Wyszukiwanie studentów uzdolnionych na podstawie danych demograficzno-społecznych

Szymon Arabas
Julia Kotowska
Przemysław Michałowski
Yauheni Semianiuk

1. Ogólne założenia

Cel utworzenia modelu
Celem utworzenia modelu jest umożliwienie predykcji przed pierwszymi egzaminami, którzy studenci mogą wykazywać szczególne zdolności w nauce. Dzięki temu modelowi będzie można wyszczególnić, a następnie objąć dodatkową opieką dydaktyczno-naukową jednostki uzdolnione, które mogłyby od samego początku reprezentować uczelnię w zawodach i konkursach.

Zbiór danych
Dane pochodzą ze strony
http://archive.ics.uci.edu/ml/datasets/Student+Performance?fbclid=IwAR3BuuYfrtsQ920re0KhvyzSRiJ8br87sf6BbPJc1ePdCDvsyt4K15oQYRg
Zgodnie z informacją w opisie zbioru, zostały one zebrane do badania: P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
[Web Link]

Zmienna prognozowana
Jako zmienną prognozowaną przyjęto binarne przekształcenie zmiennej G3. Jako klasę pozytywną zdefiniowano wyniki znajdujące się w pierwszym kwartylu najwyższych wyników tej zmiennej, a jako klasę negatywną - wszystkie wyniki poniżej tego kwartyla. Udział klasy pozytywnej w całym zbiorze wynosi 20%, minimalny wynik punktowy wymagany aby znaleźć się w klasie pozytywnej wynosi 15 punktów. Wszystkie odchylenia powinny stać się przedmiotem analiz i z dużym prawdopodobieństwem mają istotny wpływ na budowę drzewa *klasyfikacyjnego .*
Zmienna ta została wybrana jako prognozowana, ponieważ stanowi ona najlepszy dostępny kwantyfikator umiejętności studentów pośród dostępnych danych.

Spodziewane zależności i hipotezy badawcze
Przed rozpoczęciem analiz liczbowych należy wyposażyć się w odpowiednią wiedzę ekspercką. Na podstawie literatury przedmiotu i wcześniejszych badań, można spodziewać się następujących zależności:

1) Uczniowie, często spożywający alkohol będą mieli gorsze wyniki w nauce (na podstawie: Jeremy Staff, Megan E. Patrick, Eric Loken, Jennifer L. Maggs „Teenage Alcohol Use and Educational Attainment" *Journal of Studies on Alcohol and Drugs*, s. 848-858, Rutgers University, 2008)
2) Dzieci uczęszczające w przeszłości do przedszkoli mają większe szanse wykazywać szczególne uzdolnienia (Lynn A. Karoly, James H. Bigelow *The Economics of Investing in Universal Preschool Education in California*, s. 44-52, wyd. Rand, Santa Monica, 2005)

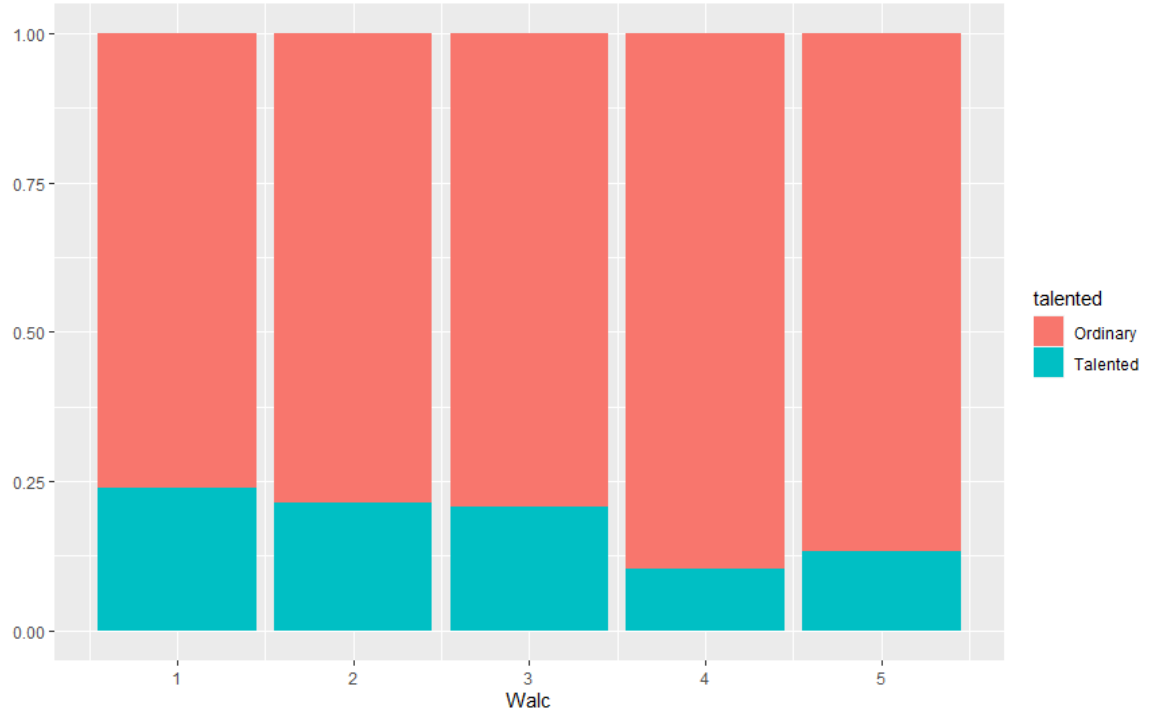
3) Fakt pobierania dodatkowych zajęć (niezależnie od ich charakteru) nie powinien być istotnie skorelowany z uzdolnieniem w skali populacji (Jorge A. Costa. Alexandre Ventura, Sara Azevedo *Private Tutoring in Portugal*, s.151-165, University of Aveiro)
4) Uczniowie spędzający bardzo mało czasu na nauce, nie będą osiągali wysokich wyników

2. Wstępna analiza zmiennych

W ramach wstępnej analizy, sprawdzono udział klasy pozytywnej dla poszczególnych wartości wszystkich atrybutów. Obliczone także Information Value. Ze względu na małą liczbę zmiennych o wysokim IV w dalszych krokach zdecydowano się na zastosowanie lasów losowych, w związku z czym wartości IV mają w tej analizie znaczenie drugorzędne.

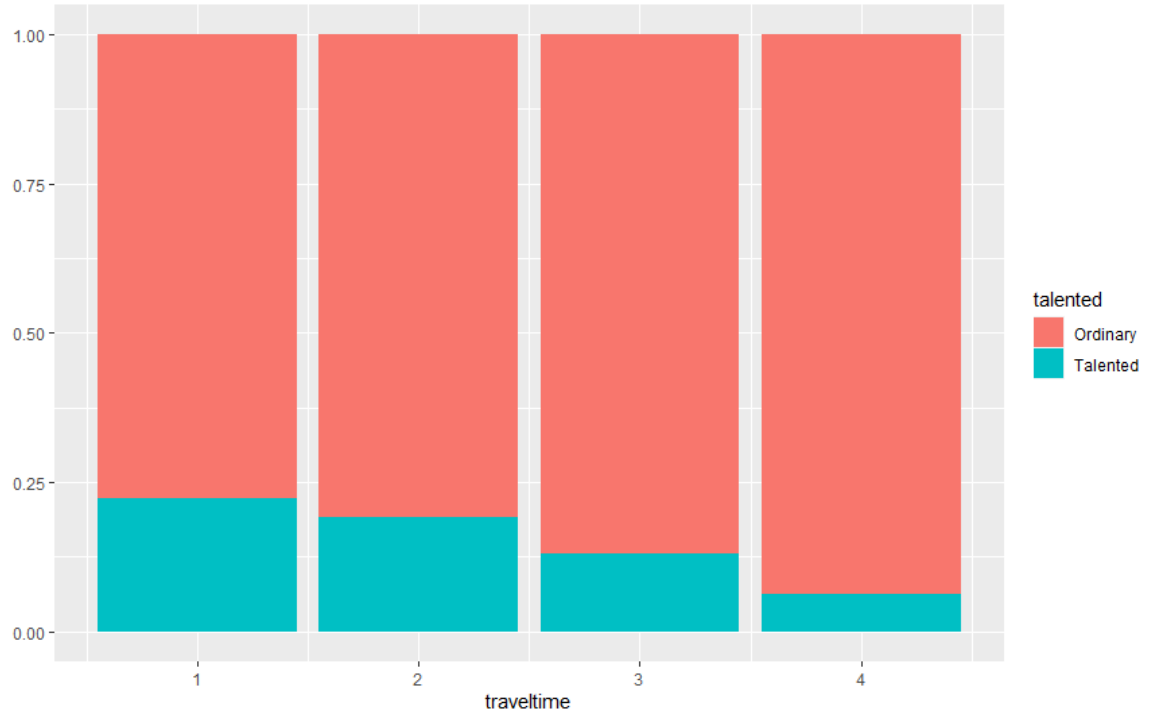
Walc

IV = 0.01. Typ zmiennej: porządkowa. Zmienna obrazująca konsumpcję alkoholu w weekendy, im wyższa liczba, tym większa konsumpcja. Jak widać, udział klasy pozytywnej dla wysokich wartości (4 i 5) walc jest wyraźnie niższy niż dla wartości mniejszych. Sygnalizuje to słabą ujemną korelację między wynikami w nauce a ilością spożywanego alkoholu.



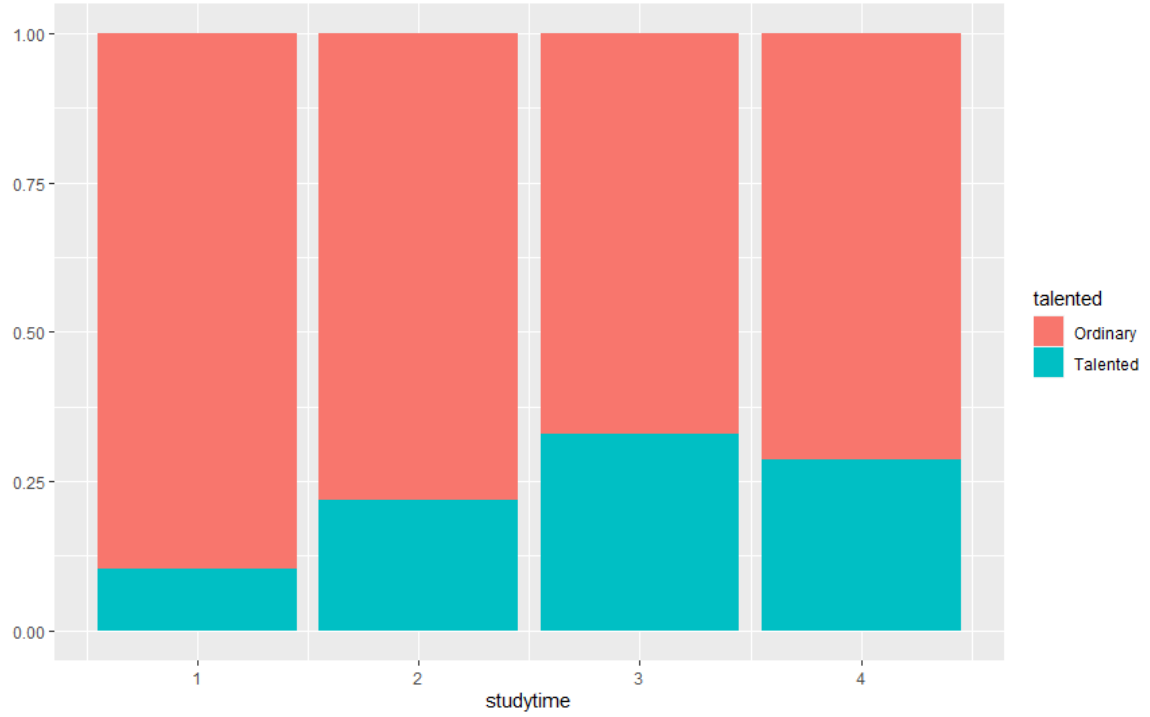
Traveltime

IV = 0.06. Typ zmiennej: porządkowa, gdzie 1 - <15 min, 2 – 15-30 min, 3 – 30-60 min, 4 - > 60min. Zmienna obrazuje czas podróży z domu do szkoły (dziennie w jedną stronę). Można zauważyć dość systematyczny spadek udziału klasy pozytywnej wraz ze wzrostem czasu podróży. Może to być spowodowane zmęczeniem oraz mniejszym czasem na naukę spowodowanym codziennym przemieszczaniem się.



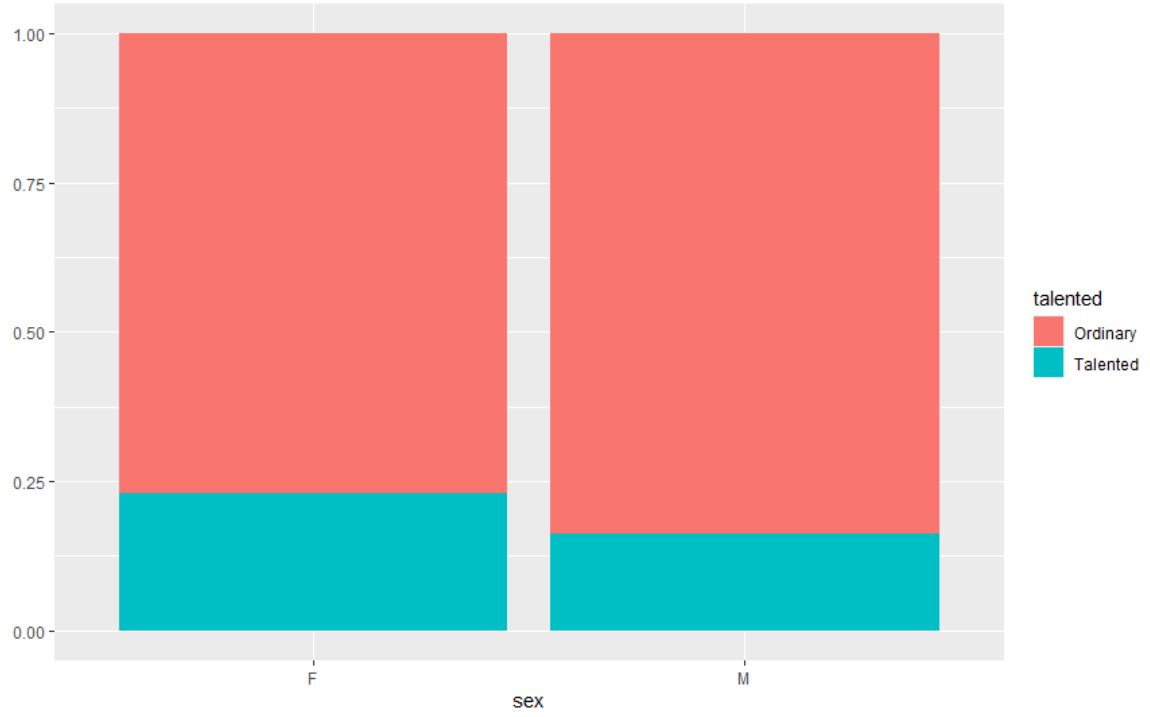
Studytime

IV = 0.25. Typ zmiennej: porządkowa, gdzie 1 - < 2 godz., 2 – 2-5 godz., 3 – 5-10 godz., 4 - >10 godz. Zmienna obrazuje tygodniowy czas poświęcony na samodzielną naukę. Za wyjątkiem najdłuższego czasu poświęconego na naukę, widać dodatnią korelację między wynikiem, a czasem poświęconym na naukę. Ponadto istnieje ujemna korelacja pomiędzy zmienną studytime a freetime, której istotność potwierdził test chi-kwadrat (p-value<0.05). Wysoka wartość IV wskazuje na przydatność tej zmiennej w predykcji.



