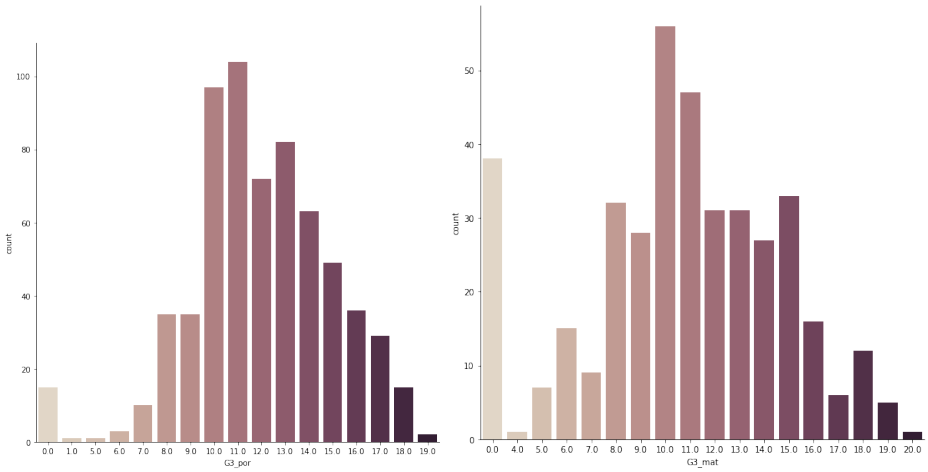
Dalej możemy już przypuszczać, że parametry 'Fjob' i 'Mjob' mogą okazać się niezbyt przydatne. Ponieważ w obu przypadkach największa liczba wartości przypada dla kategorii ‘other’ która nie mówi nam praktycznie nic, bo nie wiemy, czy będzie to stanowisko dużo lepsze niż pozostałe kategorie, czy może dużo gorsze jedyne co nam mówi to to, że jest to osoba pracująca.





Rysunek 1. Histogramy atrybutów

Po atrybutach 'freetime', 'goout', 'Dalc', 'Walc' widzimy, że są dosyć przewidywalne, spożycie alkoholu spada wraz z licznością grupy osób, a większe spożycie przypada na weekend. Rozkład czasu wolnego i wychodzenia ze znajomymi mniej więcej się pokrywają i są równomiernie rozłożone. Podczas analizy danych możemy też doszukać się informacji, które nie były wcześniej podane, na przykład, jeśli przyjrzymy się histogramom dotyczącym nieobecności (parametry 'absences_por' i 'absences_mat') widzimy zależność, w której zdecydowanie więcej osób ma parzystą liczbą nieobecności niż nieparzystą. Szanse, że wydarzyło się to przypadkiem, są praktycznie niemożliwe, możemy więc wnioskować, że na przykład zajęcia odbywają się w blokach po dwie godziny lub inaczej liczone były nieobecności usprawiedliwione i nieusprawiedliwione.



Rysunek 2. Histogramy dla ocen z portugalskiego i matematyki

Rozkłady dla ocen końcowych nie są szczególnie zaskakujące widzimy bowiem wyśrodkowanie, aczkolwiek wygląda na to, że mamy zdecydowanie lepsze oceny z Portugalskiego niż Matematyki.

Skalowanie danych

Skalowanie danych służy do wyrównania wartości pomiędzy różnymi atrybutami na etapie, na którym nie wiemy, jak duży wpływ mają atrybuty na przewidywaną wartość. Zakładając więc, że mogą być one tak samo istotne, mógłby pojawić się problem w przypadku, gdy atrybuty mają różne skale. W naszym modelu mógłby być to wiek i parametr określający płeć, czyli zakresy (15-22) i parametr binarny, czyli zakres (0-1) bez przeprowadzenia skalowania przy późniejszych wyliczeniach mogą mieć znacząco różne wpływy na wynik. Jeśli chodzi o podstawowe typy skalowania danych, rozpatrzymy użycie normalizacji (min-max) oraz skalowania (z-score).

Normalizacja $\mathrm{X}_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$

Normalizacja danych pomaga nam w przeskalowaniu danych do zakresów <0-1>, co może być zwodne przy atrybutach, które mają nierównomiernie rozłożone dane lub mają dużo wartości odstających (ang. noise) [1]. Weźmy na przykład sytuację, w której mamy 1000 wartości w zakresie 0-4 i jedną wartość równą 100. Przez jedną odstającą wartość po normalizacji skala znacząco się zmniejszy i faktyczna różnica między 1 a 4 będzie minimalna. Standaryzacja pomaga poradzić sobie z tym problemem, biorąc pod uwagę odchylenie standardowe.

Standaryzacja $X_{std} = \frac{x - \acute{x}}{\sigma_x}$

$\acute{x}$ – średnia dla atrybutu x

σ_x - odchylenie standardowe dla atrybutu x

Problemem standaryzacji jest utrata oryginalnych wartości i odchylenia standardowego, dane po normalizacji znacznie łatwiej odwrócić do wartości oryginalnych [1]. Ponieważ nie mamy szczególnie różnorodnych danych, wystarczy nam normalizacja min-max, problem pojawia się przy atrybutach ‘por’ i ‘mat’ gdyż mają one wartość 1 lub null, czyli po normalizacji pełnego zestawu danych uzyskamy dzielenie przez 0 (Xmax=1 i Xmin=1) dlatego należy pamiętać o zamianie wartości nullowych na 0

Błędne dane

W tym punkcie rozważymy dwa rodzaje potencjalnie błędnych danych

1. Błędy wynikające z odstających rekordów
2. Błędy wynikające z niepoprawnego zbierania lub wprowadzania danych

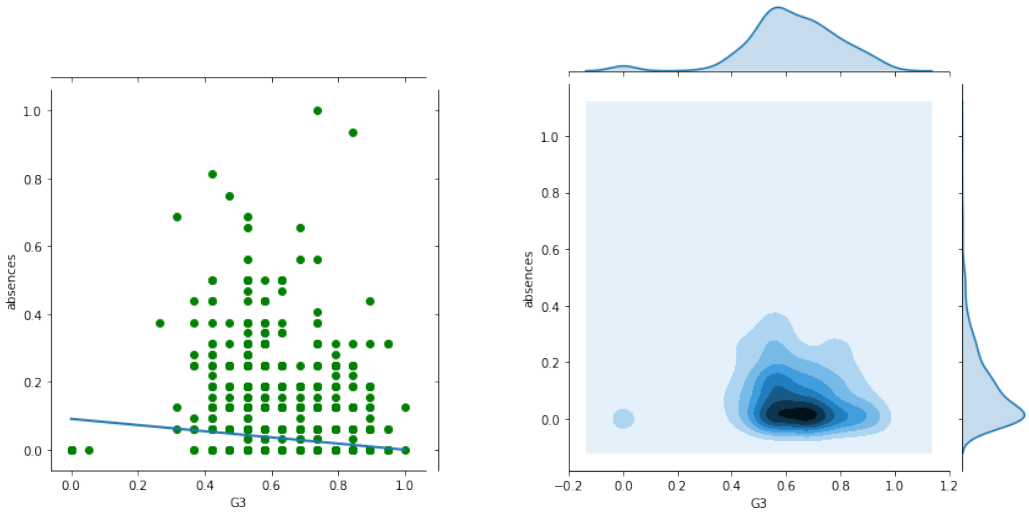
W pierwszym przypadku usuwamy rekordy, które radykalnie odstają od naszej średniej, na przykład przypadki osób wybitnie inteligentnych, które nie będą musiały się przykładać, a i tak osiągną wysoki wynik, przez co mogą one zaburzyć późniejsze obliczenia. W drugim przypadku rozpatrzone zostały przypadki ocen

pozytywnych, pojawiających się po negatywnych co w praktyce oznaczałoby, że uczeń nie zaliczył pierwszego roku, a zaliczył drugi, co jak wiemy, nie jest możliwe. Po analizie danych udało się znaleźć jeden właśnie taki rekord, który logicznie nie ma sensu, mianowicie ma oceny pozytywne za klasę 2 i 3 a ocenę negatywną za 1 nie jest to wiarygodny rekord, stąd został on usunięty.

```
-------------------- Niewiarygodny rekord ---------------------

   G1_por  G2_por  G3_por  failures_por
     0.0    11.0    11.0           0.0
```

Zajmiemy się również rekordami, które odstają od większości, Rysunek 3 przedstawia rekordy dla dwóch atrybutów 'G3' i 'abscences' oraz odwzorowanie korelacji między nimi równej -0.091. Korelacja została przedstawiona jako niebieska linia [4;12], jest ona dosyć płaska, ponieważ większość rekordów wypada w podobnym przedziale, co widać na rysunku obok . W prawym górnym rogu znajdują się dwa punkty, które odstają nie tylko od reszty punktów, ale i od głównej tendencji (korelacja ujemna). Z tej przyczyny rekordy zostały usunięte, gdyż są traktowane jako zakłócenia w danych, co oznacza, że rekordy mogą być prawidłowe, ale prawdopodobnie zaburzą wyniki przy dalszych obliczeniach. W rzeczywistości taki rekord może oznaczać osobę wyjątkowo utalentowaną, która nie musi często przychodzić na zajęcia, by mieć dobre oceny. Aczkolwiek w tym przypadku po dalszej analizie można założyć, że są to osoby, których często nie ma ze względów zdrowotnych, a nie podejścia do nauki.

```
------------ osoby z najwyższą liczbą nieobecności ------------

           G3  absences  health
274  0.736842    1.0000    0.00
525  0.842105    0.9375    0.25
```