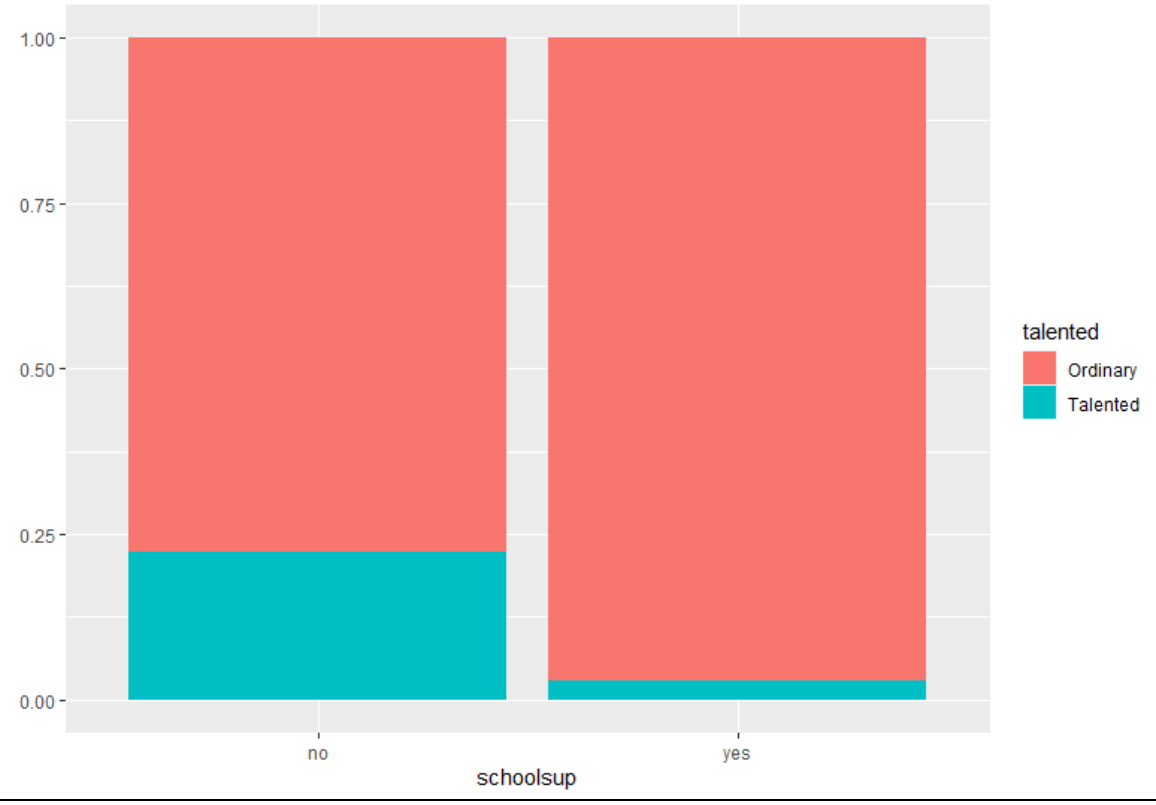
Sex

IV = 0.04. Typ zmiennej: binarna, gdzie F – kobieta, M – mężczyzna. Widać nieco zwiększony udział kobiet w klasie pozytywnej, jednak IV nie sugeruje zbyt dużego wpływu tej zminnej na predykcję.



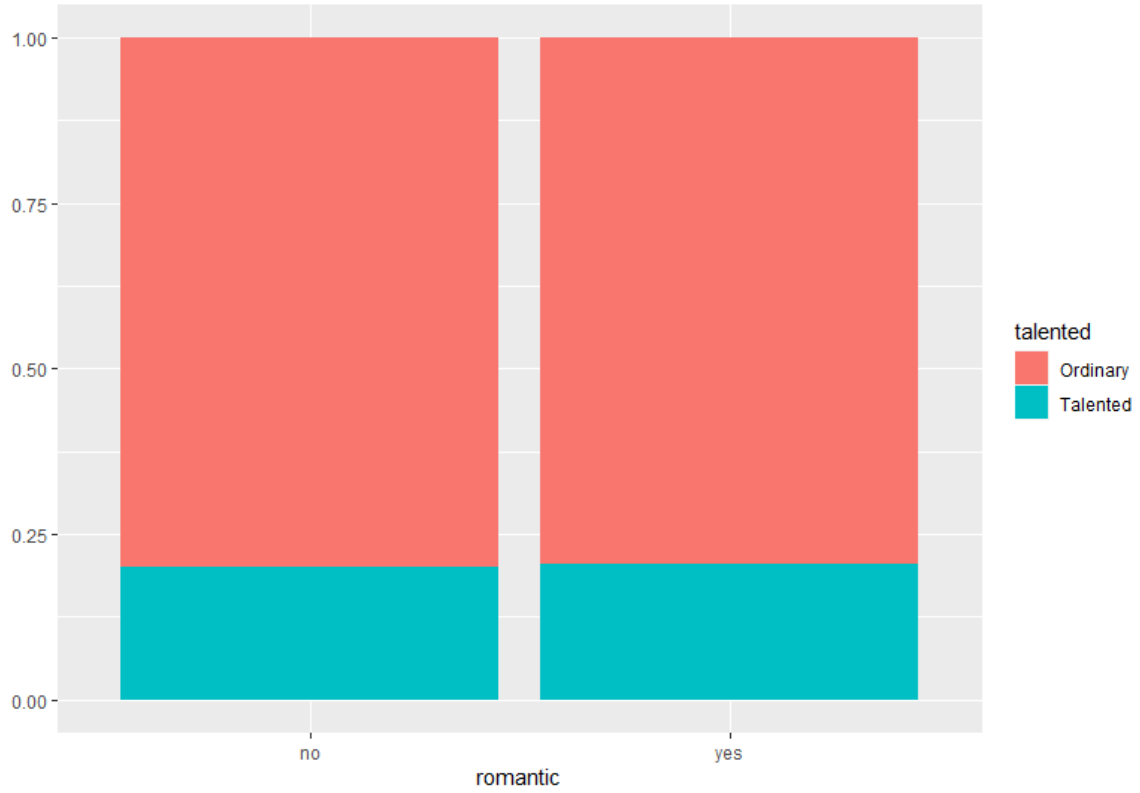
Schoolsup

IV = 0.25 Typ zmiennej: binarna. Zmienna obrazuje, czy uczeń otrzymuje dodatkowe wsparcie w nauce. Widać, że uczniowie otrzymujący takie wsparcie, znacznie rzadziej wykazują wysokie wyniki w nauce. Może to być skutkiem tego, że wsparciem tym otacza się głównie słabszych uczniów. Wysoka wartość IV wskazuje na przydatność tej zmiennej w predykcji.



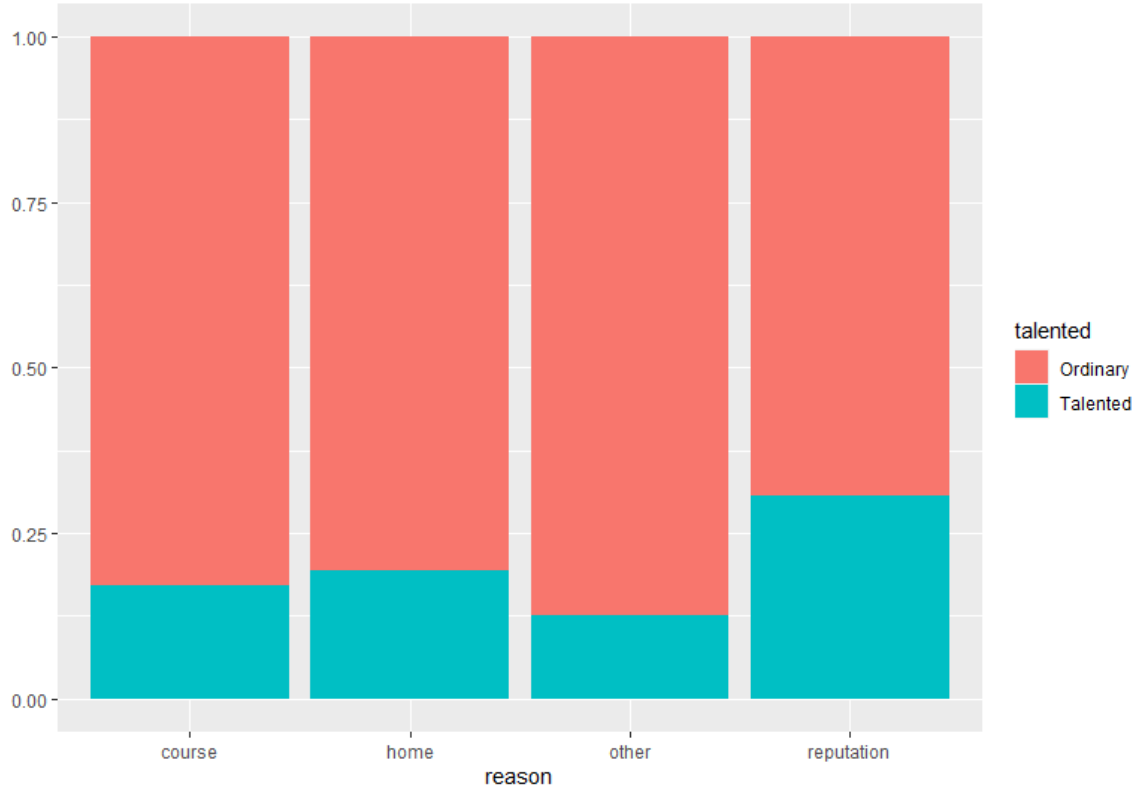
Romantic

IV < 0.01. Typ zmiennej binarna. Zmienna obrazuje, czy badana jednostka znajduje się w związku. Nie widać istotnego wpływu na zmienną objaśnianą, co potwierdza IV.



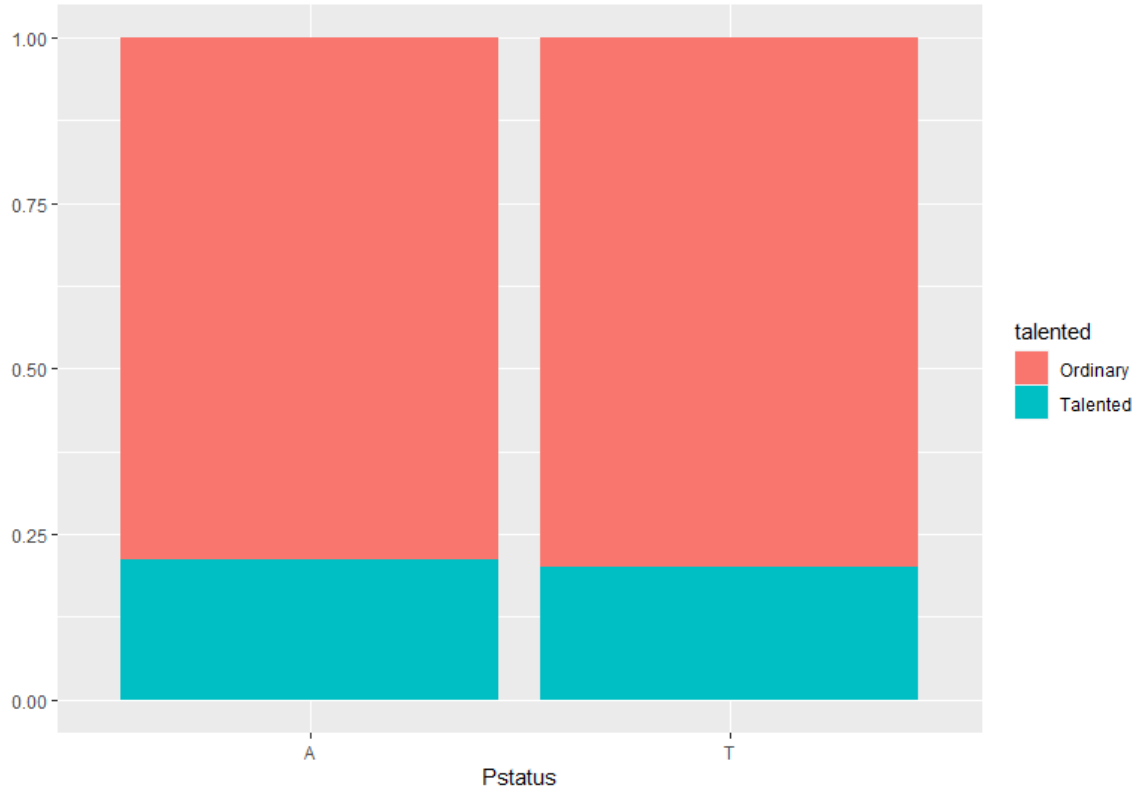
Reason

IV = 0.13. Typ zmiennej: nominalna, gdzie course – preferencja kierunku, home – odległość od domu, reputation – reputacja szkoły, other – inne. Zmienna obrazuje powody wybrania szkoły, w której odbyło się badanie. Widać, że osoby, które wybrały tę szkołę ze względu na jej reputację, osiągają znacznie częściej bardzo dobre wyniki (i nie są to obserwacje odstające!). Ponadto osoby, które wybrały odległość od domu jako powód wyboru, rzeczywiście dojeżdżają krócej (traveltime), co potwierdził test t dla dwóch prób niezależnych (podział ze względu na wartość reason, liczona średnia z traveltime). Przeciętna wartość IV może wskazywać na to, że któryś z powodów rzeczywiście może mieć wysoki wpływ na wynik predykcji.



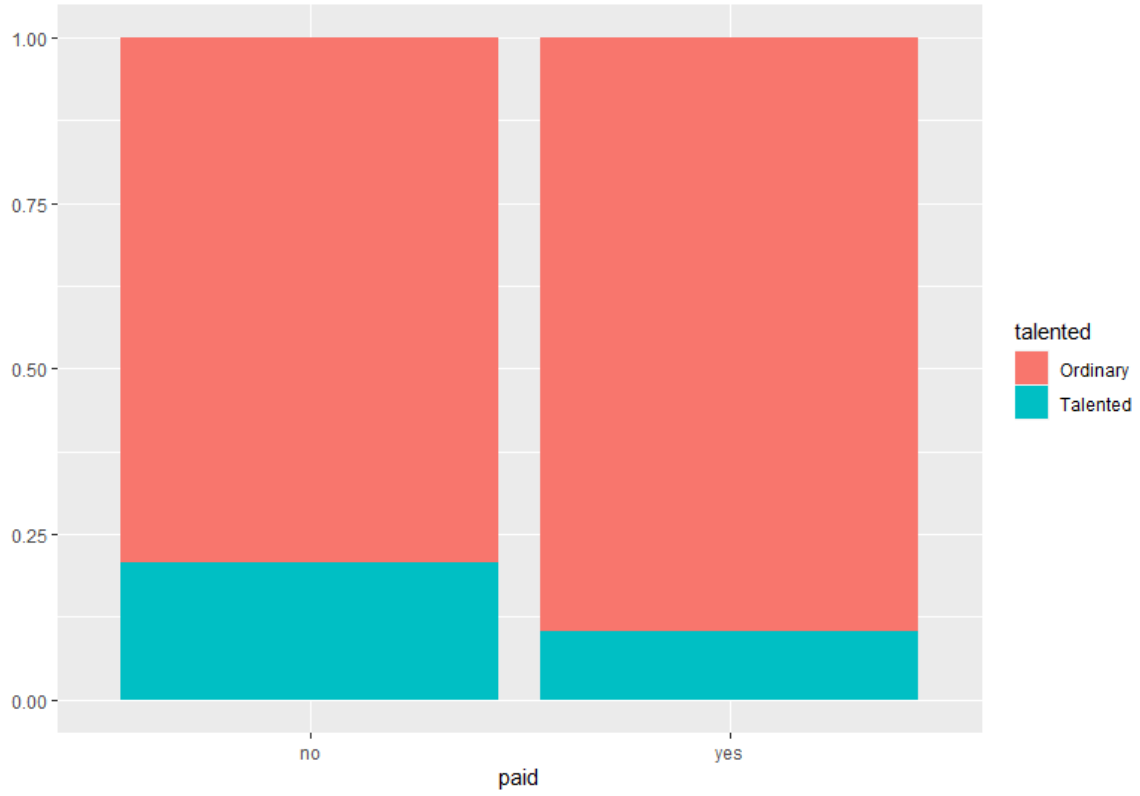
Pstatus

IV < 0.01 Typ zmiennej: binarna, gdzie A- osobno, T – razem. Zmienna obrazuje, czy rodzice respondenta żyją razem, czy osobno. Nie widać istotnych różnic w udziale klasy pozytywnej, co potwierdza IV.



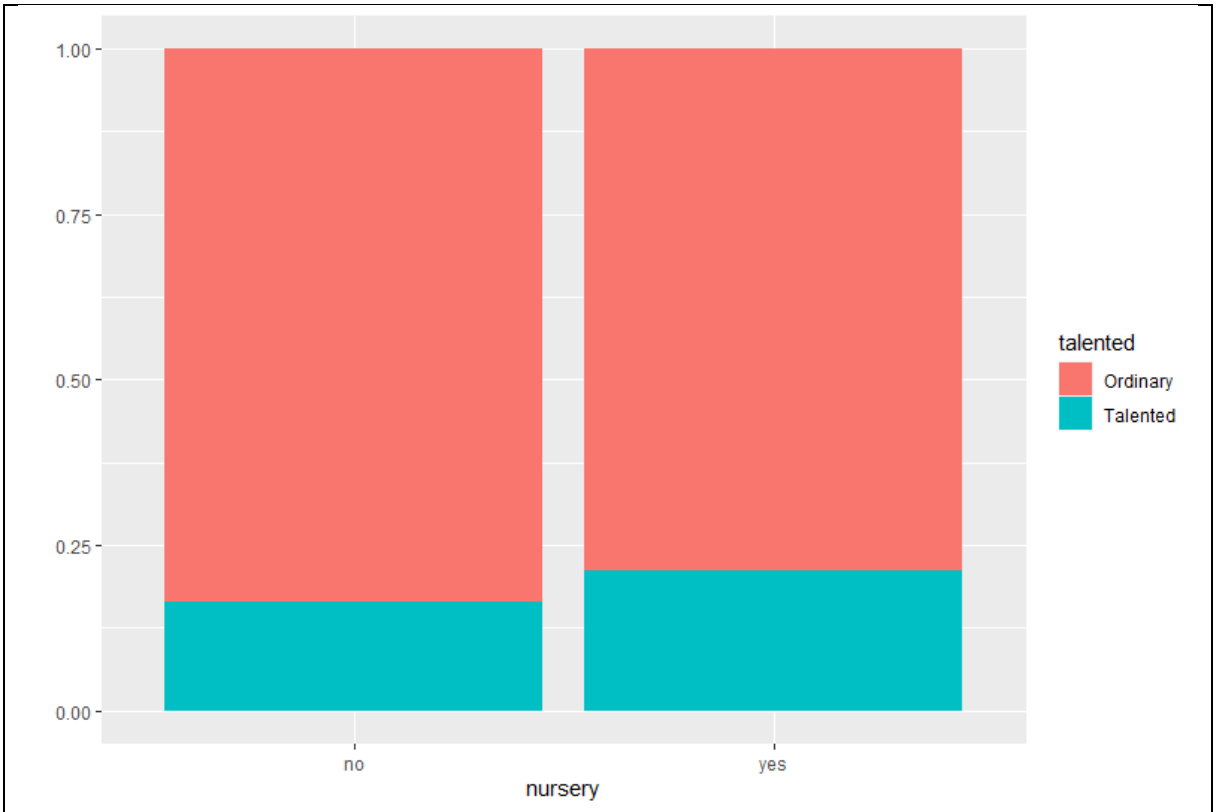
Paid

IV = 0.03. Typ zmiennej: binarna. Zmienna obrazuje, czy uczeń pobiera płatne korepetycje z analizowanego przedmiotu. Podobnie jak w przypadku zmiennej schoolsup widać mniejszy udział klasy pozytywnej dla osób pobierających dodatkowe zajęcia, jednak siła predykcyjna tej zmiennej jest niższa. Może to wynikać z mniejszej liczebności odpowiedzi ‘yes’ dla zmiennej paid niż dla zmiennej schoolsup oraz większego zróżnicowania udziału klasy pozytywnej pomiędzy wartościami ‘yes’ i ‘no’ odpowiednich zmiennych.



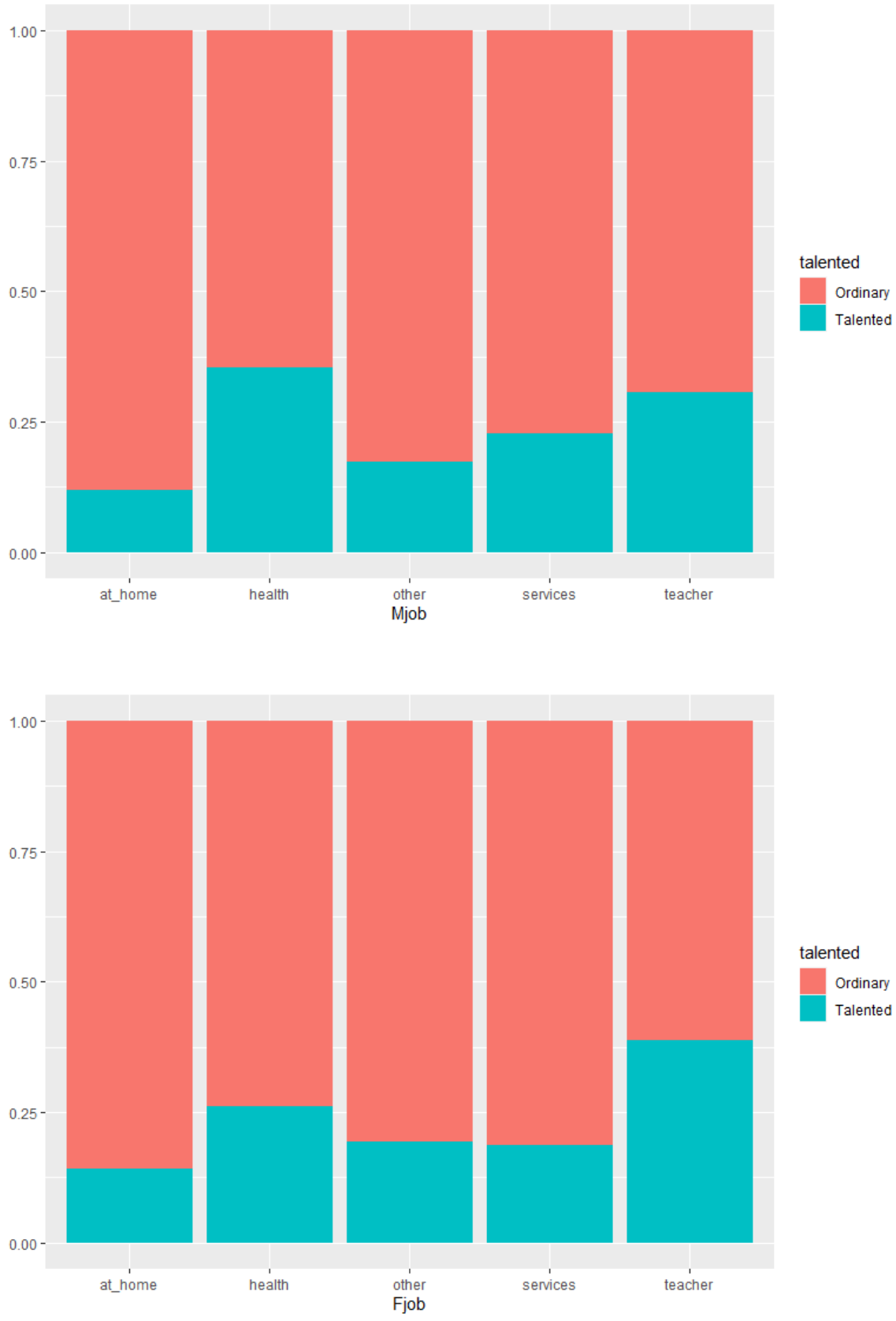
Nursery school

IV = 0.01. Typ zmiennej: binarna. Zmienna pokazuje, czy uczeń uczęszczał w przeszłości do przedszkola. Widać nieco większy udział klasy pozytywnej wśród osób, które do przedszkola uczęszczały, nie wydaje się być on jednak istotnie większy, co potwierdza IV.



Mjob, Fjob

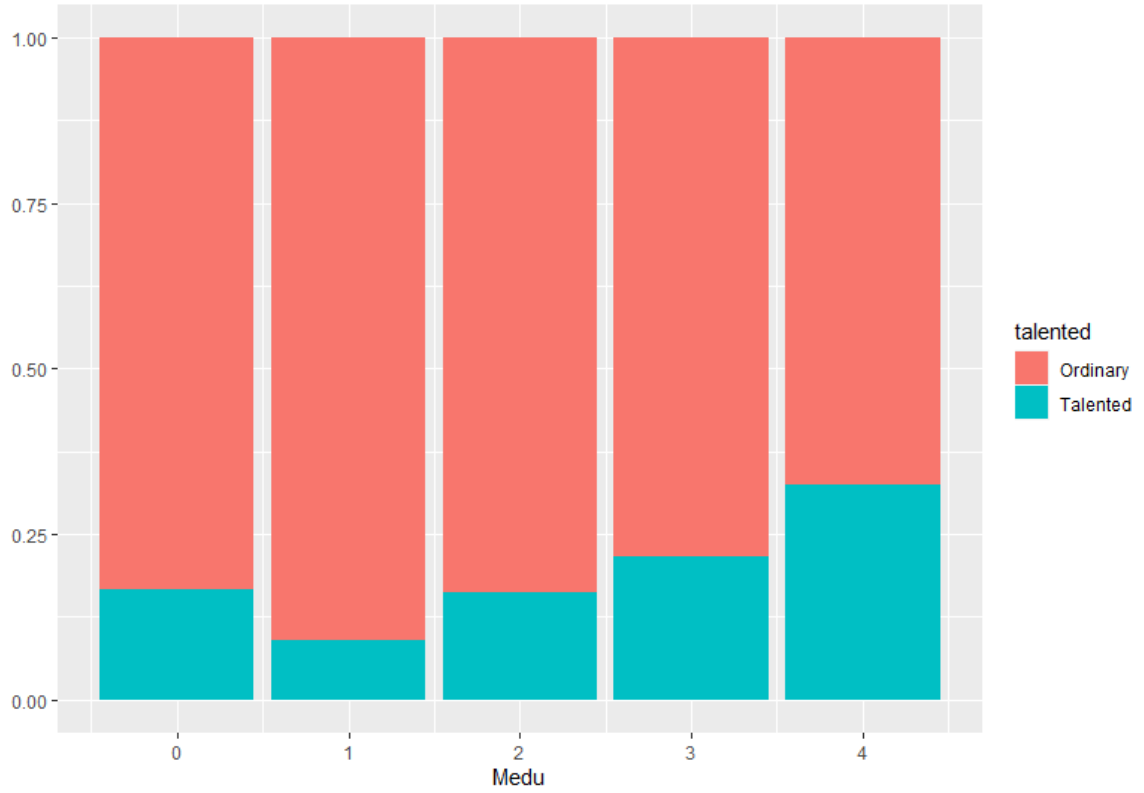
IV(Mjob) = 0.18, IV(Fjob)=0.08. Typy zmiennych: nominalne, gdzie: teacher – nauczyciel, health – służba zdrowia, civil – służba cywilna (np. policja) i administracja, at_home – bezrobotna, other – inne. Zmienne opisujące branżę lub zawód, w którym pracuje matka (Mjob) lub ojciec (Fjob) ucznia. Widać zwiększony odsetek klasy pozytywnej wśród dzieci nauczycieli i pracowników służby zdrowia. Może to być spowodowane tym, że praca w tych zawodach wymaga wykształcenia, a więc ich dzieci również są lepsze w nauce. Przypuszczenia te częściowo potwierdzają umiarkowane wartości IV.