Rysunek 3. Asocjacja między parametrami abscences i G3

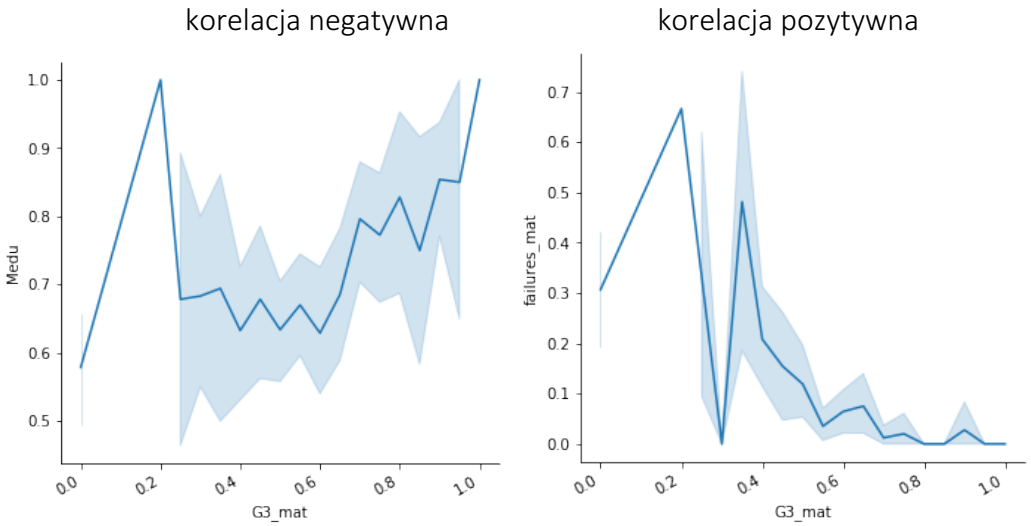
Korelacje atrybutów

Różnice w korelacji między przedmiotami

Rozpatrzenie różnic między ocenami dla obu przedmiotów matematyki i portugalskiego, by móc dokonać jakiejkolwiek analizy zależności między obydwoma zbiorami, musimy operować na zbiorze danePrzeciecie, gdyż żeby porównanie było wiarygodne, będziemy chcieli mieć te same osoby i rozpatrzeć jak różne mają one oceny z poszczególnych przedmiotów. Do porównania owych atrybutów użyjemy macierzy korelacji, która dla n atrybutów będzie macierzą nxn gdzie kolumny i wiesze to wszystkie atrybuty więc na przekątnej znajdzie się podobieństwo atrybutu do samego siebie, każda wartość podobieństwa jest wyliczana ze wzoru [2]:

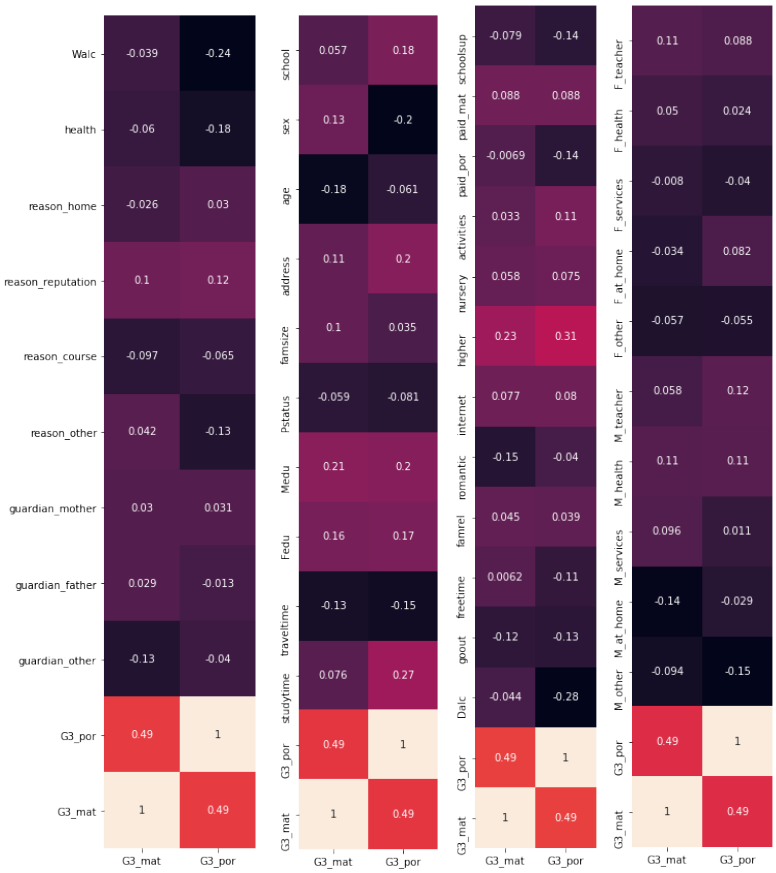
$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B}$$

gdzie:
A i B to nasze atrybuty
σ_A - odchylenie standardowe
$\bar{A}$– średnia atrybutu A
a_i – kolejne wartości atrybutu A
Jako wynik otrzymujemy wartość z zakresu <-1,1> gdzie 0 to brak jakiejkolwiek korelacji a -1 i 1 to maksymalna korelacja, z tym że dla 1 jest to korelacja pozytywna (wartości razem rosną) a dla -1 negatywna (jedna z wartości rośnie, gdy druga maleje)[10]. Więc porównując siłę korelacji, skupiamy na się na wartości bezwzględnej wyników, zanim przejdziemy dalej warto zwizualizować jak takie korelacje wyglądają.



Rysunek 4. Przykłady korelacji negatywnej i pozytywnej

Porównajmy więc wyliczone wcześniej wartości, dla atrybutów i ocen końcowych.



Rysunek 5. Korelacje atrybutów predykcyjnych z atrybutem wynikowym

parametr dotyczący czasu poświęconego na naukę ma zdecydowanie większy wpływ na przewidzenie ocen z portugalskiego. Możemy zatem wysunąć potencjalne wnioski,

- W przedmiotach ścisłych predyspozycje mają większe znaczenie niż w przedmiotach humanistycznych, gdzie szanse są bardziej wyrównane i liczy się bardziej czas włożony w naukę.
- Może też oznaczać większe zróżnicowanie w tempie pracy przy różnych przedmiotach, zakładając na przykład, że wypracowanie wszyscy uczniowie napiszą w zbliżonym czasie a zadania matematyczne część studentów zrobi zdecydowanie szybciej niż inni, widać tu również pewien mankament owego parametru, który skupia się na czasie poświęconym na dany przedmiot, nie mówiąc nic o ilości faktycznie wykonanej pracy.
- Może też oznaczać, że warto zmienić podejście prowadzenia przedmiotów ścisłych.

Porównanie wszystkich atrybutów

Rozpatrzyliśmy już które atrybuty mają zróżnicowany wpływ na oceny końcowe, przejdźmy więc do rozpatrzenia, które z atrybutów mają niski wpływ na obie oceny końcowe. Dla progu wartości bezwzględnej mniejszej niż 0.05 te atrybuty to:

```
---------------- ATRYBUTY O MAŁEJ KORELACJI -------------------

                   G3_mat    G3_por
famrel           0.044868  0.038807
reason_home     -0.026417  0.029768
guardian_mother  0.029581  0.030828
guardian_father  0.029050 -0.013242
F_services      -0.008042 -0.040039
```

Oznacza to, że są to atrybuty, których możemy się pozbyć, by ułatwić sobie późniejsze obliczenia oraz zredukować wielkość naszych danych. Natomiast atrybuty o największym znaczeniu, których wartość bezwzględna korelacji jest większa niż 0.2 dla obu przedmiotów to:

```
---------------- ATRYBUTY O DUŻEJ KORELACJI -------------------

                 G3_mat    G3_por
Medu           0.214327  0.203379
failures_mat -0.374683 -0.370106
higher         0.227307  0.307190
G1_mat         0.804905  0.559566
G2_mat         0.906443  0.514668
G3_mat         1.000000  0.492559
G1_por         0.516138  0.837542
G2_por         0.550597  0.889608
G3_por         0.492559  1.000000
```

Fakt, że edukacja matki jest istotniejsza niż edukacja ojca może mieć tu wiele potencjalnych przyczyn.

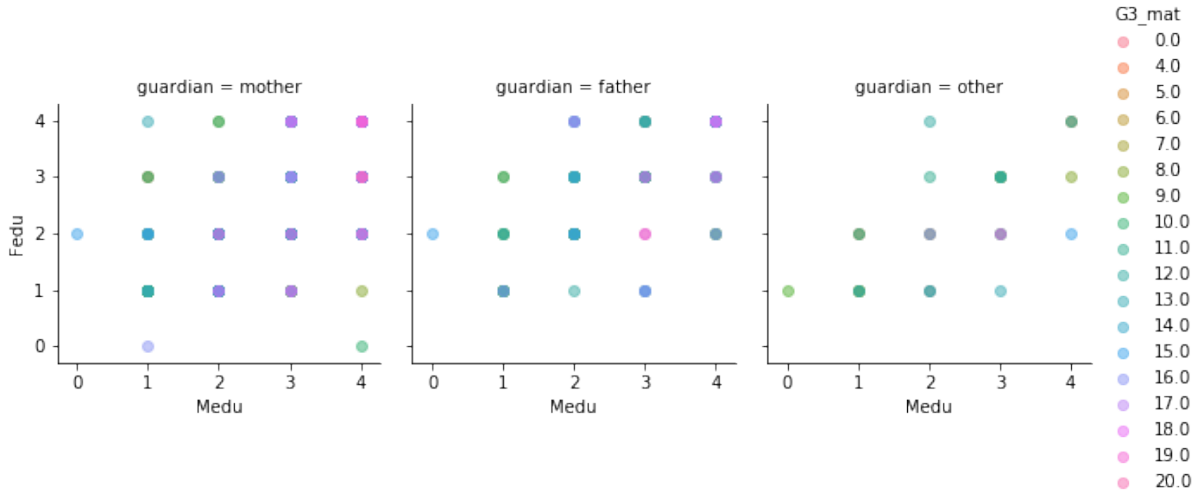
- samotne matki wychowujące dzieci
- większy wkład matki w edukacje dziecka
- ewentualne genetyczne predyspozycje

Rozważania te możemy zweryfikować dzięki parametrowi ‘Pstatus’ który określa czy rodzice są razem, czy nie

```
-------------------------- WSZYSCY ----------------------------

        G3_mat    G3_por
Medu  0.214327  0.203379
Fedu  0.161218  0.167544
```

```
-------------- OSOBY KTÓRYCH RODZICE SĄ RAZEM -----------------

        G3_mat    G3_por
Medu  0.219150  0.210245
Fedu  0.153506  0.166613


------------ OSOBY KTÓRYCH RODZICE  NIE SĄ RAZEM -------------

        G3_mat    G3_por
Medu  0.067504 -0.089203
Fedu  0.193997  0.113177
```

Takie wyniki w pierwszej kolejności wykluczają czynnik genetyczny, bo gdyby to on miał mieć znaczenie, dla wszystkich grup otrzymalibyśmy takie same wyniki, dodatkowo widzimy drastyczną różnicę między zależnościami w przypadku edukacji matki, gdy rodzice są razem, a gdy są osobno. Można by na tym etapie zagłębić dalej osoby, których rodzice nie są razem w zależności od prawnego opiekuna, lecz z racji niewielkiej ilości rekordów nie daje nam to wiarygodnych wyników. Natomiast ciekawą zależność widać, gdy weźmiemy wszystkich uczniów niezależnie od 'Pstatus' widzimy, że parametry dotyczące edukacji zarówno ojca, jak i matki odwzorowują się na ocenie ucznia najprecyzyjniej, gdy opiekunem dziecka jest matka a najgorzej, gdy jest nim ktoś inny niż rodzice.