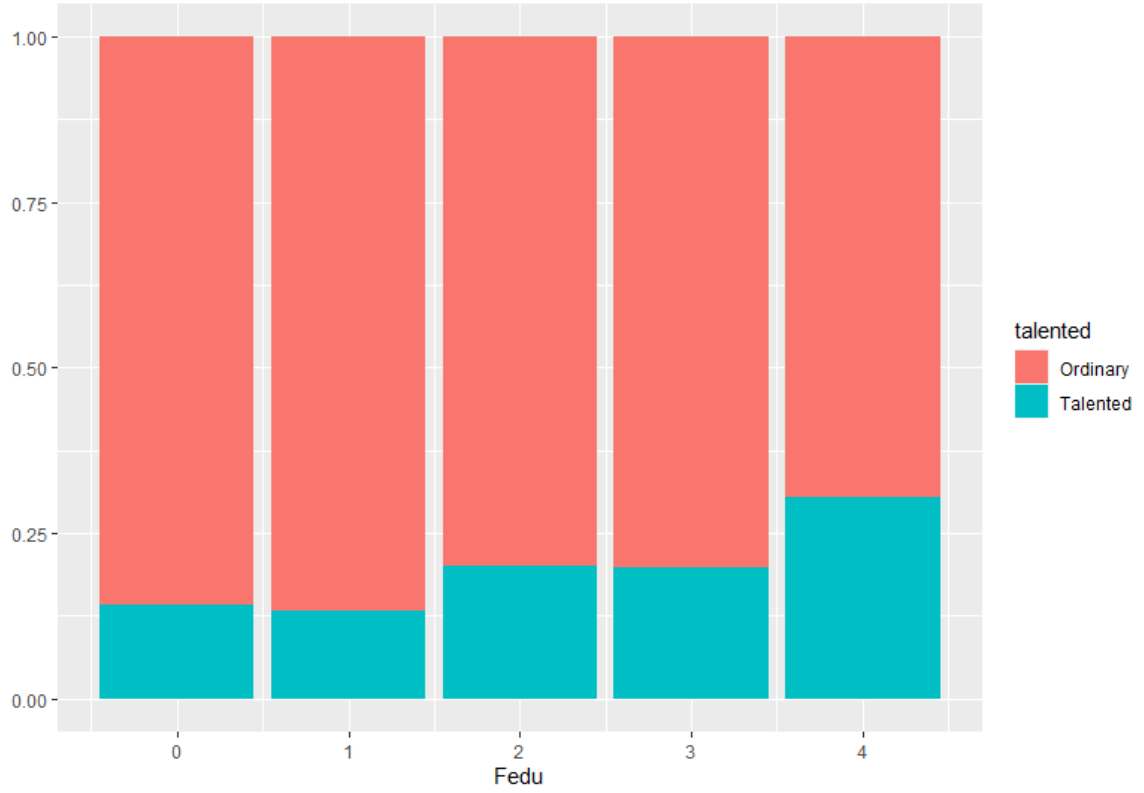


Medu, Fedu

IV(Medu) = 0.3, IV(Fedu) = 0.13. Zmienne porządkowe, gdzie 0 – brak, 1 – podstawowe (do 4 klasy), 2 – podstawowe (5-9 klasa), 3 – średnie, 4 – wyższe. Zmienne obrazujące wykształcenie matki (Medu) lub ojca (Fedu) ucznia. Widać wyraźną dodatnią korelację między stopniem wykształcenia a

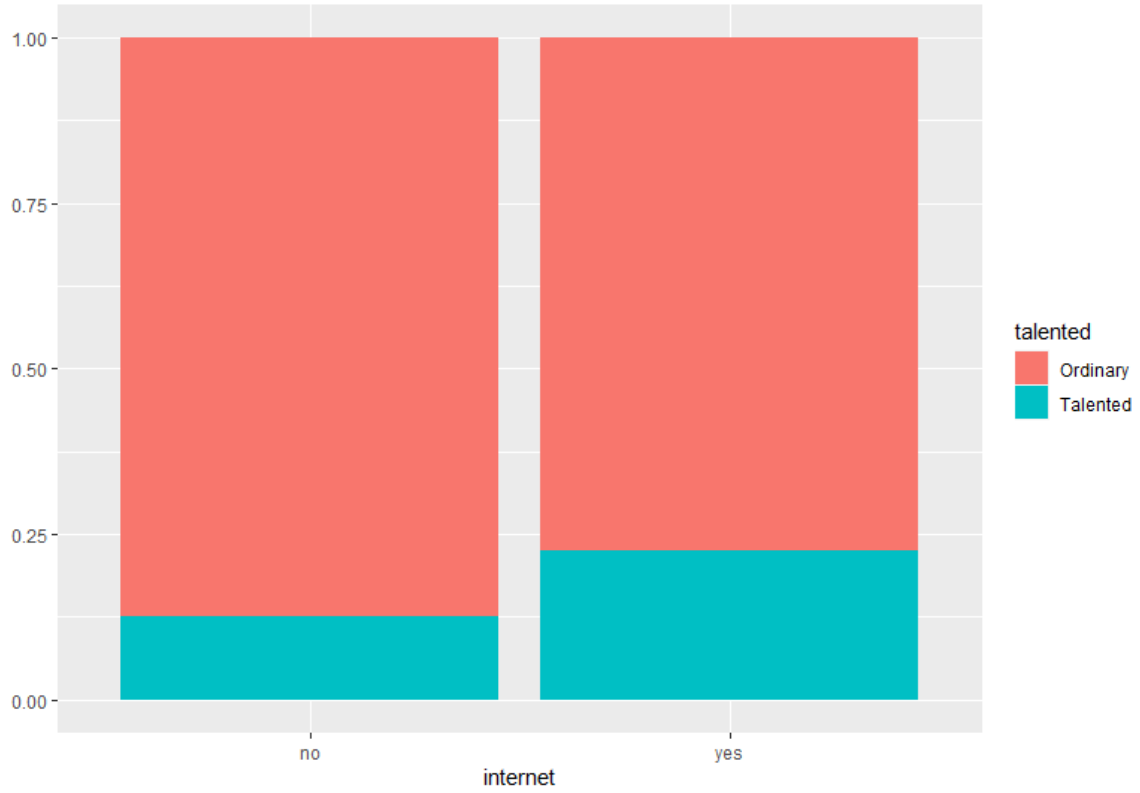
udziałem klasy pozytywnej, co potwierdza również wniosek wyciągnięty dla zmiennych Mjob i Fjob. Wartości IV sugerują przydatność omawianych zmiennych (w szczególności Medu) w procesie predykcji.





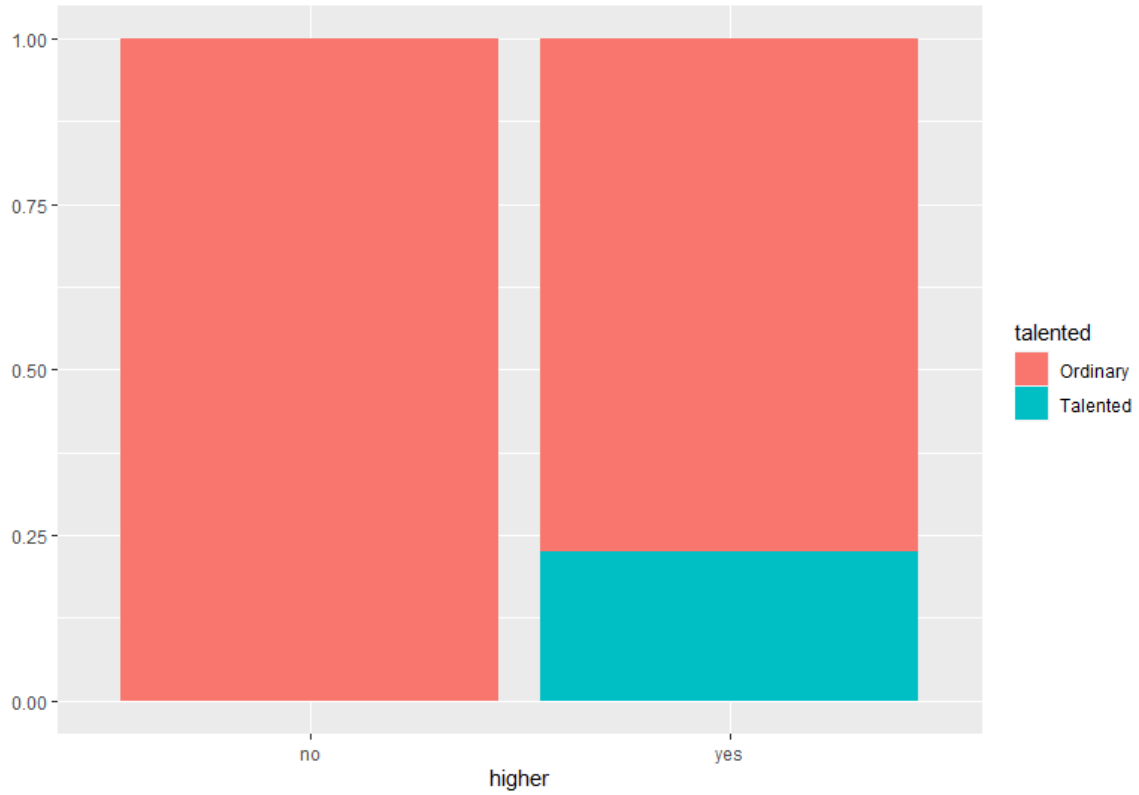
Internet

IV = 0.08. Typ zmiennej: binarna. Zmienna obrazująca, czy uczeń ma w domu dostęp do internetu. Widać zwiększony udział klasy pozytywnej wśród osób, które mają dostęp do internetu, IV jednak sugeruje, że nie jest to bardzo istotny czynnik.



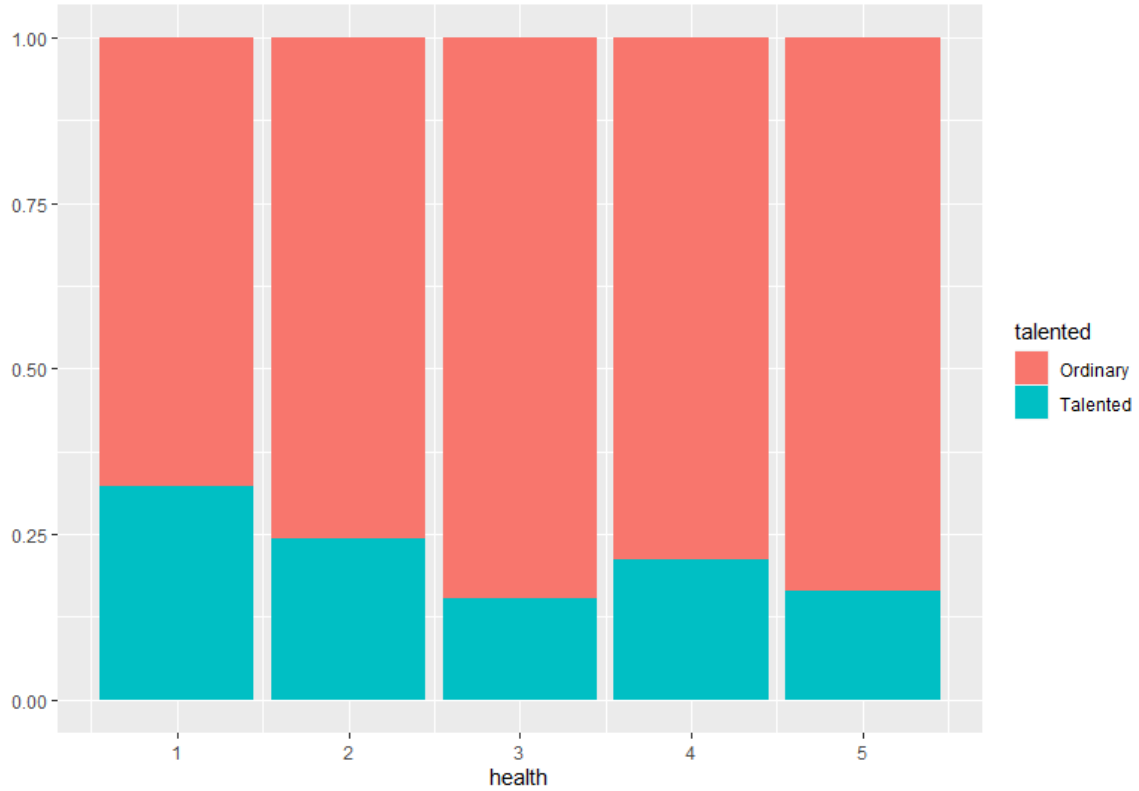
Higher

IV = 0.38. Typ zmiennej: binarna. Zmienna obrazująca, czy uczeń chce w przyszłości podjąć edukację wyższą. Widać, że osoby, które nie chcę jej podejmować, prawie nigdy nie osiągają wysokich wyników. Można to tłumaczyć tym, że skoro nie są w stanie osiągać choćby średnich wyników, to i tak nie mają szans na podjęcie takiej edukacji, bądź też nie zależy im na wynikach, skoro i tak nie zamierzają tej edukacji podjąć. Obie kategorie mają duże liczebności. IV przyjmuje bardzo wysoką wartość, co potwierdza spostrzeżenie dotyczące rozkładu danych na wykresie.



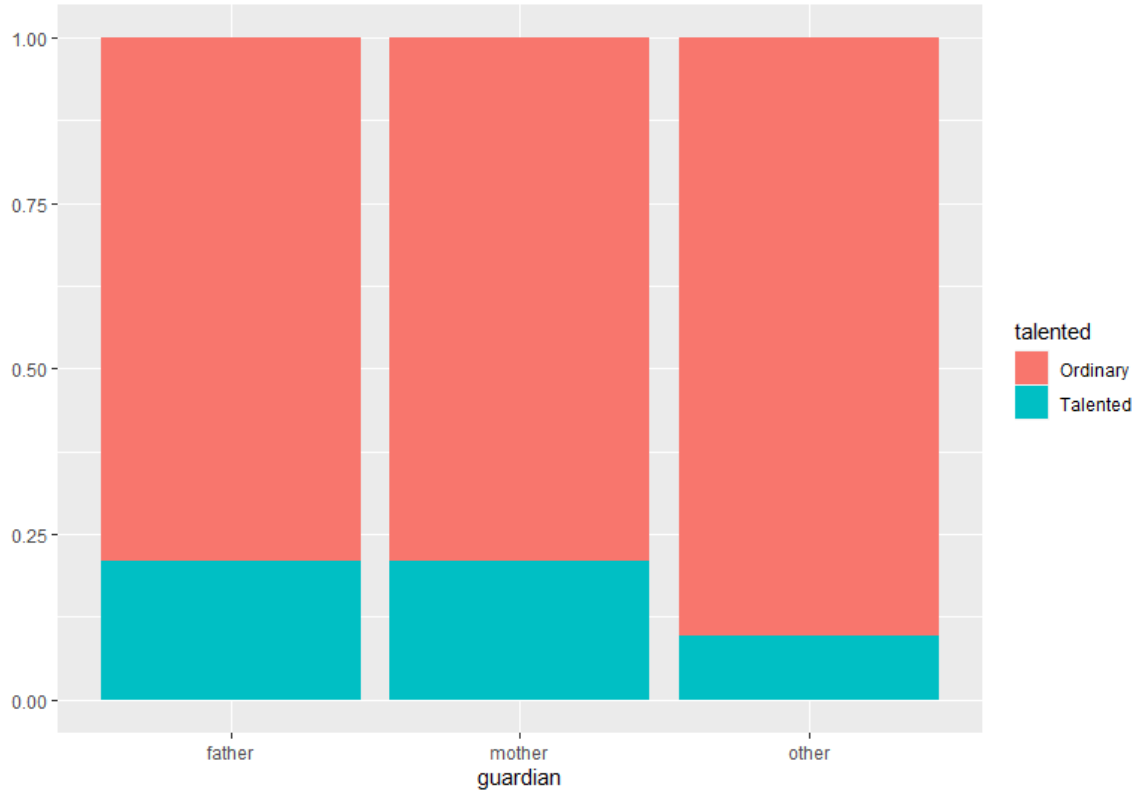
Health

IV = 0.12. Typ zmiennej: porządkowa. Zmienna obrazująca ocenę swojego stanu zdrowie, gdzie 1 – bardzo zły, 5 - bardzo dobry. Około 40% ankietowanych określiła stan swojego zdrowia jako bardzo dobry, wystarczające liczebności zostały zachowane również dla pozostałych liczebności. IV wskazuje na przeciętną wartość predykcyjną tej zmiennej.



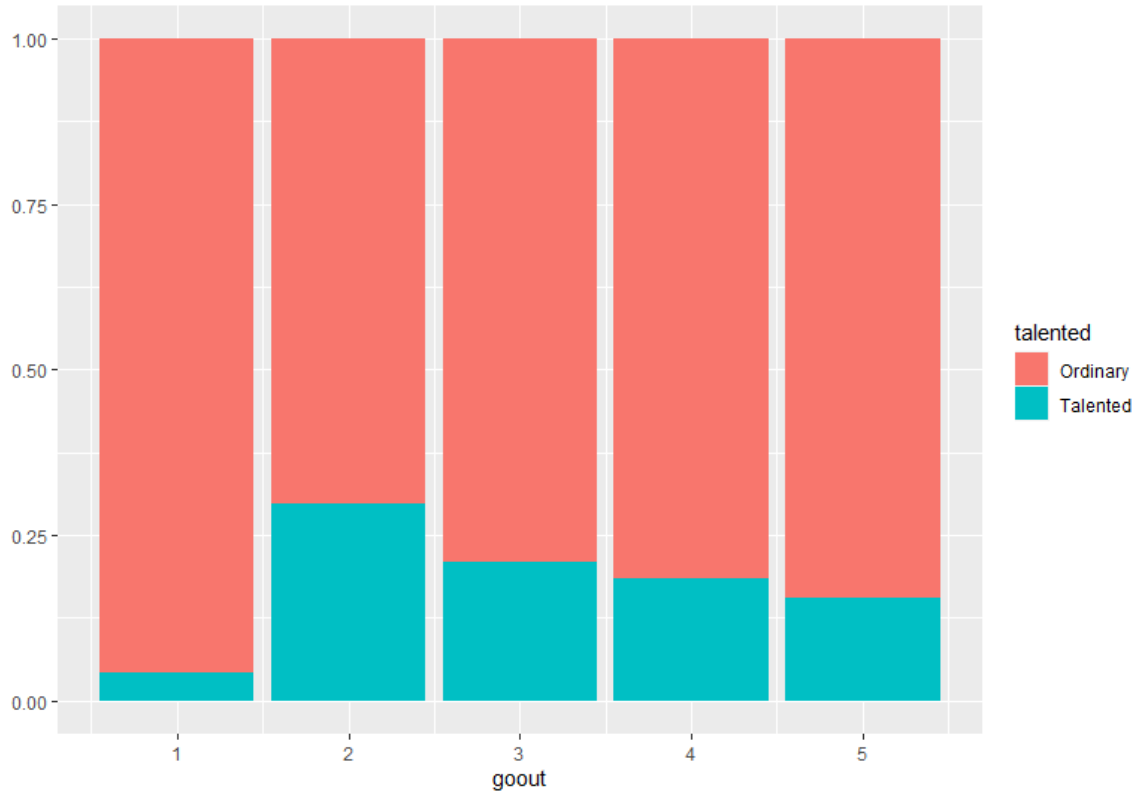
Guardian

IV = 0.04. Typ zmiennej: nominalna. Zmienna obrazująca, kto został wskazany jako główny opiekun ucznia w domu. IV wskazuje na bardzo małą wartość predykcyjną tej zmiennej



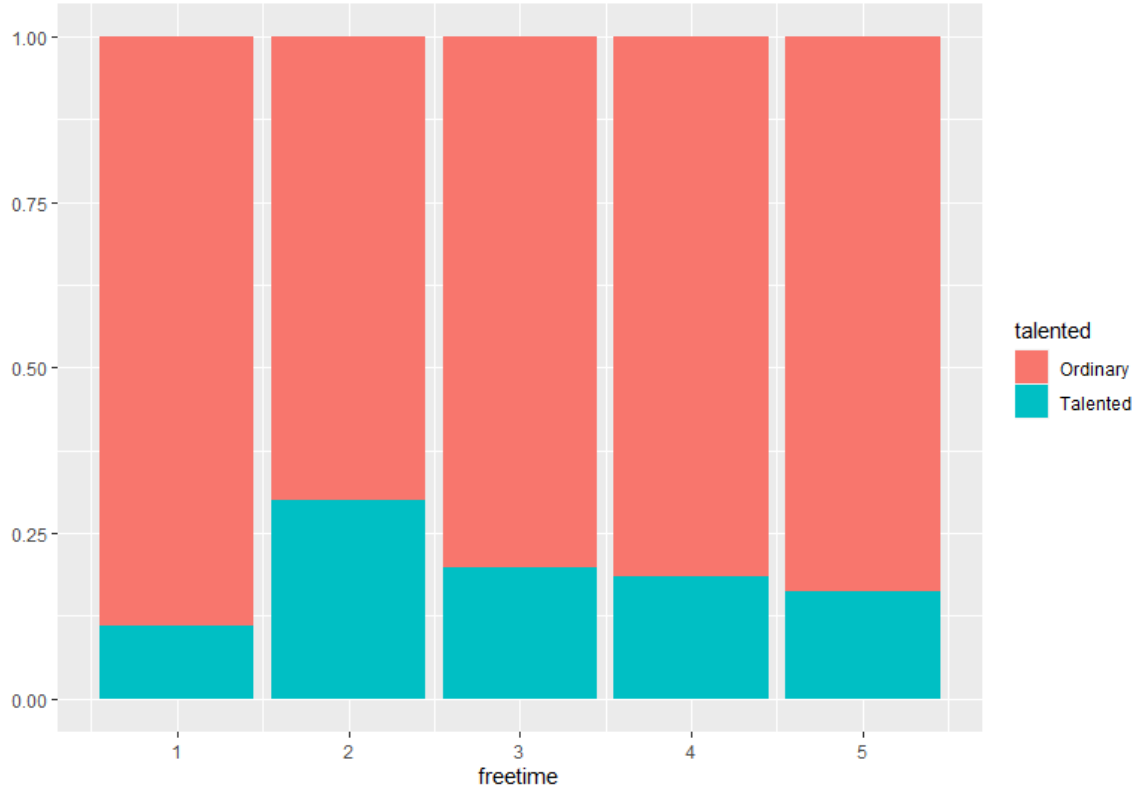
Goout

IV = 0.22. Typ zmiennej: porządkowa, gdzie 1- bardzo rzadko, 5- bardzo często. Zmienna prezentująca jak często uczeń wychodzi na spotkania z przyjaciółmi. Poza mniejszym udziałem klasy pozytywnej dla osób wychodzącym bardzo rzadko, widać niewielki trend spadkowy wraz z wzrostem częstotliwości wyjść. Istotne liczebności zostały zachowane. Wartość IV jest dość wysoka.



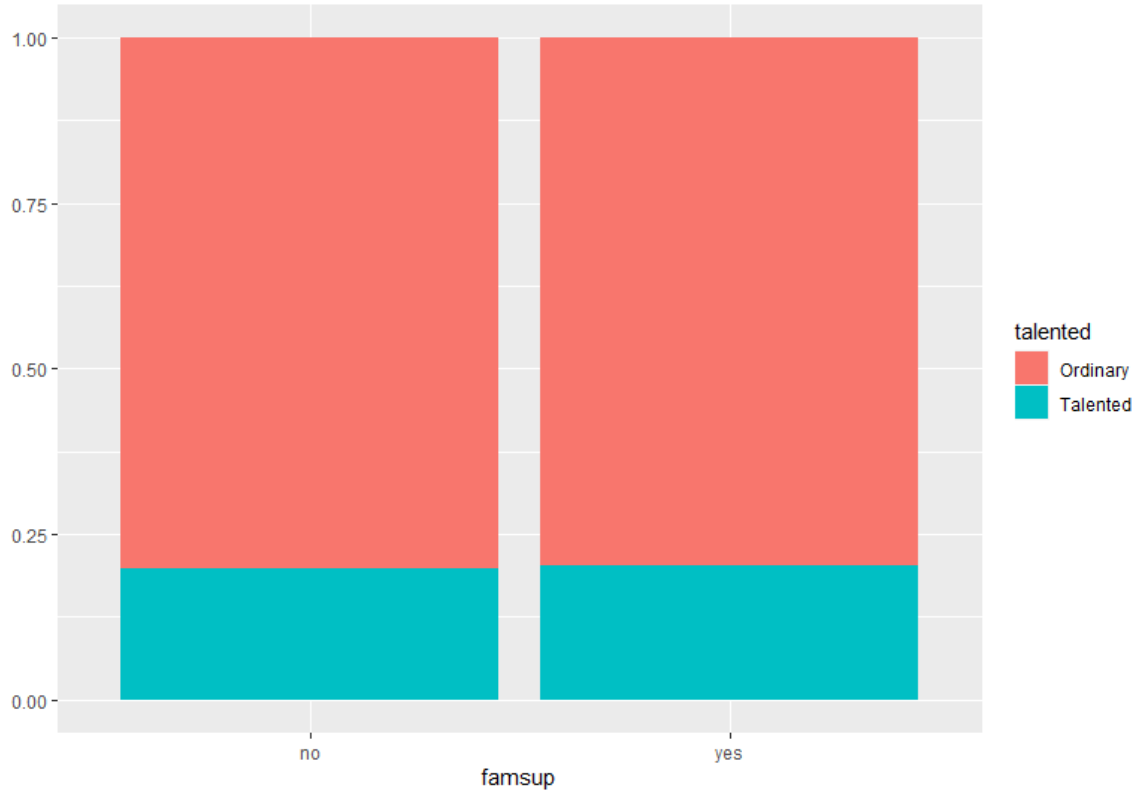
Freetime

IV = 0.09. Typ zmiennej: porządkowa, gdzie 1 – bardzo mało, 5 – bardzo dużo. Zmienna obrazująca jak uczeń ocenia ilość wolnego czasu po zajęciach. Rozkład udziały klasy pozytywnej jest podobny do tego dla zmiennej goout, dodatnią korelacje pokazuje również tabela krzyżowa. Może to być skutkiem tego, że w wolnym czasie uczniowie wychodzą się spotykać z przyjaciółmi. Wartość IV jest jednak dla tej zmiennej niższa (ale nie bardzo niska). Ponadto istnieje ujemna korelacja pomiędzy zmienną studytime a freetime, której istotność potwierdził test chi-kwadrat (p-value<0.05).