Rysunek 6. Powiązania między parametrami Medu, Fedu i guardian

Analiza Koszykowa

Ciekawą formą analizowania danych jest analiza koszykowa, która ma na celu analizę prawdopodobieństwa wystąpienia produktów X przy wystąpieniu produktów Y. Warto zaznaczyć, że jeśli X daje Y, nie koniecznie oznacza to, że Y daje X. Weźmy pod uwagę przykład lampy i żarówki, jeśli kupuje lampę prawdopodobnie kupie też żarówkę, ale kupno żarówki już nie indykuje kupowania lampy. Załóżmy więc, że szukamy reguły X => Y wtedy w pierwszym kroku wyliczamy

jej wsparcie (ang. support), czyli stosunkowo jak często obie rzeczy występują razem .

$$support = \frac{frq(X,Y)}{N}$$

Gdybyśmy jednak poprzestali na tym kroku, nie biorąc pod uwagę, jak często X występuje ogólnie, może się okazać, że X wystąpi w każdym rekordzie, przez co wywnioskujemy, że skoro Y zawsze występuje z X, to znaczy, że jest między nimi relacja. By uniknąć tego problemu, wyliczmy pewność (ang. confidence), która sprawdza również częstość występowania X [2;9].

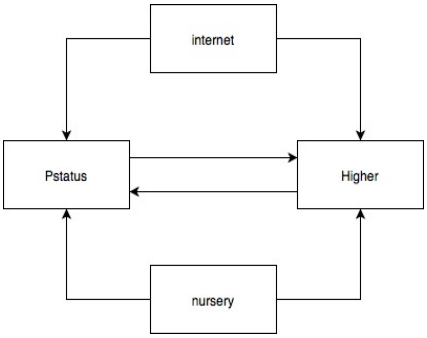
$$confidence = \frac{frq(X,Y)}{frq(X)}$$

W naszym przypadku taka analiza ma sens tylko dla atrybutów binarnych, jeśli ustawimy próg akceptowalności reguł na confidence > 0.85 i support >0.68 otrzymamy asocjacje.

$$lift = \frac{support}{support(X) * support(Y)}$$

antecedents	consequents	Support	confidence	lift	leverage	conviction
(Pstatus)	(higher)	0.787037	0.896309	1.003123	0.002451	1.026915
(higher)	(Pstatus)	0.787037	0.880829	1.003123	0.002451	1.023014
(nursery)	(higher)	0.722222	0.900000	1.007254	0.005201	1.064815
(nursery)	(Pstatus)	0.700617	0.873077	0.994295	0.004020	0.960531
(internet)	(higher)	0.695988	0.905622	1.013546	0.009302	1.128251
(internet)	(Pstatus)	0.682099	0.887550	1.010778	0.007273	1.084160

Możemy z nich wyczytać zarówno informacje o chęci do wyższej edukacji, co jest celem tej pracy, ale i możemy się dowiedzieć, co utrzyma rodziców razem, co pośrednio też ma wpływ na chęć do podjęcia wyższej edukacji. Dla czytelności wyniki zostały przedstawione na grafie, z którego możemy wyczytać, że internet w domu, chodzenie do przedszkola czy rodzice mieszkający razem prowadzą do chęci posiadania wyższego wykształcenia. Mniej istotne w naszym przypadku, ale moim zdaniem dużo ciekawsze okazuje się fakt, że mamy większe prawdopodobieństwo rodziców do zostania razem, jeśli poślą dziecko do przedszkola i mają w domu internet. Dodatkowo widzimy, że chęć bycia studentem i chęć rodziców do bycia razem jest jedyną relacją symetryczną, co oznacza, że chęci wpływają na siebie z obu stron.



Rysunek 7. Wizualizacja wyników analizy koszykowej

MODEL PREDYKCYJNY

Przy doborze odpowiedniego modelu warto przede wszystkim przyjrzeć się danym. W naszym przypadku większość atrybutów to atrybuty binarne zarówno te, które były binarne od początku, jak i te, które zostały przekształcone na potrzebę dalszych obliczeń. Reszta atrybutów to atrybuty numeryczne, przy czym na tym etapie warto zwrócić uwagę na inny sposób podziału danych, ze względu na ich ilość mamy atrybuty ciągłe i dyskretne. Ciągłe to takie, które mogą mieć nieskończenie wiele wartości jak na przykład waga. Natomiast w naszym zbiorze danych występują tylko atrybuty dyskretne, czyli o określonej liczbie wartości, dodatkowo jak widzieliśmy w fazie analizy danych, zakresy te są stosunkowo niewielkie. Podsumowując, pracujemy na danych, dla których istnieje określona liczba możliwych rekordów, zależny nam więc na dobrym podziale danych, a nie na wyliczaniu konkretnych wartości. Tak więc rozpatrzymy drzewo decyzyjne oraz Kernel SVM, gdyż oba algorytmy zajmują się podziałem danych o skomplikowanym kształcie, a wszystko wskazuje na to że to będzie jednym z większych problemów podczas klasyfikacji. Przed przystąpieniem do modelowania musimy jednak zająć się odpowiednim podziałem danych.

Podział danych

Train test split

By mieć jak określić skuteczność naszego modelu będziemy dzielić dane losowo na zbiór uczący i testowy. Podejście to zapobiega zjawisku, jakim jest ‘overfitting’ czyli sytuacja, w której nasz model świetnie radzi sobie z danymi, do których mieliśmy dostęp, ale przy nowych danych nie jest w stanie sobie poradzić [1] . Dzieje się tak ze względu na to, że nasz model odwzorowuje wtedy nasze dane, a nie tendencje naszych danych. Gdyby przerzucić się ze sztucznej na naszą inteligencję wyglądałoby to, jak nauczenie się rozwiązań wszystkich zadań na pamięć, zamiast zrozumieć, na czym owo zadanie polega. Znając na pamięć wszystkie rozwiązania, perfekcyjnie poradzimy sobie ze znanymi zadaniami, ale gdy pojawią się nowe dane, nie będziemy w stanie wykonać zadania. Natomiast jeśli chodzi o nasze dane mamy ich stosunkowo niewiele, dla przewidywanego atrybutu mamy dużą rozbieżność między wartością minimalną a maksymalną i dodatkowo wiemy, że ma ona nierówny rozkład. Aby uwzględnić te nierówności, podczas podziału danych użyjemy metody ‘próbkowania’ czyli wybierania fragmentu danych z dużych zbiorów, jest to metoda, która najwiarygodniej stara się odwzorować całe dane.

Stratified sampeling

Metoda ta polega na uwzględnieniu klas wynikowych i ich wielkości, to znaczy dobraniu danych tak, żeby każda klasa znalazła się w wybranej części danych. Dodatkowo by dobrze odwzorować dane, zależy nam, aby stosunek ilości rekordów należących do danych klas odzwierciedlał tę proporcję, które występowały w oryginalnym zbiorze danych [2]. By koncepcyjnie zrozumieć tę metodę, można wyobrazić sobie, jak wyglądają predykcje wyników wyborów w kraju. Wiadomo, że ankieterzy nie zapytają każdego obywatela więc dobierają

osoby zróżnicowane społeczno-ekonomicznie, w takich ilościach by odpowiadały one proporcjonalnie do społeczeństwa. W naszym wypadku będziemy wybierali dane na zbiór treningowy, tak by uwzględnić wszystkie klasy oraz częstość ich występowania.

Drzewo decyzyjne

Ogólną koncepcją drzewa decyzyjnego jest podział danych w każdym z węzłów ze względu na wartość jednego z atrybutów i wyznaczenie oczekiwanej wartości dla końcowych grup danych znajdujących się w liściach. Wyliczanie zaczynamy w korzeniu z całym zbiorem danych i odpalamy procedurę która:
W pierwszym kroku decyduje czy podzielić dane, czy nie.

a. Jeśli nie to szacuje wartość tego liścia.
b. Jeśli tak to szuka atrybutu, który najlepiej podzieli te dane.

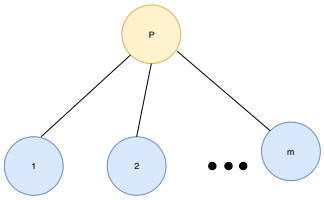
Mając wybrany atrybut, dzieli dane na grupy, po czym dla każdej grupy wywołujemy tę procedurę. Koncepcja jest dosyć prosta, problem pojawia się przy podejmowaniu decyzji.

1. Kiedy przestać dzielić dane? Ustalamy, że nie chcemy dalej dzielić danych zazwyczaj w przypadkach, gdy zachodzi jeden z warunków.
   a. Liczba rekordów w węźle jest poniżej jakiejś ustalonej wcześniej wartości minimalnej.
   b. Wszystkie lub prawie wszystkie rekordy mają taką samą wartość oczekiwaną.
2. Jak dobierać atrybut, według którego będziemy dzielić dane w węźle? Dzielimy dane w taki sposób, który najbardziej rozdzieli wartości oczekiwane, czyli zyskać jak najwięcej informacji w węźle (ang. Information Gain), to znaczy szukamy atrybutu, dla którego Information Gain będzie największa [1].