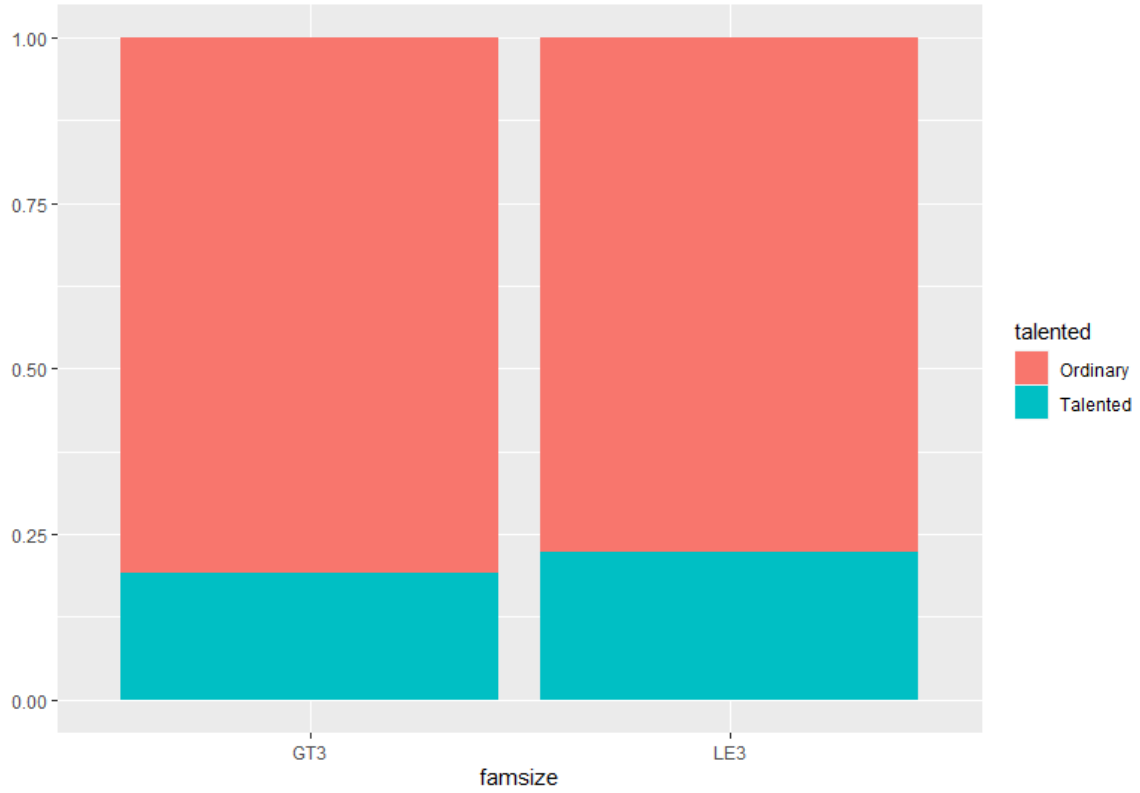
Famsup

IV < 0.01. Typ zmiennej: binarna. Zmienna obrazująca, czy uczeń otrzymuje wsparcie w nauce od członków rodziny. Rozkład udziału klas jest niemalże jednolity dla obu wartości tej zmiennej. Brak istotności dla predykcji potwierdza IV.



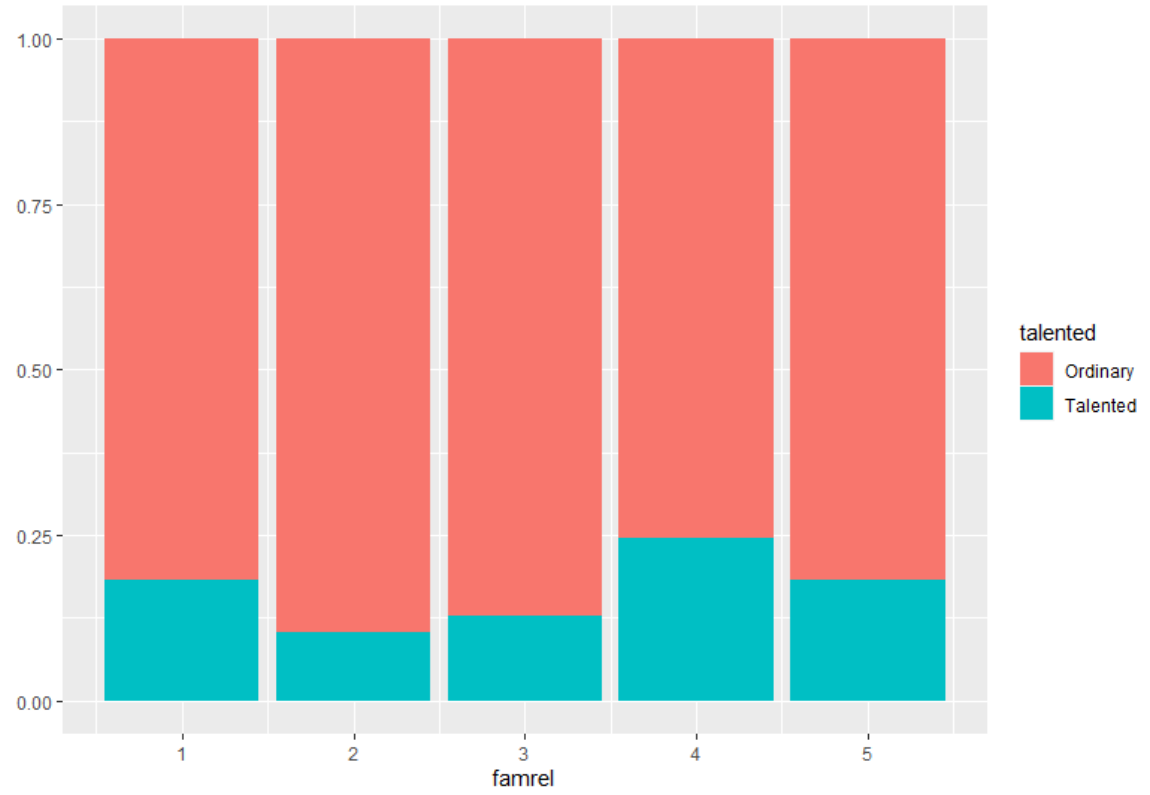
Famisze

IV < 0.01. Typ zmiennej: binarna, gdzie GT3 – więcej niż 3, LS3 – równo lub mniej niż 3. Zmienna obrazująca liczbę członków gospodarstwa domowego ucznia. Rozkład udziału klas jest niemalże jednolity dla obu wartości tej zmiennej. Brak istotności dla predykcji potwierdza IV.



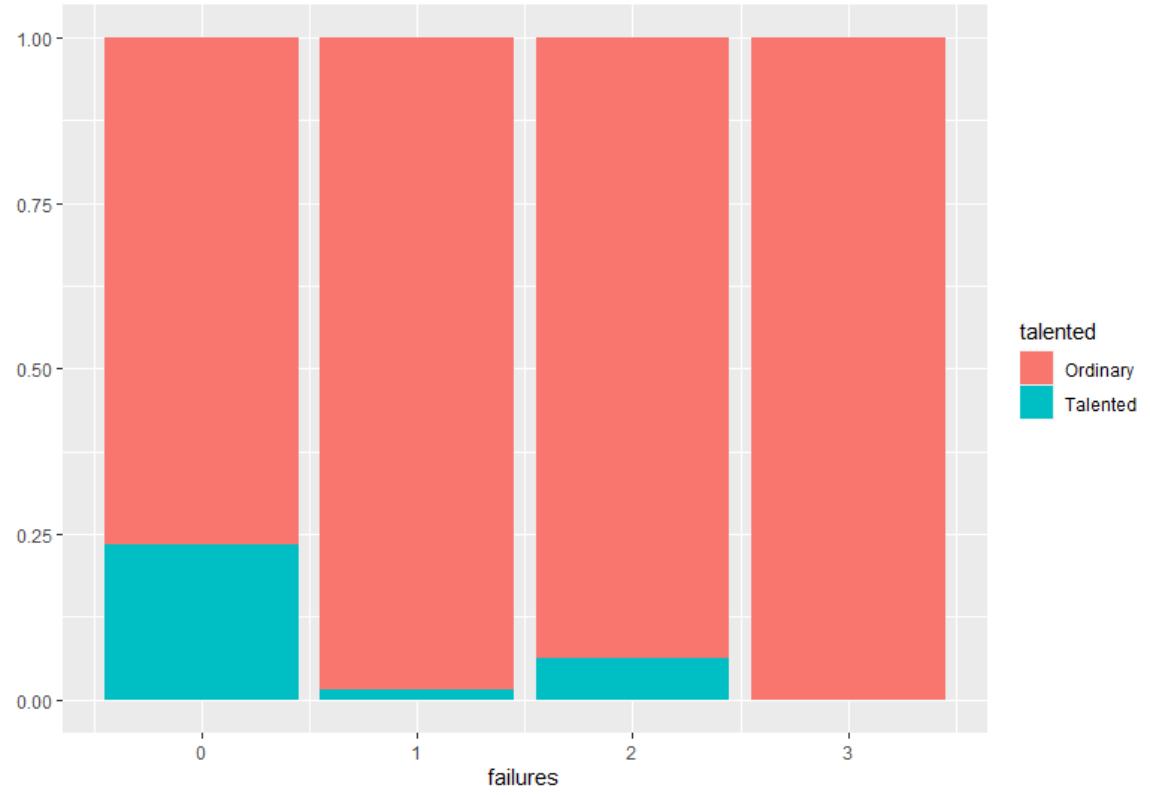
Famrel

IV = 0.1. Typ zmiennej: porządkowa, gdzie 1 – bardzo złe, 5 – bardzo dobre. Zmienna obrazująca ocenę relacji rodzinnych. Widać, że uczniowe z rodzin o dobrych (4) lub bardzo dobrych (5) relacjach osiągają nieco częściej wysokie wyniki. IV pokazuje przeciętne znaczenie zmiennej dla predykcji.



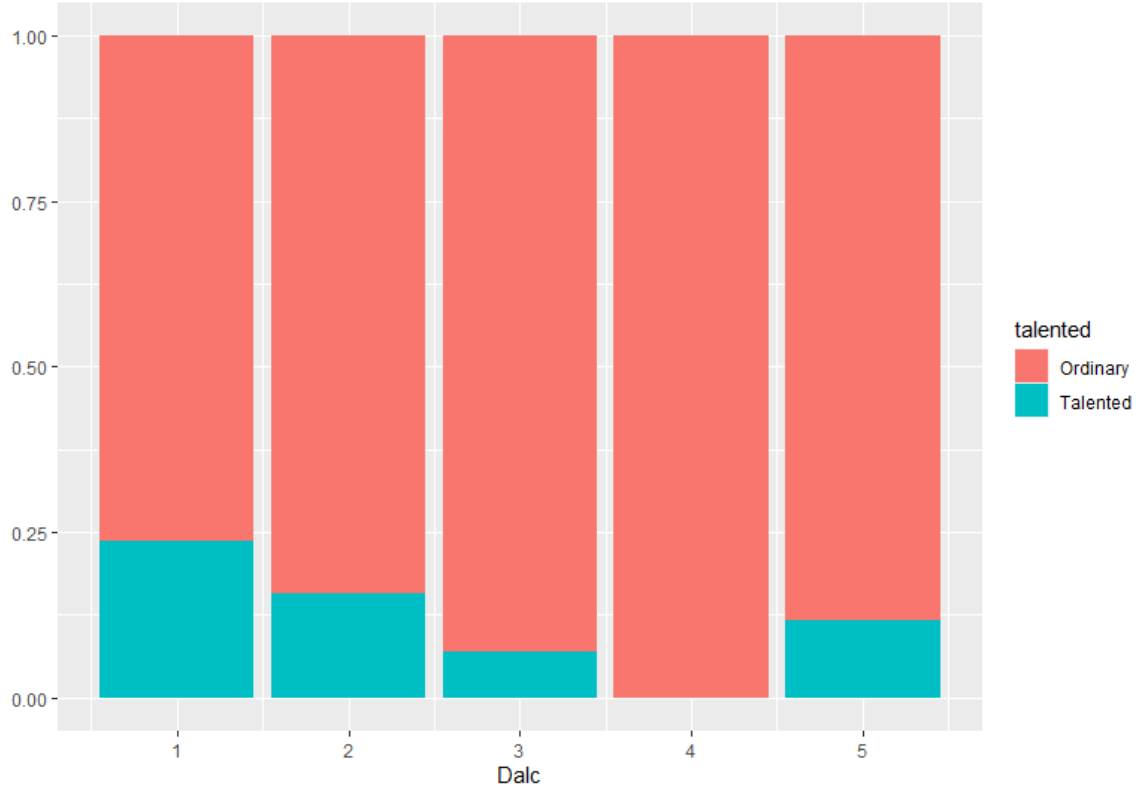
Failures

IV = 0.44. Typ zmiennej: porządkowa (zbliżona do liczbowej dyskretnej). Zmienna ukazująca ile razy uczeń w przeszłości nie zdał (gdzie 3 oznacza 3 lub więcej razy). Widać, że na wysokie wyniki można liczyć głównie wśród uczniów, którym zawsze udawało się zdać. Przydatność tej zmiennej w predykcji potwierdza bardzo wysoka wartość IV.



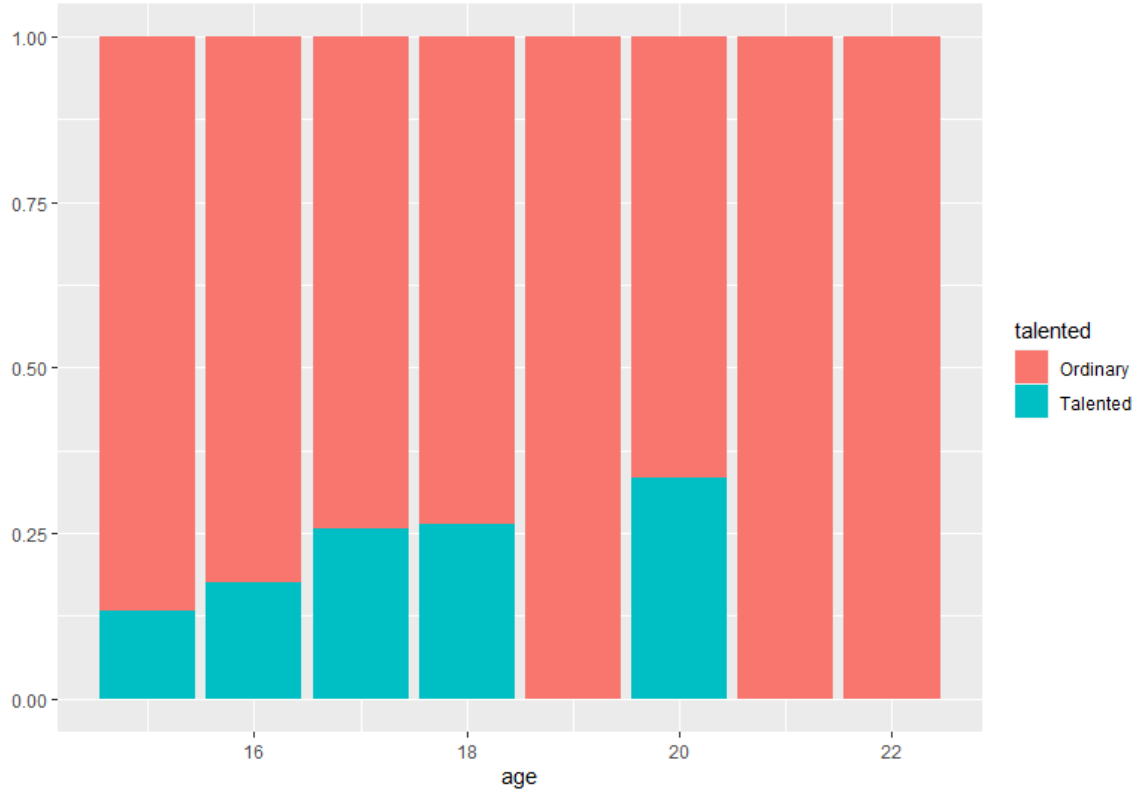
Dalc

IV = 0.16. Typ zmiennej: porządkowa, gdzie 1 – bardzo niskie, 5 – bardzo wysokie. Zmienna ukazująca spożycie alkoholu w dni powszednie. Widać wyraźnie ujemną korelację między ilością spożywanego alkoholu w dni powszednie a szansą na osiąganie wysokich wyników. IV pokazuje przeciętną użyteczność tej zmiennej w predykcji.



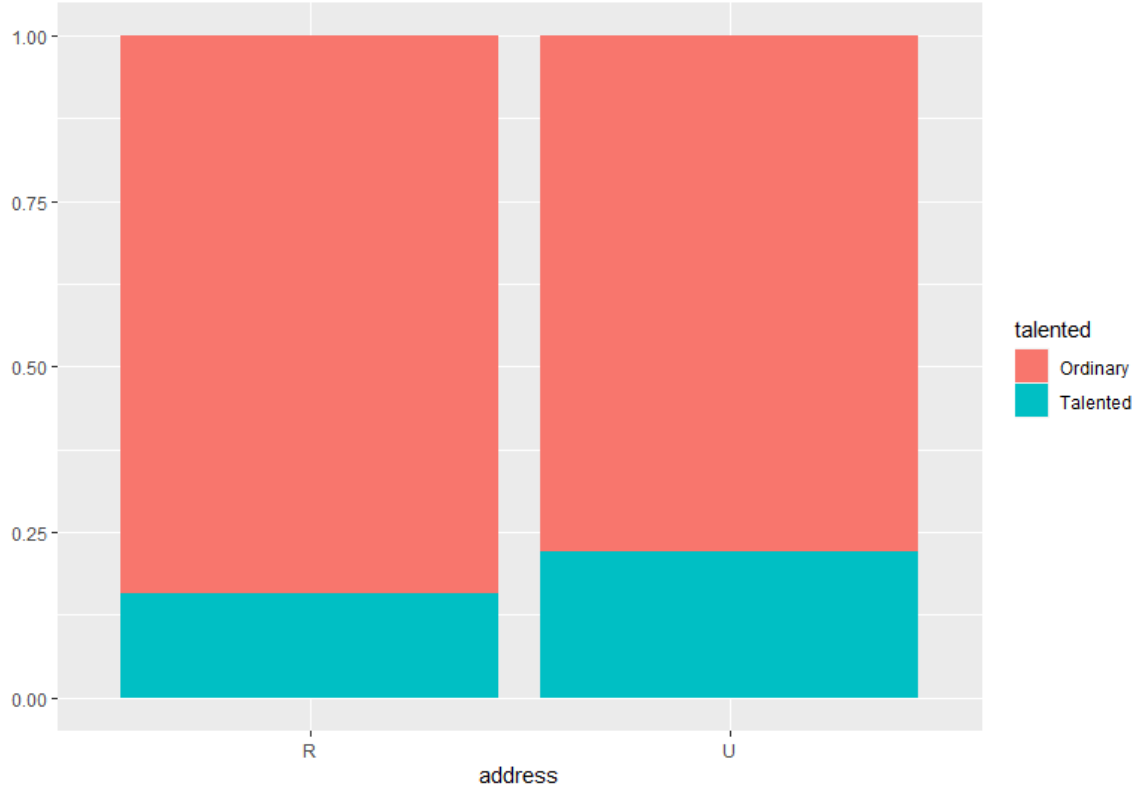
Age

IV = 0.23. Typ zmiennej: liczbowa dyskretna. Zmienna obrazująca wiek respondenta**.** W wieku 15-18 lat otrzymano zadowalające liczebności odpowiedzi (powyżej 100 respondentów w każdej kategorii). Osób w wieku 19 lat jest dokładnie 32, co również jest zadowalającą liczebnością. Osób powyżej 19 roku życia jest 9, a więc wyniki w trzech ostatnich kolumnach można uznać za obserwacje odstające. Tabele krzyżowe ze zmienną failures pokazały, że osoby w wieku 19 lat, to głównie osoby, które nie zdały, a więc niski udział klasy pozytywnej dla age=19 można tłumaczyć tymże faktem. Wartość IV sugeruje dużą przydatność tej zmiennej w predykcji, ale należy pamiętać o zawyżeniu wynikającym z dużej liczby zbioru wartości zmiennej objaśniającej.



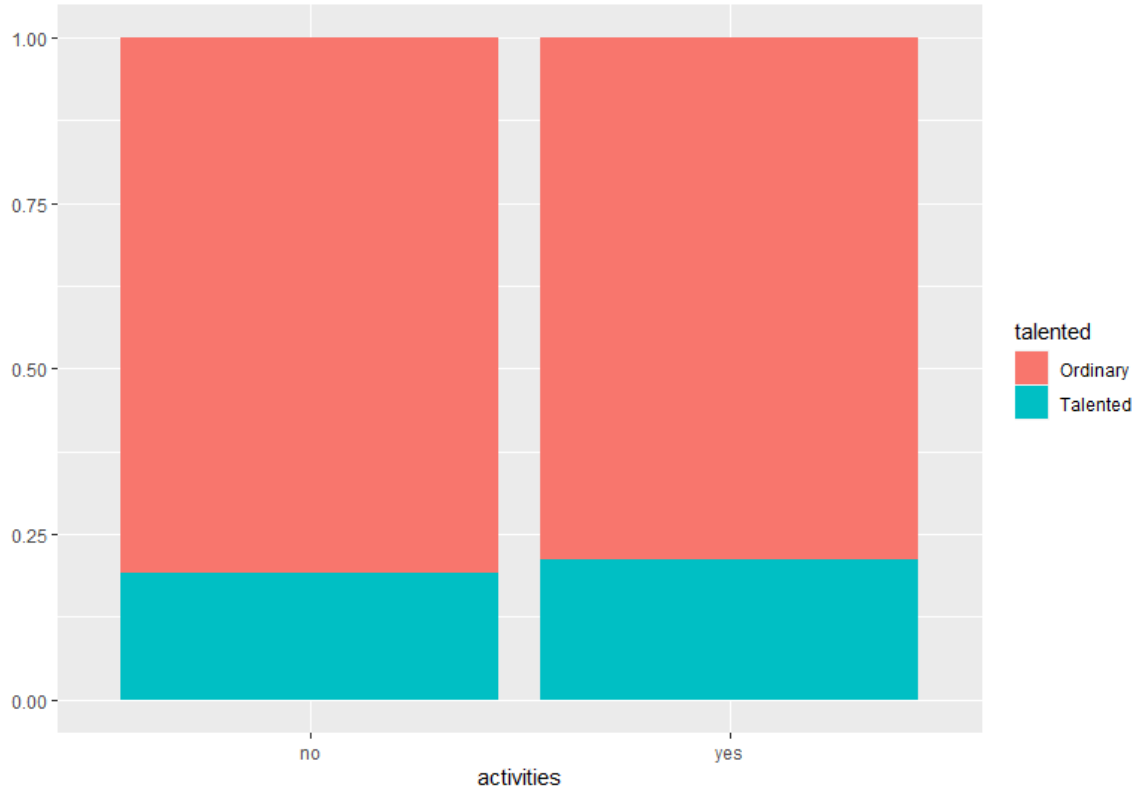
Address

IV = 0.04. Typ zmiennej: binarna, gdzie R – wieś, U – miasto. Zmienna obrazująca miejsce zamieszkania ucznia. Widać, że uczniowie z terenów miejskich nieco częściej osiągają wysokie wyniki niż osoby z terenów wiejskich. Niewielkie znaczenie tej różnicy potwierdza niska wartość IV.



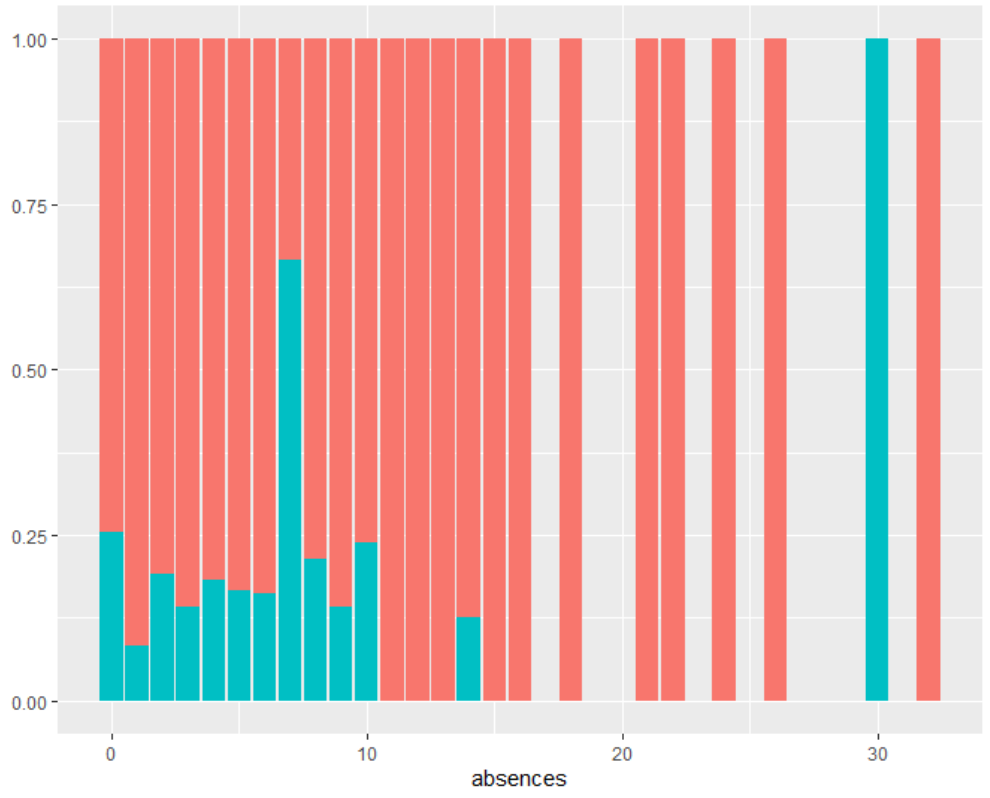
Activities

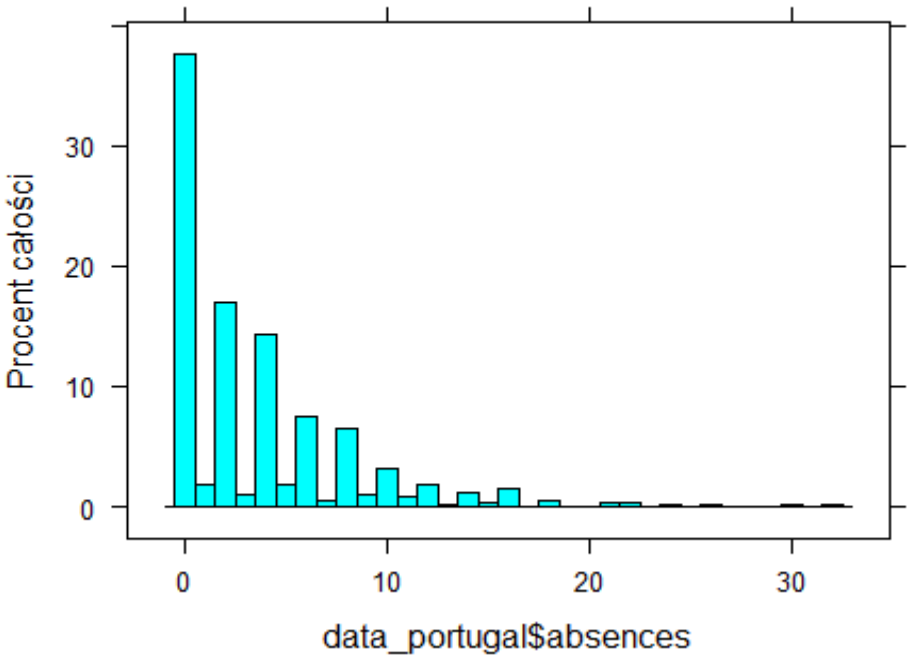
IV < 0.01Typ zmiennej: binarna. Zmienna obrazująca, czy uczeń angażuje się w zajęcia poza programem szkolnym. Widać, że uczniowie angażujący się w taką działalność nieco częściej osiągają wysokie wyniki, wartość IV pokazuje jednak niewielkie znaczenie tejże różnicy w predykcji.



Absences

IV = 0.16. Typ zmiennej: liczbowa dyskretna. Zmienna obrazująca liczbę nieobecności na zajęciach szkolnych. Około 38% uczniów nie opuściło żadnych zajęć, widać zmniejszone liczebności dla wartości nieparzystych (można domyślać się, że powodem tego był układ planu zajęć), można się spodziewać raczej niewielkiej siły predykcyjnej, co sugeruje również histogram i IV (należy pamiętać o zawyżeniu spowodowanym dużą liczbą kategorii zmiennej objaśniającej).


1.00
0.75
0.50
0.25
0.00
0
10
20
30
absences
talented
Ordinary
Talented


Procent całości
30
20
10
0
0
10
20
30
data_portugal$absences

4

3. Konstrukcja modeli

Wyniki w nauce języka portugalskiego uzyskane przez uczniów zostaną zamodelowane z wykorzystaniem nadzorowanego podejścia – binarnej klasyfikacji. Do budowy klasyfikacji został wybrany algorytm lasów losowych

Wszystkie symulacje przedstawione w niniejszym raporcie przeprowadzono z użyciem języka R. W badaniach wykorzystano następujące biblioteki:

- *dplyr* – do manipulacji danymi zapisanymi zarówno w formie ramek danych, jak i przechowywanych w bazach,

- *caret* – do budowy modeli, testowania, wyboru zmiennych,

- *tictoc* – do zagnieżdżenia funkcji czasowych,

- *pROC* – do wizualizacji, wygładzania i porównywania charakterystyk krzywych ROC.