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

Oznacza ono, że dla węzła p w którym się znajdujemy, szukamy największej 'I' - miary zanieczyszczenia w tym węźle minus suma miar zanieczyszczenia z potencjalnych węzłów potomnych. Gdybyśmy jednak zatrzymali się na tym etapie, węzeł zawierający 2000 rekordów miałby taki sam wpływ na wynik jak węzeł zawierający 20 rekordów.
By uniknąć tego zaburzenia w proporcji danych, każdą miarę mnożymy przez jej zawartość w węźle rodzica, $\frac{N_j}{N_p} = \frac{|D_j|}{|D_p|}$ czyli ilość rekordów w węźle dzielona przez ilość rekordów w węźle rodzica.

f - cecha/atrybut na podstawie którego zostanie wykonany podział w węźle
D_p – zestaw danych węzła nadrzędnego
D_j – zestaw danych j-tego węzła potomnego
I - miara zanieczyszczenia
N_p – całkowita liczba próbek węzła nadrzędnego
N_j – całkowita liczba próbek j-tego węzła potomnego



Rysunek 8. Przykład węzła

Dalej pojawia się nam pytanie, jak obliczyć miarę zanieczyszczenia
I tu pojawia się wiele możliwych odpowiedzi, omówimy trzy najpopularniejsze, każda z nich bazuje na wyliczaniu $p(i|t)$, czyli proporcji między próbkami należącymi do klasy a wszystkimi próbkami w węźle t. Czyli jeśli w węźle mamy 100 rekordów, z czego 30 ma oczekiwaną wartość równą 1 to $p(i = 1|t) = 30/100$

Wskaźnik Giniego (ang. Gini impurity) – zajmuje się minimalizacją prawdopodobieństwa nieprawidłowej klasyfikacji [1].

$$I_G(t) = \sum_{i=1}^{c} p(i|t)\big(1 - p(i|t)\big)$$

Entropia –zajmuje się maksymalizacją wzajemnych informacji w drzewie [1].

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) log_2 p(i|t)$$

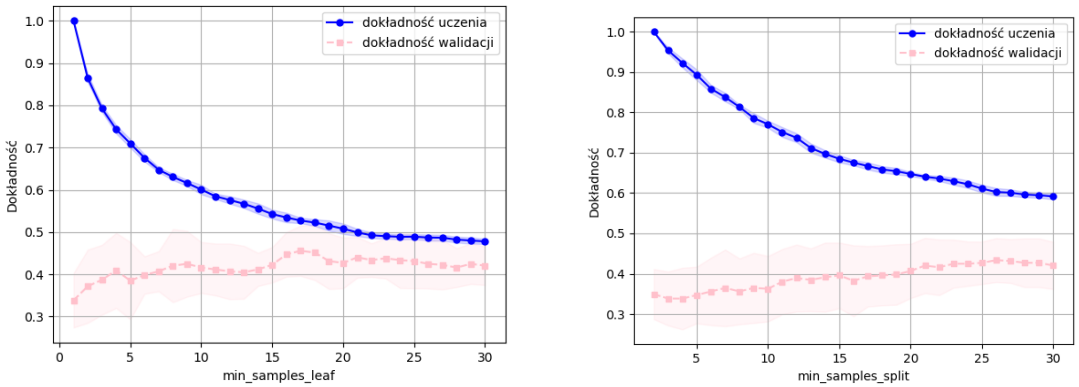
Błąd klasyfikacji – zajmuje się znalezieniem najlepszego podziału [1].

$$I_E(t) = 1 - max\,\{p(i|t)\}$$

Wskaźnik Giniego oraz entropia zazwyczaj mają podobne wyniki, natomiast dużo informacji możemy uzyskać dzięki błędowi klasyfikacji z racji, iż Wskaźnik Giniego, oraz entropia zajmują się doborem atrybutu do konkretnego węzła (to samo zadanie nieco innymi metodami). Natomiast błąd klasyfikacji wyznacza granice podziału w danym węźle, czyli dla danego atrybutu szuka optymalnego podziału. Oczywiście metoda ta ma sens jedynie przy atrybutach numerycznych i nominalnych gdzie dane według jakiegoś atrybutu mogą być podzielone na wiele możliwości, przy atrybutach binarnych podział jest oczywisty.

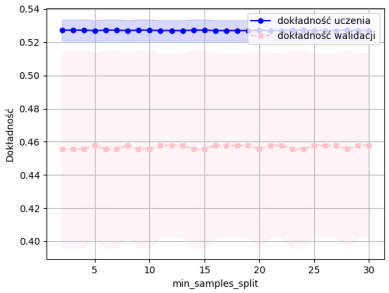
Dobór parametrów

Jak już wiemy, kryterium do tworzenia liści w drzewie może być minimalna liczba rekordów do podziału czy też minimalna liczba rekordów w liściu. W tym podrozdziale zajmiemy się próbą odnalezienia tych parametrów, które niestety będą musiały być dostrojone eksperymentalnie. Do budowy drzewa użyjemy metody sklearn.tree.DecisionTreeClassifier, dla której postaramy się znaleźć optymalne wartości dla parametrów „min_samples_split" oraz „min_samples_leaf" Jak już wiemy, w modelu zależy nam na dokładności uczenia oraz dokładności jej testowania, czyli, innymi słowy, walidacji szukamy więc kompromisu między dokładnością uczenia a dokładnością walidacji [1]. Wartości dla parametrów znajdujemy na wykresie, który bierze pod uwagę dokładność oraz wartość parametru. Widzimy, że parametr „min_samples_leaf" znacząco łatwiej dobrać, bo widać dosyć jasno, że wartość 17 zbliża nam bardzo wykresy dokładność walidacji i uczenia przy zachowaniu stosunkowo wysokiej dokładności uczenia .



Rysunek 9. Pierwszy etap doboru parametrów dla drzewa decyzyjnego

Po dobraniu wartości dla parametru „min_samples_leaf" możemy zobaczyć jak jego dobór zmieni zależności dla parametru „min_samples_split".

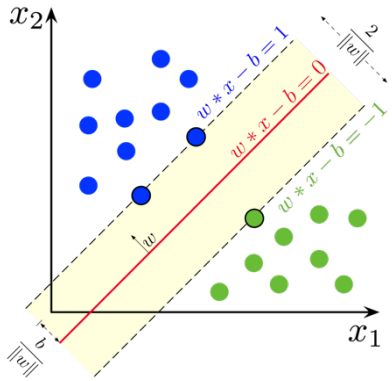


Rysunek 10. Drugi etap doboru parametrów dla drzewa decyzyjnego

Widzimy, że nie będzie on odgrywał aż tak dużej roli, jak parametr „min_samples_leaf", ale widać dokładnie, że wartość 5 da nam najlepsze wyniki.

Kernel SVM

Większość algorytmów zajmuje się jedynie podziałem danych Support Vector Machine stara się te dane podzielić jak najlepiej. Czyli znajduje prostą, która jest pod takim kątem, który maksymalizuje sumę odległości wszystkich rekordów od tej prostej. W praktyce minimalizujemy sumy odległości rzutów rekordów na prostą od środka układu współrzędnych. Prosta ta działa jak perceptron, który klasyfikuje dane na dwie grupy. Łatwo zobrazować to wyznaczając dwie proste równoległe do prostej dzielącej dane takie, że pierwsza przechodzi przez najbliższy punkt podzbioru A a druga najbliższy punkt podzbioru B, te proste wyznaczają tak zwany margines, który maksymalizujemy.



Rysunek 11. Support Vector Machine

Źródło: https://en.wikipedia.org/wiki/Supportvector_machine#/media/File:SVM_margin.png

Przy nieliniowych problemach, gdy nie jesteśmy w stanie bezbłędnie podzielić zbiorów, używamy klasyfikacji przy użyciu miękkiego marginesów (ang. Soft margin classification) [7 ; 1], który minimalizuje:

$$\frac{1}{2}\|w\|^2 + C\left(\sum_i \xi_i\right)$$

Z rysunku 10 widać, że szerokość marginesu jest równa 2 dzielone przez długość „w", maksymalizacja tego wyrażenia jest równoznaczna z minimalizacją $\frac{1}{2}\|w\|^2$,czyli podstawowego SVM. Dane niepodzielne liniowo dzielone prostą będą obarczone jakimś błędem, aby uwzględnić minimalizacje błędnych klasyfikacji wprowadzamy ξ, które jest odległością błędnie zaklasyfikowanego punktu do krańca marginesu, na którym powinien się znaleźć, wszystkie te błędy sumujemy. Mamy więc teraz maksymalizację marginesu przy minimalizacji błędów. Nachodzi wiec pytanie, czym jest parametr C, jest on niczym innym niż zwykłym współczynnikiem wyznaczanym przez nas eksperymentalnie który będzie decydował o stosunku istotności minimalizacji błędów do maksymalizacji marginesu. Kernel-SVM służy nam w przypadkach, gdzie chcemy znaleźć najlepszy podział danych, ale mamy niepodzielne liniowo dane. W takim przypadku przenosimy dane do wyższego wymiaru, tak by było możliwe znalezienie płaszczyzny, która nam te dane podzieli. Łatwo zobrazować to, wyobrażając sobie zdjęcie góry zrobione z lotu ptaka i chęci liniowego oddzielenia czubka góry od

reszty, jest to niemożliwe. Aczkolwiek gdy dodamy nowy wymiar, czyli dostaniemy nową perspektywę mianowicie zdjęcie z ziemi, bez problemu znajdziemy płaszczyznę, która oddzieli wierzchołek od reszty góry. Przejdźmy więc do wyliczeń, w pierwszym kroku mapujemy dane do większej przestrzeni, w której dane są już rozdzielne liniowo. Dzieje się to dzięki nieliniowym kombinacjom cech ,na przykład ich iloczynu ,przy użyciu funkcji $\phi: R^d \rightarrow R^k gdzie\ (k \gg d)$, po przekształceniu wszystkich rekordów otrzymamy macierz K .

$$K = \begin{bmatrix} \phi(x_1)^T\phi(x_1) & \cdots & \phi(x_1)^T\phi(x_n) \\ \vdots & \ddots & \vdots \\ \phi(x_n)^T\phi(x_1) & \cdots & \phi(x_n)^T\phi(x_n) \end{bmatrix}$$

W wyższej przestrzeni możemy użyć algorytmu SVM do podziału tych danych.
Problematyczna staje się ilość obliczeń niezbędna do wyliczenia nowych cech szczególnie przy wielowymiarowych danych. By ominąć ten problem, użyjemy tak zwanej sztuczki jądra.
Zakłada ona następujące cechy macierzy K

- symetryczność ($K = K^T$)
- dodatnio półokreśloność - czyli dla jakiejś wartości jest zawsze dodatnia lub zerowa.

Dzięki tym cechom dla par rekordów w tej macierzy $x_n i\ x_m$ możemy wyznaczyć ich nieliniowe podobieństwo , przekształcając oba przy użyciu funkcji $\phi(x_n)^T\phi(x_m)$, oznacza to, że istnieje taka funkcja $k(x_n, x_m)$, która dla dwóch zmiennych jest w stanie wyliczyć to powiązanie [1;6]. Dzięki czemu jesteśmy w stanie obejść się bez wyliczania ϕ, jednak potrzebujemy znać funkcje jądra $k(x_n, x_m)$, która jest wyznacznikiem podobieństwa $x_n\ do\ x_m$, czyli iloczynem skalarnym tych wektorów , jej interpretacje to między innymi:

Wielomianowa funkcja jądra

$$k(x_n, x_m) = ({x_n}^T x_m + \theta)^p$$

θ - próg
p - parametr wyznaczany ręcznie

Jądro tangensa hiperbolicznego

$$k(x_n, x_m) = tanh(\eta\ {x_n}^T x_m + \theta)$$

Jądro radialnej funkcji bazowej(RBF) = jądro gaussowskie

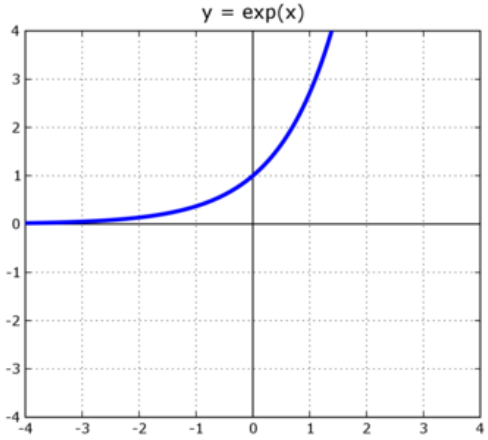
$$k(x_n, x_m) = exp\left(\frac{-\|x_n - x_m\|^2}{2\sigma^2}\right)$$

Można je uprościć do

$$k(x_n, x_m) = exp(-\gamma\|x_n - x_m\|^2)$$

dla $\gamma = \frac{1}{2\sigma^2}$, który będzie przedmiotem naszej optymalizacji [1]. Jądro gaussowskie jest dosyć intuicyjne, jeśli popatrzymy na $\|x_n - x_m\|^2$, szukamy

odległości będącej różnicą wektorów, czyli ich podobieństwa, tyle że w zależności od rozkładu tych punktów wartość ta może się znacząco różnić. Aby to unormować, znajdujemy eksperymentalnie γ , które unormuje nam te wartości , mimo unormowania wartości ciągle ich zakres jest nam nieznany, a dla macierzy podobieństwa zależałoby nam na wartościach od 0 do 1, gdzie 1 to porównanie tych samych wektorów. Jeśli przyjrzymy się wykresowi funkcji eksponencjalnej . f(x)=$e^x = exp\ (x)$



Rysunek 12. Wykres funkcj eksponencjalnej

Źródło: https://pl.wikipedia.org/wiki/Funkcja_wykładnicza

Widzimy, że dla $x\ \epsilon\ (-\infty, 0>$ zbiór wartości znajduje się w zakresie <0,1> , więc dla iloczynu γ ($\gamma \geq 0$) oraz odległości podniesionej do kwadratu otrzymamy nieujemny wynik, po przemnożeniu całości przez -1 nasz wynik wpasowuje się w dziedzinę $x\epsilon\ (-\infty, 0>$ .Po dodaniu funkcji exp() mamy gwarancje wyniku w zakresie <0,1>.

ANALIZA POPRAWNOŚCI MODELU

Dokładność

k-fold cross validation , hold-out cross validation

By zacząć mierzyć dokładność naszego modelu, jak już wcześniej omówiliśmy, musimy podzielić dane na treningowe i testowe. Problem może pojawić się przy dzieleniu naszego zbioru danych, bo ciężko ocenić, w jakich proporcjach warto je podzielić. Zakładamy, że aby model był skuteczny, wolimy raczej większą część użyć jako zbiór uczący, zazwyczaj jest to 70 / 30 taką metodę nazywamy hold-out cross validation i sprawdzi się ona świetnie przy bardzo dużych zbiorach[11]. Aczkolwiek gdy nasza liczba rekordów nie jest za duża, przy takim podejściu możemy mieć mocno zniekształcony obraz poprawności modelu. Ponieważ przy mniejszej ilości danych mamy większe prawdopodobieństwo, że na przykład jakaś grupa rekordów cała wyląduje w danych testowych. W naszym

przypadku może być to dużym problemem z racji małej ilości rekordów i stosunkowo dużej rozbieżności w wartościach atrybutu, który chcemy przewidzieć.
Tu na ratunek przychodzi nam metoda k-fold cross validation, która dzieli zbiór danych na k części $x_1, x_2, x_3 \dots x_k$,tej samej wielkości i przeprowadza k walidacji krzyżowych. Używając kolejno jako zbiór testowy $x_1, x_2, x_3 \dots x_k$ a jako zbiór treningowy resztę danych następnie wyciągamy średnią z wszystkich wyników co jest naszym ostatecznym wynikiem. Dużą zaletą tej metody jest fakt, iż każdy rekord był użyty jako dane testowe i treningowe.
Istnieje również wariant tej metody dla bardzo małych zbiorów danych zwana 'leave-one-out'[5] która dla bazy danych zawierającej n rekordów wykonuje k-fold cross validation dzieląc dane na n grup , przez co na dane testowe składa się jeden rekord, oczywiście użycie takiej metody dla większej ilości danych nie da dobrych wyników i będzie bardzo czasochłonny.
Istnieje też metoda OOT 'out of time' która bierze pod uwagę aktualność danych i to jak mogą się zmieniać, aczkolwiek metoda ta nie nadaje się zupełnie w naszym przypadku [5,14].

Po przeprowadzeniu k-fold cross validation możemy porównać dokładność obu modeli, która dla drzewa decyzyjnego wynosi w przybliżeniu 46,5 % a dla Kernel SVM jest to już 41%. Oba wyniki mogą wydawać się niezbyt wysokim osiągiem, aczkolwiek dokładność modelu to nie wszystko. Można by rozważyć który model byłby lepszy taki o 0% dokładności, który zawsze myli się tylko o 1 dla osoby, która powinna dostać wynik 17 dostanie 18 lub 16. Czy model skuteczny w 75% który dla pozostałych 25% będzie dawał zupełnie inne wyniki niż powinien, na przykład osobie, która powinna mieć 2 dostanie 20. Nie ma prostej odpowiedzi na pytanie, który z modeli byłby lepszy, bo zawsze zależy to od postawionego problemu, w naszym przypadku pewnie chcielibyśmy być gdzieś pomiędzy. Znając systemy edukacyjne w większości krajów, możemy przypuszczać, że oceny zamykają się w skali 1-6 z ewentualnymi plusami czy minusami , a nam została przedstawiona skala ocen 0-20. Tak więc pewnie zależałoby nam, by dokładnie oszacować ocenę, ale czy będzie to ocena z plusem, czy minusem nie jest już takie istotne. Do tego podejścia do problemu wrócimy później, teraz zajmijmy się zobrazowaniem tego, jak bardzo nasz model jest omylny w aspekcie nie tego, jak wiele danych jest błędnie klasyfikowane, a jak bardzo źle są one klasyfikowane. Do zobrazowania tych błędów służą metryki skuteczności.

Metryki Skuteczności

Metryki skuteczności pozwolą nam na wgląd do jakości klasyfikacji modelu, tudzież zobaczymy, w jakim stopniu się myli oraz w których miejscach. Użyjemy do tego macierzy pomyłek (ang. confusion matrix), w której jako kolumny przedstawione są rekordy, które należą do danej klasy, a jako wiersze rekordy, które zostały do tej klasy zakwalifikowane, jak więc nie trudno się domyślić na przekątnej znajdują się wszystkie rekordy prawidłowo przypisane do danej klasy [1].

```
Macierz pomyłek:

[ 4  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
[ 0  0  0  1  0  0  0  0  0  0  0  0  0  0  0]
[ 0  0  0  1  2  0  0  0  0  0  0  0  0  0  0]
[ 3  0  0  4  2  2  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  9  1  0  0  0  0  0  0  0  0]
[ 1  0  0  1  1 16  5  3  1  1  0  0  0  0  0]
[ 0  0  0  0  0  6 21  2  2  0  0  0  0  0  0]
[ 0  0  0  0  0  0  6  3 11  2  0  0  0  0  0]
[ 0  0  0  0  0  0  0  6 15  4  0  0  0  0  0]
[ 0  0  0  0  0  0  0  1 13  3  2  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  3 12  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  1  5  4  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  2  1  6  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  4  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  1  0  0]
```

Sprawdzając poprawność klasyfikacji naszego modelu, patrząc kolejno na każdą z klas możemy uzyskać 4 potencjalne wyniki klasyfikacji.

true positive – w przypadku gdy rekord należy do klasy i został tak zaklasyfikowany (przekątna macierzy)
true negative- w przypadku gdy rekord nie należy do klasy i został tak zaklasyfikowany (suma wszystkich wartości bez kolumny i wiersza danej klasy)
false positive – w przypadku gdy rekord nie należy do klasy ale został do niej zaklasyfikowany (suma wartości w kolumnie - true positive)
false negative – w przypadku gdy rekord należy do klasy ale nie został do niej zakwalifikowany (suma wartości w wierszu - true positive)
Po wyliczeniu tych parametrów dla każdej z klas, otrzymamy następujące wyniki:

```
true positive:  [4, 0, 0, 4, 0, 16, 21, 3, 15, 3, 12, 4, 6, 0,
0]
false negative:  [0, 1, 3, 7, 10, 13, 10, 19, 10, 16, 3, 6, 3,
4, 1]
false positive:  [4, 0, 0, 3, 5, 17, 12, 12, 27, 11, 9, 1, 5, 0,
0]
true negative:  [186, 193, 191, 180, 179, 148, 151, 160, 142,
164, 170, 183, 180, 190, 193]
```

PRECYZJA - (ang. precision) Określa jaka część rekordów zaklasyfikowanych do tej klasy została zaklasyfikowana prawidłowo w zakresie [0,1]

$$PRE = \frac{TP}{TP + FP}$$

```
------------------------- precision ----------------------------

0.5        0.         0.         0.57142857 0.         0.48484848
 0.63636364 0.2        0.35714286 0.21428571 0.57142857 0.8
 0.54545455 0.         0.
średnia dla wszystkich klas:  0.3253968253968254
```

PEŁNOŚĆ - (ang. recall) Określa jaka część rekordów z wszystkich dobrze zaklasyfikowanych należy do tej klasy.

$$PEŁ = \frac{TP}{FN + TP}$$

```
-------------------------- recall ----------------------------

1.         0.         0.         0.36363636 0.         0.55172414
 0.67741935 0.13636364 0.6         0.15789474 0.8         0.4
 0.66666667 0.         0.
średnia dla wszystkich klas:  0.3569136597519011
```

PARAMETR F1: (ang. F1 score) szczególnie przydatny przy różnorodnych danych określa średnią harmoniczną (ang. harmonic mean) między pełnością a precyzją. Używamy tu średniej harmonicznej zamiast zwykłej średniej arytmetycznej, gdyż w przypadku dużych różnic w wartościach między precyzją a pełnością da nam lepszy obraz faktycznej prawidłowości modelu [1] . Na przykład w przypadku pełności=0. 1 i precyzji=0.0001 średnia arytmetyczna wyniesie 0,05005 a harmoniczna = 0,0001998001998

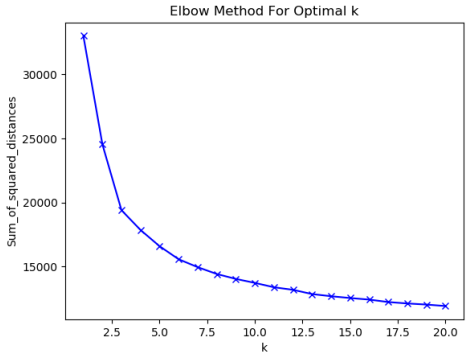
$$F1 = \frac{2 * PRE * PEŁ}{PRE + PEŁ}$$

```
---------------------------- F1 -------------------------------

0.66666667 0.         0.         0.44444444 0.         0.51612903
 0.65625    0.16216216 0.44776119 0.18181818 0.66666667 0.53333333
 0.6        0.         0.
średnia dla wszystkich klas:  0.3250154454252913
```

Wróćmy jednak do naszych rozważań co do szerszego spojrzenia na dane ze skali 0-20 na 1-6 ,oczywiście by przejść dalej, warto potwierdzić naszą hipotezę. Jednym z pomysłów może być próba zgrupowania podobnych rekordów. Jeśli okaże się, że najlepiej grupują się one do 6 klas, będzie to indykator, że możemy mieć racje. Oczywiście nie jest to stuprocentowe potwierdzenie naszej teorii z racji, gdyż może się okazać, że podobieństwo jest kwestią podobnych tendencji dla niektórych osób bez odzwierciedlenia w ocenach końcowych. Poszukamy więc optymalnej liczby klas dla atrybutu oceny końcowej. Posłużymy się tu algorytmem k-means, który działa następująco:

0. Znajdź k losowych punktów (czyli rekordów) – te rekordy nazywamy centroidami.
1. Wszystkie rekordy przyporządkuj do najbliższego (w praktyce najbardziej podobnego) centroidu .
2. Wylicz średnią dla każdej grupy, jeśli są takie jak centroidy idź do 3, jak nie to ustaw je jako nowe centroidy i idź do 1.
3. Wylicz i zapamiętaj wariancje i idź do 0.

Warunki stop mogą być różne, zależy nam na podziale oferującym najmniejszą wariancję [16]. Nas interesuje ile to k byłoby równe przy optymalnej klasyfikacji, metoda punktu łokcia zajmuje się tym właśnie problemem. Wiemy, że wariancja będzie malała wraz ze wzrostem liczby centroidów (k) aż do momentu, w którym każdy rekord stanie się centroidem i wariancja wyniesie 0. Zależy nam na minimalizacji wariancji, ale oczywiście nie chcemy doprowadzić do sytuacji, w której każdy rekord jest osobnym klastrem. Znalezienie złotego środka w metodzie punktu łokcia polega na znalezieniu kolokwialnego 'łokcia' czyli momentu, od którego wzrost K nieznacząco zmniejsza wariancje. Oczywiście jest to metoda raczej dająca nam jakieś przybliżenie nie 100% pewność wyboru [17].



Rysunek 13. Dobór ilości klas do algorytmu kmeans

Mając niewielkie potwierdzenie naszych przypuszczeń, możemy przejść do próby lepszego poznania kontekstu naszych danych, a mianowicie systemu oceniania uczniów w liceum z kraju, z którego pochodzą nasze dane. Dzięki tej wiedzy oprócz potwierdzenia naszych przypuszczeń możemy również, dowiedzieć się jak dane te powinny być zgrupowane. Gdyż 21 ni jak nie dzieli się na 6 równych zakresów i tu okazuje się, że portugalski system następująco przelicza wyniki ocen.

```
<0 ,3.4> -> 1 ,<3.5,9.4> -> 2, <9.5 ,13.4>  -> 3
<13.5,15.4>  -> 4, <15.5,17.4>  -> 5, <17.5,20>      -> 6
```

Ponieważ operujemy na ocenach całkowitych nasze przedziały będą mapowane następująco.

```
<0 ,3>     -> 1
<4,9>      -> 2
<10 ,13>   -> 3
<14,15>    -> 4
<16,17>    -> 5
<18,20>    -> 6
```

Po takim mapowaniu danych dokładność naszego modelu wzrasta do 75% , wzrost precyzji był oczekiwany z racji zmniejszenia liczby potencjalnych klas, aczkolwiek tak duży wzrost znacząco wspiera naszą teorię dotyczącą 6 klas i ich większego podobieństwa w obrębie klasy niż przypadkowych porównań o tej samej różnicy co to oznacza? Jeśli mamy racje to rekordy, a i b, między którymi różnica wynosi 1, będą do siebie bardziej podobne, jeśli wylądują w tej samej klasie na przykład 18 bardziej podobne do 19 niż 17, bo i 18 i 19 należą do klasy 6, a 17 należy do klasy 5.

Mając obliczoną dokładność takiego modelu, wróćmy do metryk skuteczności i zobaczmy, jakie dadzą nam wyniki po mapowaniu naszych klas.

```
Macierz pomyłek:

 [  6   0   0   0   0   0]
 [  2  15   8   0   0   0]
 [  1   1 111   2   0   0]
 [  0   0  13  15   0   0]
 [  0   0   0   6   5   3]
 [  0   0   0   0   3   3]

true positive:  [6, 15, 111, 15, 5, 3]
false negative:  [0, 10, 4, 13, 9, 3]
false positive:  [3, 1, 21, 8, 3, 3]
true negative:  [185, 168, 58, 158, 177, 185]



------------------------- precision --------------------------

0.66666667 0.9375     0.84090909 0.65217391 0.625      0.5
średnia dla wszystkich klas:  0.7037082784365393


-------------------------- recall ----------------------------

1.         0.6        0.96521739 0.53571429 0.35714286 0.5
średnia dla wszystkich klas:  0.6596790890269151


---------------------------- F1 ------------------------------

0.8        0.73170732 0.89878543 0.58823529 0.45454545 0.5
średnia dla wszystkich klas:  0.6622122484729145
```

SYSTEM PREDYKCYJNY

W celu wprowadzenia utworzonego modelu do życia oraz realnego pozytywnego jej wpływu na poprawę wysiłków edukacyjnych stworzony został prosty system który na podstawie modelu jest w stanie określić szacowaną ocenę końcową. Na etapie czystego szacowania wyników taki system nie wnosił by wiele do poprawienia końcowych ocen studentów. Z tego też powodu z modelu zostały wyizolowane atrybuty mające szanse na poprawę jego ocen (w zależności od zachowania ucznia lub pomocy z zewnątrz którą mógł by on otrzymać). Są to parametry ‘schoolsup' , 'famsup','paid','activities','higher' ,'internet' ,'studytime' ,'freetime' , 'goout' , 'Dalc' ,'Walc' ,’failures’,'absences’ i będziemy się do nich odnosić jako atrybuty zmienne . Z punktu widzenia państwa czy też ośrodka edukacyjnego parametry dotyczące wsparcia finansowego czy dostępu dziecka do Internetu mogą wnieść ważne informacje dla takich placówek co pomoże im w

oszacowaniu opłacalności inwestycji w tych sektorach. Natomiast reszta parametrów jest przeznaczona do eksperymentów dla studenta a przede wszystkim do podbudowy jego motywacji na przykład jak już było omawiane w nauce języka ojczystego dużym faktorem był czas włożony w naukę , posiadając taki system dany student mógł by zobaczyć jak podskoczyła by jego ocena końcowa gdyby wkładał więcej czasu w naukę owego przedmiotu.

System 1.0

System ma wczytywać dane danej osoby i umożliwiać zmienianie wartości atrybutów zmiennych. Na podstawie nowo powstałych rekordów będziemy przewidywać, w jakiej z klas powinny się one znaleźć. Mamy możliwe dwa podejścia co do wyliczanych ocen, możemy je zostawić w zakresie 0-20 lub zgrupować je do 1-6 , by zaobserwować zmiany, które mogą okazać się niewielkie, zostaniemy jednak przy zakresie 0-20, szczególnie że był on w stanie określać nam wyniki z błędem maksymalnie o 3 klasy. Aby przetestować system sprawdzimy, czy zmieniając atrybuty zależne od studenta na "najgorsze" lub "najlepsze" przyniesie to zmianę w estymacji oceny. Jako „najgorsze" wartości atrybutów zmiennych zakładamy najwyższe wartości dla parametrów o ujemnej korelacji oraz najniższe wartości dla parametrów o dodatniej korelacji, analogicznie postępujemy z „najlepszymi wartościami". Mamy więc po trzy wartości dla każdej osoby: pierwotną, po dobrych zmianach i po złych zmianach. Sprawdzone zostało ile rekordów się zmienia , ile rekordów zmienia się dobrze (lepsza ocena po dobrych zmianach niż po złych), ile na źle oraz różnice ocen po dobrych zmianach i po złych

```
------------------------ KERNEL SVM ---------------------------

liczba zmienionych rekordów: 466
liczba nie zmienionych rekordów: 110
liczba źle zmienionych rekordów: 11
liczba dobrze zmienionych rekordów: 455
średnia różnica: 3.223175965665236


------------------------ DECISION TREE ------------------------

liczba zmienionych rekordów: 175
liczba nie zmienionych rekordów: 401
liczba źle zmienionych rekordów: 75
liczba dobrze zmienionych rekordów: 100
średnia różnica: 1.0
```

Po utworzeniu systemu opartego na przewidywaniu oceny widzimy, dużą rozbieżność w wynikach między drzewem decyzyjnym a kernel SVM. To który z tych modeli lepiej estymuje zależy od tego, jak bardzo wkład danej osoby może mieć faktyczny wpływ na jej oceny . Szczególnie w przypadku drzewa decyzyjnego zmiana parametrów zależnych od studenta nie ma dużego wpływu na wyniki, co wydaje się nieintuicyjne, szczególnie biorąc pod uwagę wyliczone wcześniej korelacje. Na tym etapie warto się zastanowić czy dla dwóch osób o takim samym podłożu społeczno-

ekonomicznym i włożonych wysiłkach w naukę otrzymamy identyczne wyniki . Czy każde rodzeństwo zachowujące się tak samo, to znaczy mające takie same wartości parametrów zmiennych będzie miało takie same oceny ? Oczywiście nie bo jest nam wiadome, że każda osoba ma inne predyspozycje, geny, łatwość skupienia, efektywność itd. Dobrym przykładem że taki rekord mógłby zdarzyć się przy większej ilości danych, jest rekordów różniących się tylko ocenami i edukacją matki .

---------- Dwa podobne rekordy ----------

	Osoba1	Osoba2
Medu	2	3
G1	10	11
G2	11	12
G3	10	12

System 2.0

Podejmiemy się więc utworzenia alternatywnej wersji systemu, która będzie brała pod uwagę fakt, że mimo zmieniających się atrybutów osoba pozostaje taka sama.
Koncepcja talentu
Wprowadzimy więc pomocniczy parametr określony jako 'talent' aczkolwiek warto tu odnotować, że nie staramy się zdefiniować znaczenia słowa talent jest to tylko pomocnicza nazwa.Jest to sztucznie wytworzony parametr na użytek naszego systemu, który nie będzie jawny dla uczniów. Będzie miał największe wartości u osób, które mają najgorsze warunki, najmniej wkładają czasu , nie dbają o naukę, ale mają dobre oceny . Natomiast najmniejsze wartości u osób z dobrym podłożem społeczno-ekonomicznym, które się starają, wkładają czas itd. ale mają słabe oceny.
Po pierwsze spróbujemy policzyć, "wkład studenta" i wyliczyć czy miał "łatwy" start– parametry dotyczące rodziny, zdrowia czy zamieszkania. Wykorzystamy do tego wyliczoną wcześniej macierz korelacji, tyle że teraz interesuje nas tylko lista korelacji (corr) wszystkich atrybutów z oceną końcową (oczywiście odrzucamy korelacje oceny końcowej z samą sobą). Czyli dla n atrybutów i wartości tych atrybutów $x_1, x_2 \dots x_n$, będzie to wyglądało następująco.

$$\sum_{i=1}^{n} corr_i * x_i$$

Warto tu zaznaczyć, że wszystkie wartości atrybutów uległy wcześniej normalizacji stąd dostaniemy idealnie wyważoną sumę w zależności od istotności. Dodatkowo jak wiemy, korelacje mogą być dodatnie, jak i ujemne, dzięki czemu zapewniamy, że parametry, które mają negatywną korelację, obniżają sumę, a parametry o korelacji dodatniej ją podwyższają.

Pojawia się jednak problem, zakres wartości tej sumy jest dla nas nieznany, by tego uniknąć, dokonamy jej normalizacji, znalezienie największej i najmniejszej potencjalnej sumy będzie dosyć proste, gdyż mamy znormalizowane wartości dla x. Więc krańcowe wartości sumy będą zawsze dla wszystkich x równych 1. Tak więc maksymalna suma maxCorr będzie równa sumie wszystkich dodatnich korelacji

a minCorr będzie równa sumie wszystkich minimalnych korelacji dochodzimy wiec do wzoru.

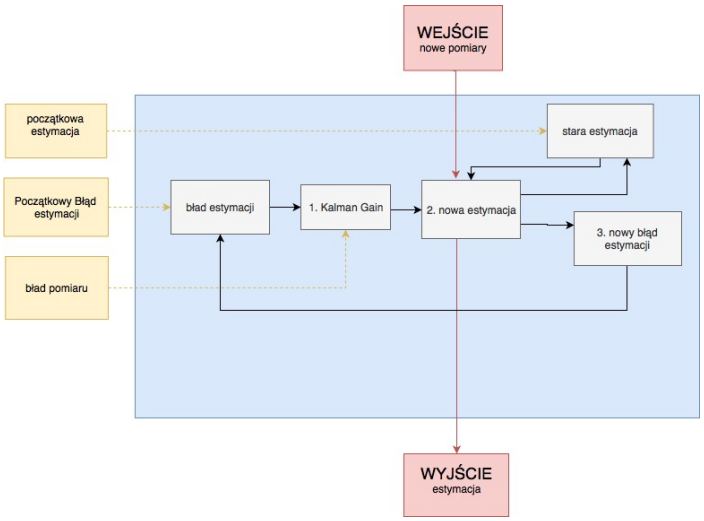
$$\frac{\sum_{i=1}^{n} corr_i * x_i - minCorr}{maxCorr - minCorr}$$

Teraz zależy nam na stosunku tego, co wyliczyliśmy do oceny , chcemy uzyskać duży talent dla dużej oceny i złych warunków, a jak na razie mamy odwrotną sytuację gdzie 0 to wartość dla najgorszych warunków, a 1 dla najlepszych. Odwrócimy więc tę zależność, wystarczy odjąć ją od 1 i tak otrzymamy ostateczny wzór na talent.

$$talent = ocena * \left(1 - \frac{\sum_{i=1}^{n} corr_i * x_i - minCorr}{maxCorr - minCorr}\right)$$

Koncepcja filtru Kalmana

Do utworzenia naszego systemu użyjemy koncepcji filtru Kalmana, który bazuje na połączenie pomiaru i predykcji. Zakładając, że model użyty przy budowie Systemu 1.0 przewiduje nam ocenę, ale budując system, chcemy umożliwić użytkownikowi obserwacje zmieniających się atrybutów, które będą naszym pomiarem. Mamy więc imitacje całego systemu, w którym na początku musimy dostarczyć estymacje oceny, błąd oceny oraz błąd pomiaru. Początkowa estymacja oceny przed jakimikolwiek zmianami wartości parametrów jest bezproblemowa, bo zaczynamy, znając wartość oceny, a skoro znamy ocenę, to błąd możemy ustawić jako 1. Problemem pozostaje jednak błąd pomiaru, bo nie jesteśmy w stanie go stu procentowo wyliczyć, a będzie on oddziaływał na nasze wyniki przez cały czas działania algorytmu, ponieważ jeśli chodzi o estymacje i błąd estymacji nawet jak byśmy pomylili się na starcie, zostaną one wyrównane w późniejszych iteracjach, a błąd pomiaru ustalamy z góry[13]. Błąd został wyznaczony eksperymentalnie by zmieścić się w średnich wahaniach w ocenach



Rysunek 14. Schemat filtru Kalmana

Mając wyznaczone wartości, możemy przystąpić do iteracyjnego wykonywania algorytmu, którego schemat został przedstawiony poniżej. Dla naszych potrzeb jest

to znacząco uproszczony system, w którym na poziomie każdej z iteracji mamy do wyliczenia 3 zmieniające się parametry, które zostały zaznaczone na rysunku i opisane poniżej.

Kalman Gain - K_G

miara która reprezentuje rozkład naszej pewności co do dwóch estymacji w zakresie <0,1>. Dąży do 1 gdy zakładamy że poprzednia estymacja jest mniej wiarygodna niż estymacja pomiaru, czyli obarczona większym błędem (E_{est}niż $(E|pom)$, gdzie pom – pomiar, est-estymacja [15]

$$K_G = \frac{E_{est}}{E_{est} + E_{pom}}$$

Nowa estymacja -$ocena_t$

wyliczamy ocene w momencie t z poprzedniej oceny oraz oceny wyliczonej z nowych pomiarów, przy uwzględnieniu Kalman gain [15].

$$ocena_t = ocena_{t-1} + K\left(ocena_{pom} - ocena_{t-1}\right)$$

Po rozpisaniu wzoru łatwiej zobaczyć, że jest to faktycznie wyważona średnia miedzy obiema ocenami.

$$ocena_t = ocena_{t-1} + K * ocena_{pom} - K * ocena_{t-1}$$
$$ocena_t = (1 - K) * ocena_{t-1} K * ocena_{pom}$$

Dzięki wprowadzonemu wcześniej stałemu parametrowi, jakim jest talent, po przekształceniu wzoru możemy uzyskać równanie wyliczające ocenę z uwzględnieniem zmieniających się wartości atrybutów.

$$ocena_{pom} = \frac{talent}{1 - \frac{\sum_{i=1}^{n} corr_i * x_i - minCorr}{maxCorr - minCorr}}$$

Nowy błąd estymacji - E_{est_t}

estymacja błędu w momencie t [18].

$$E_{est_t} = \frac{E_{est_{t-1}} * E_{pom}}{E_{est_{t-1}} + E_{pom}}$$

Jest to wzór zakładający rozkład normalny oraz stopniowe zmniejszanie się błędu, natomiast my zakładamy, że jesteśmy w stanie dobrze oszacować ocenę na początku a wraz z zmianami estymacja ta staje się coraz mniej wiarygodna i coraz bardziej musimy polegać na pomiarze. Dlatego też użyjemy funkcji odwrotnej do przedstawionej wyżej by błąd estymacji ciągle rósł. Zakładamy rozkład normalny, bo niewielkie zmiany mogą przynieść na duże rezultaty ale jest mało prawdopodobne

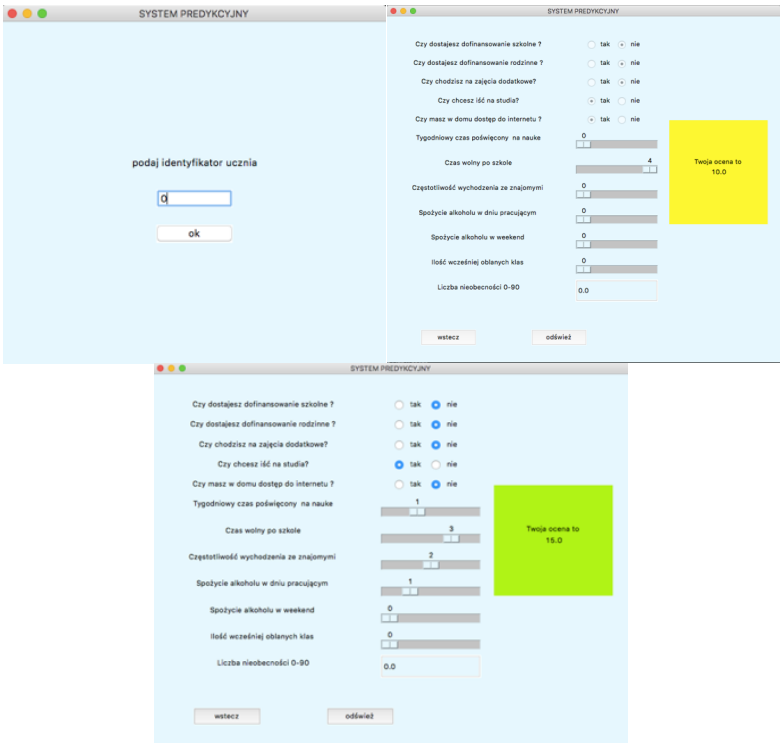
żeby działały one z taką samą siłą przy zmianie wszystkiego . Ciężko jest ocenić poprawność działania takiego systemu, gdyż nie posiadamy informacji zwrotnej. Nie wiemy, jak duży wpływ mają osobiste predyspozycje. Aczkolwiek jest to skala, którą w przyszłości dałoby się wyliczyć, wystarczyłyby nam wartości wszystkich atrybutów zbierane od tych samych osób przez jakiś okres. Widzielibyśmy wtedy, jak bardzo są oni w stanie wpływać na zakres swoich ocen. Na chwilę obecną nasz najlepszy szacunek to wyliczenie średniej z różnic najlepszej i najgorszej oceny każdego ucznia co sugeruje, że wahania oscylują w zakresie równym 2. Warto tu odnotować, że z danych wynika, że statystyczny uczeń będzie miał wyższą ocenę końcową, niż dwie poprzednie co też ma sens, biorąc pod uwagę wiek i prawdopodobnie chęć dobrych ocen na świadectwie końcowym lub do pokazania przy rekrutacji na uczelnie wyższe. Oczywiście możemy to sprawdzić.

```
---------- Średnie osób nie chcących iść na studia ------------

G1:  4.6231884057971016
G2:  4.942028985507246
G3:  6.0144927536231885
Różnią G1 i G3  1.391304347826087


----------- Średnie osób  chcących iść na studia --------------

G1:  7.7417677642980935
G2:  7.925476603119584
G3:  9.325823223570191
Różnią G1 i G3  1.584055459272098
```

Jak widać, nie ma wielkiej różnicy, co sugeruje albo chęć dobrych ocen końcowych, łatwiejszy materiał lub odpowiedzialniejsze podejście. Podsumowując, jest to system czysto hipotetyczny, którego wyniki mogą być akceptowalne tylko i wyłącznie po zebraniu większej ilości danych w przyszłości.

Obsługa systemu

Zarówno system 1.0, jak i 2.0 dla użytkownika wyglądają tak samo i powinien być tak samo obsługiwany. System został wyposażony w prosty interfejs, by każdy uczeń mógł sprawdzić, jak zmieniłyby się jego oceny, gdyby zmienił jakieś ze swoich zachowań. W pierwszym kroku uczeń podaje swój identyfikator by wczytać oddane wcześniej przez siebie dane. Następnie ma możliwość zmiany wartości atrybutów oraz obserwacji zmiany oceny , dodatkowo wraz ze wzrostem oceny zmienia się kolor okna od czerwieni po zieleń. Student może też zresetować całość i sprawdzić wynik dla innych atrybutów [8].



Rysunek 15. Graficzny Interfejs Użytkownika

BIBLIOTEKI

Całość pracy została wykonana w pythonie przy użyciu następujących bibliotek

Pandas – Biblioteka zapewnia typ obiektu, jakim jest DataFrame, który zapewnia szybki i przejrzysty sposób radzenia sobie ze zbiorem danych. Między innymi do usuwania, dodawania rekordów, wybierania podzbiorów czy wszelakich metod łączenia danych.

Numpy- Narzędzie ułatwiające operacje matematyczne w szczególności algebraiczne

Wizualizacja danych – Do obserwacji danych użyta została biblioteka Matplotlib.pyplot oraz bazująca na niej biblioteka Seaborn. Seaborn idealnie sprawdziła się w przypadkach, gdy zależy nam na szybkim i profesjonalnym zobrazowaniu danych , podczas gdy matplotlib sprawdza się przy potrzebie bardziej wyspecyfikowanych wykresów

SciKit learn – W zakresie uczenia maszynowego prawdopodobnie jedna z najistotniejszych bibliotek przydatna do przygotowania danych, tworzenia modelu i jego ewaluacji. Do przygotowania danych użyte zostały między innymi funkcje do losowego podziału danych na zbiór uczący i testowy oraz losowania warstwowego (Stratified Sampeling), jak i funkcja MinMaxScaler do normalizacji danych.

Wykorzystane zostały modele drzewa decyzyjnego oraz kmeans. A na końcu model był oceniony dzięki wyliczeniu metryk skuteczności oraz dostrojony przy użyciu krzywych uczenia i walidacji.

Tkinter – biblioteka zapewniająca tworzenie prostych interfejsów graficznych

PODSUMOWANIE

Jako pierwszy krok dane zostały przeanalizowane pod kątem typu oraz stopniowania, dzięki czemu mogły zostać mapowane w sposób, w który były przejrzysty do analizy. Dzięki analizie właśnie zostały odkryte ciekawe powiązania oraz odsłonione zostały obszary, w których mamy nie wystarczająco informacji. Następnie usunięte zostały wszystkie nieprawidłowe rekordy i zostały wyliczone korelacje między atrybutami a ocenami końcowymi z obu przedmiotów. Korelacje wskazały na wiele różnic między przedmiotami, ale i korelacje na tyle niewielkie dla obu przedmiotów, że bez utraty na jakości modelu mogliśmy się ich pozbyć. Następnie dane zostały podzielone i utworzone zostały dwa modele predykcyjne, które zostały później przeanalizowane pod względem nie tylko dokładności, ale i skuteczności, bo jak się przekonaliśmy skuteczność wyniku to nie wszystko. Podczas całego procesu próby zrozumienia danych i doboru modeli wypłynął istotny brak w danych, który pomógłby nam ocenić wpływ zmiany zachowania na zmianę oceny. Pojawiło się więc pytanie, czy jest to coś, co bylibyśmy w stanie jakkolwiek zmierzyć. Czy istnieje metryka pozwalająca nam uzyskać tego typu rozwiązanie? Jest to na pewno pole do rozwoju w przyszłości możliwe, że dałoby się zebrać dane co do czyjegoś IQ lub odnaleźć jakieś powiązania w genetyce. Jednak ciągle są to tylko hipotezy i nie wiemy, czy któraś z tych wartości byłaby w stanie odzwierciedlić to, o co nam chodzi. Na potrzebę tej pracy podjęliśmy się wyliczenia talentu, lecz jest to miara bez potwierdzenia, szczególnym problemem jest przeskalowanie wagi talentu, oszacowania jak bardzo jesteśmy w stanie coś zmienić.

BIBLIOGRAFIA

1.„Python Uczenie maszynowe” , Sebastian Raschka , Wyd. HELION , 2018

2.„Data mining Concepts and Techniques” , Jiawei Han, Micheline Kamber , Jian Pei ,third edition, Elsevier,2012

3.Źródło danych: https://archive.ics.uci.edu/ml/datasets/student%2Bperformance

4. „Geometric interpretation of a correlation ”,Zeszyty Naukowe Warszawskiej Wyższej Szkoły Informatyki Nr 9, , 2013 Zenon Gniazdowski, https://zeszyty-naukowe.wwsi.edu.pl/zeszyty/zeszyt9/Geometric_interpretation_of_a_correlation.pdf

5.“Cross Validation concepts for modeling”, Gopal Prasad Malakar , https://www.youtube.com/watch?v=BEQdMea5usw, [dostęp: 16.05.2017]

6. „Nonlinear Dimensionality Reduction: KPCA”, David Thompson,caltech https://www.youtube.com/watch?v=HbDHohXPLnU , [dostęp: 25.05.2018]

7.„Lecture 70 — Soft Margin SVMs”,Leskovec,Rajaraman,Ullman,Stanford University, https://www.youtube.com/watch?v=8xbnLHn4jjQ, [dostęp: 13.04.2016]

8.“Python - GUI Programming (Tkinter) ”, https://www.tutorialspoint.com/python/python_gui_programming.htm

9. „A Gentle Introduction on Market Basket Analysis — Association Rules”, Susan Li, https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce [dostęp: 25.09.2017]

10.„Why Feature Correlation Matters …. A Lot!”, Will Badr, https://towardsdatascience.com/why-feature-correlation-matters-a-lot-847e8ba439c4, [dostęp: 18.01.2019]

11.“Hold-out vs. Cross-validation in Machine Learning”, Eijaz Allibhai , https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f, [dostęp: 03.10.2018]

12.“Data visualization”, https://material.io/design/communication/data-visualization.html#

13.“Kalman Filtering - Theory and Practice Using MATLAB ”, third edition, Mohinder S. Grewal ,California State University at Fullerton , Angus P. Andrews , Rockwell Science Center (retired) , John Wiley & Sons, Inc , 2008

14. „Why isn't out-of-time validation more ubiquitous?" , Tomas Dvorak https://towardsdatascience.com/why-isnt-out-of-time-validation-more-ubiquitous-7397098c4ab6, [dostęp: 12.02.2019]

15." ONE-DIMENSIONAL KALMAN FILTER WITHOUT THE PROCESS NOISE", Alex Becker, 2018, https://www.kalmanfilter.net/kalman1d.html

16." StatQuest: K-means clustering" , Josh Starmer, https://www.youtube.com/watch?v=4b5d3muPQmA, [dostęp: 23.05.2018]

17. „10 Tips for Choosing the Optimal Number of Clusters" , Matt.0 https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92, [dostęp: 27.01.2019]

18. „The Kalman Filter: An algorithm for making sense of fused sensor insight", Sharath Srini, https://towardsdatascience.com/kalman-filter-an-algorithm-for-making-sense-from-the-insights-of-various-sensors-fused-together-ddf67597f35e , [dostęp: 18.04.2018]