Zbiór danych został podzielony na treningowy i testowy. 75% obserwacji zostało przypisanych do zbioru treningowego, natomiast pozostałe obserwacje (25%) do zbioru testowego. W wyniku podziału, 488 i 161 obserwacji przypisano odpowiednio do grupy treningowej i testowej.

Opracowano dwa modele oparte na algorytmach lasów losowych. W pierwszym modelu optymalne hiperparametry zostały obliczone przy wykorzystaniu algorytmu *grid search*. Z kolei parametry obecne w modelu drugim są wynikiem algorytmu *random search*. W obydwu przypadkach w konstrukcji modeli wykorzystano metodę *ranger*.

Model wykorzystujący algorytm *grid search*

Optymalne hiperparametry uzyskano za pomocą algorytmu *grid search*. W pierwszej kolejności szukano wartości wyrażających maksymalną liczbę zmiennych użytych do budowy drzew, minimalną wielkość liścia oraz metody podziału obserwacji, by następnie znaleźć odpowiednią liczbę drzew.

Do uzyskania wyników predykcyjnych zastosowano metodę 5-krotnej walidacji krzyżowej w trzech seriach.

Wartości optymalnych hiperparametrów wybranych do budowy modelu są następujące:

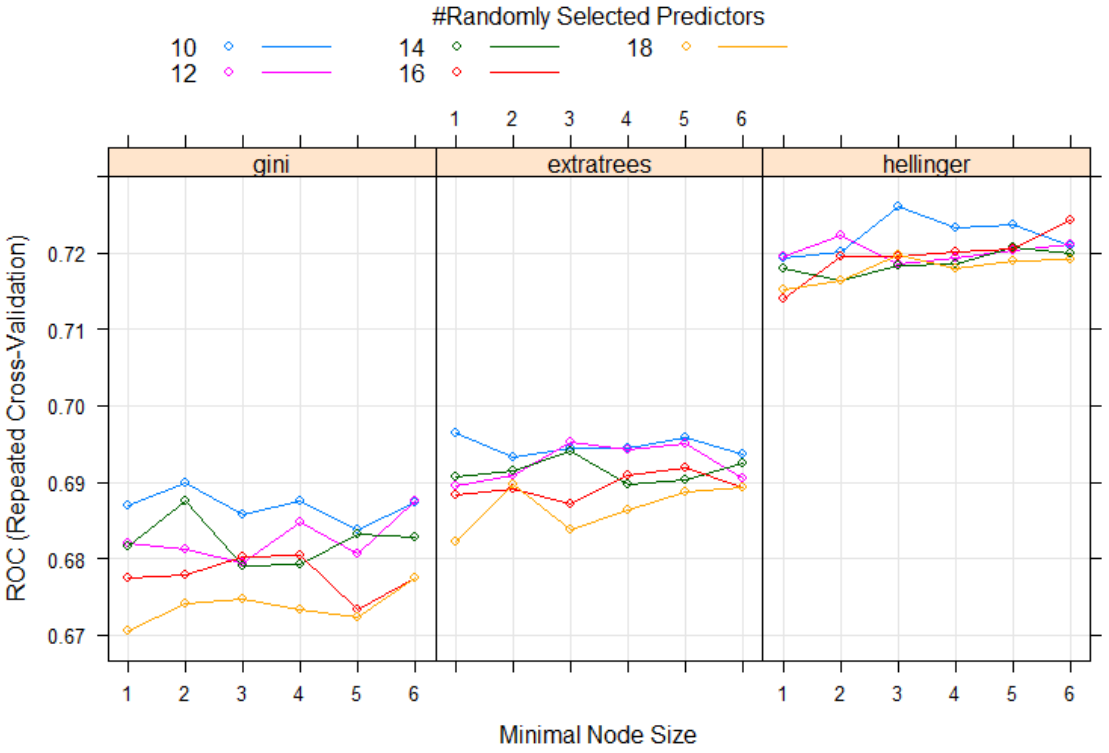
- maksymalna liczba użytych zmiennych = 10,

- metoda podziału = *hellinger,*

- minimalna wielkość liścia = 3.

Wyboru optymalnego modelu dokonano na podstawie porównań krzywych ROC. Rysunek nr 1 przedstawia wszystkie modele brane pod uwagę.

Bez wątpienia metoda podziału *hellinger* pozwala uzyskać znacznie wyższe wartości pola dla krzywych ROC. Modelem odznaczającym się na wykresie jest ten skonstruowany z 10 zmiennych. Pozostałe metody podziału: *gini* oraz *extratrees* są mniej efektywne.



Rysunek 1 – wielkości pól pod wykresem krzywej ROC w zależności od metod podziału i minimalnej wielkości liścia z wykorzystaniem grid search

Większość wartości znajduje się na diagonalnej macierzy błędów, co wskazuje na dobre dopasowanie modelu (patrz: Tabela 1)

Macierz pomyłek

Predykcja	Ordinary	Talented
Ordinary	129	9
Talented	0	23

Tabela 1 - Macierz pomyłek modelu z modelu z metodą podziału "hellinger" z dziesięcioma zmiennymi, wyznaczonego przy użyciu grid search

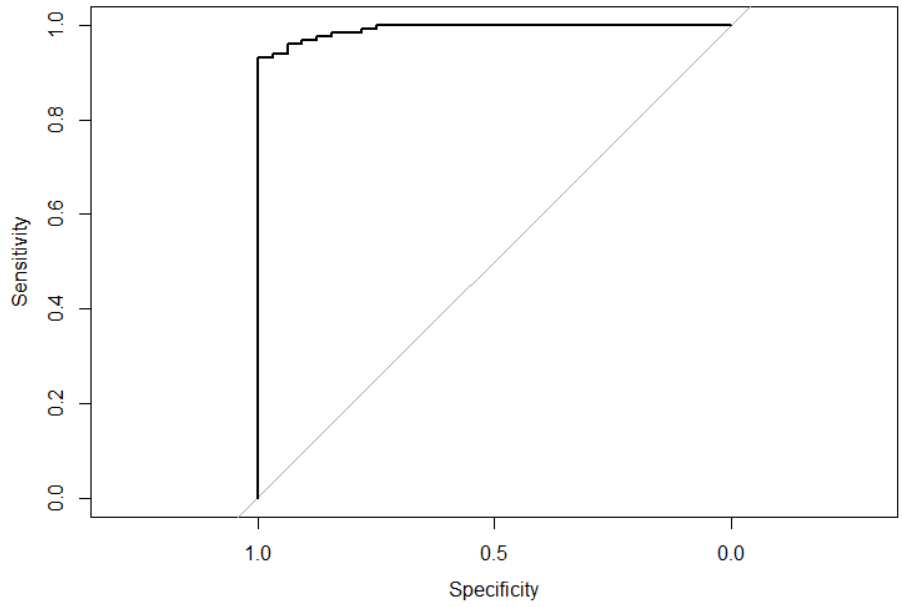
Klasyfikator prawidłowo określa klasę pozytywną, miara *Sensitivity* wyniosła 0,935 (patrz: Tabela 2). Jeszcze lepsza okazała się jego zdolność do określania klasy negatywnej, gdyż miara *Specificity* wyniosła 1. Klasyfikator nie popełnia błędu niepoprawnej klasyfikacji do klasy pozytywnej. Z kolei błąd drugiego rodzaju występuje rzadko. Prawdopodobieństwo poprawnej klasyfikacji jest na poziomie 0,944.

Statystki pochodne:

Statystyka	
SE	0,935
SP	1
FPR	0
FNR	0,065
ACC	0,944

Tabela 2 - statystyki pochodne modelu z metodą podziału "hellinger" z dziesięcioma zmiennymi, wyznaczonego przy użyciu grid search

Krzywa ROC modelu prezentuje się następująco:

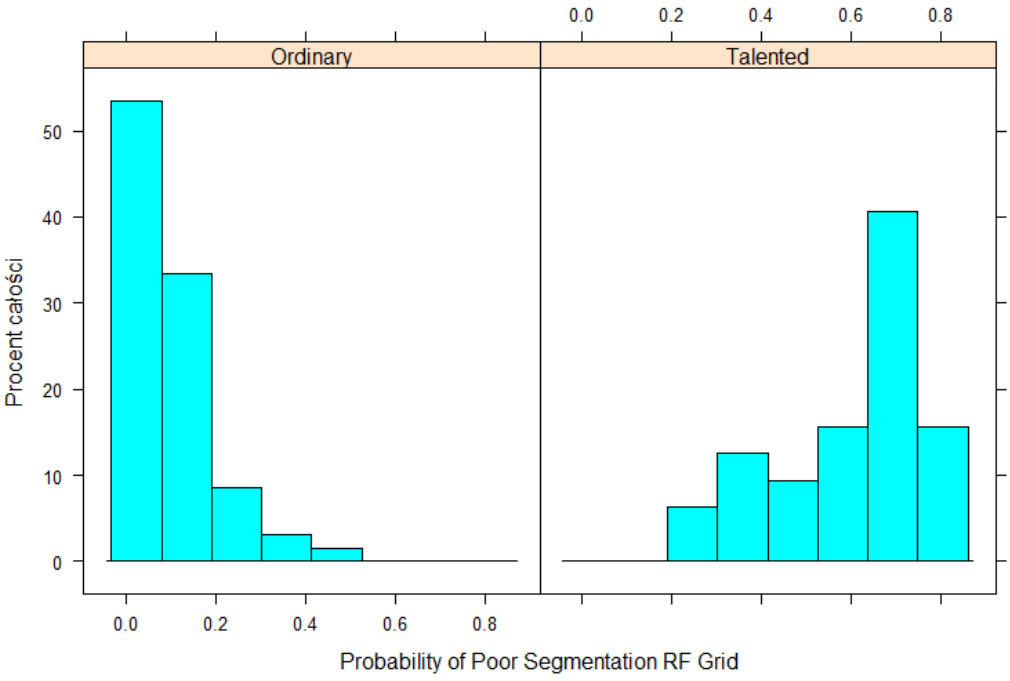


Rysunek 2 - krzywa ROC modelu z metodą podziału "hellinger" i dziesięcioma zmiennymi, wyznaczonego przy pomocy grid search

AUC = 0.9918

Klasyfikator odznacza się znakomitą jakością.

Rozkład prawdopodobieństwa błędnej klasyfikacji (Rysunek 3)



Rysunek 3 - Rozkład Prawdopodobieństwa błędnej klasyfikacji modelu z metodą podziału z metodą "hellinger" i dziesięcioma zmiennymi wyznaczonego przy pomocy grid search

Wykorzystując optymalne hiperparametry uzyskane w pierwszym kroku, przystąpiono do ustalenia optymalnej liczby drzew w modelu.

Uzyskano 5 modeli o różnej liczbie drzew. Uśrednione miary jakości modeli porównano w tabeli.

Tab.

Liczba drzew	500	1000	1500	2000	2500
AUC	0.7642065	0.7730016	0.7706849	0.7662693	0.7679753
Sensitivity	0.9648463	0.9682762	0.963148	0.9674548	0.9682873
Specificity	0.0945614	0.1175439	0.1045614	0.08701754	0.1152632

Tabela 3 - porównanie jakości lasów losowych dla modelu wyznaczonego przy pomocy grid search

Model zawierający 1000 drzew został wybrany jako najbardziej optymalny w oparciu o miarę najlepszej średniej precyzji.

Model końcowy z optymalnymi parametrami

Uzyskawszy zestaw optymalnych hiperparametrów, zbudowano finalny model.

Optymalne parametry:

- zmienne: 10,

- metoda podziału obserwacji: *hellinger*,

- minimalna wielkość liścia: 3,

- liczba drzew: 1000.

Macierz błędów

Predykcja	Ordinary	Talented
Ordinary	127	30
Talented	2	2

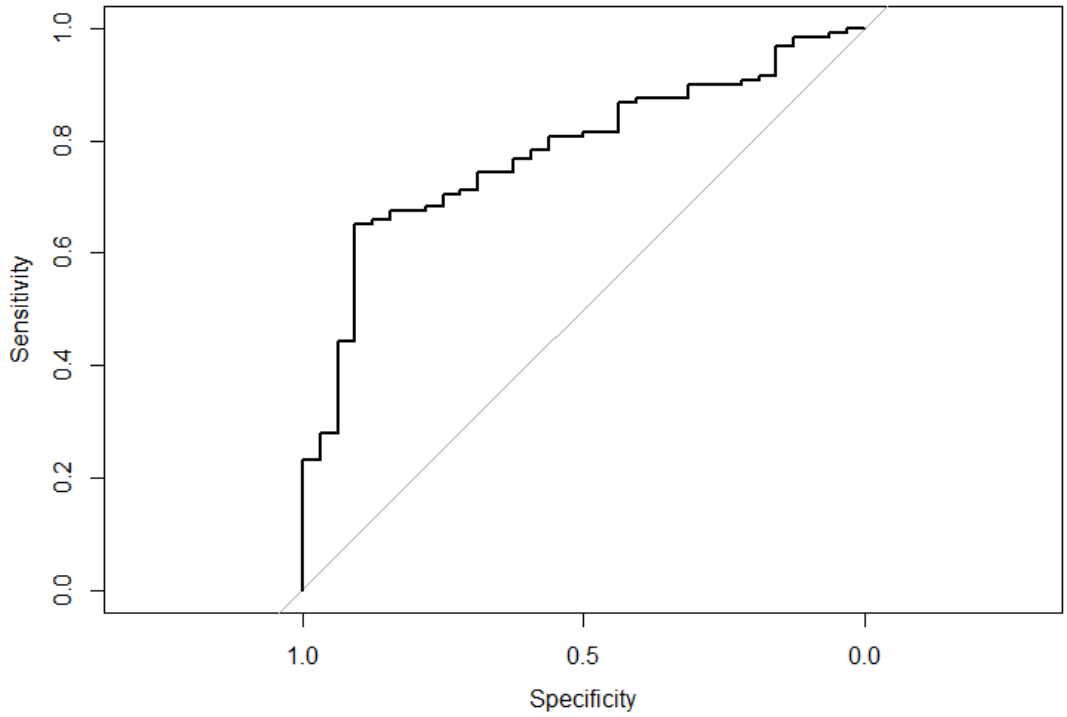
Tabela 4 - macierz błędów ostatecznego modelu z wykorzystaniem grid search

Statystyki pochodne

Statystyka	
SE	0,985
SP	0,5
FPR	0,5
FNR	0,015
ACC	0,801

Tabela 5 - statystyki pochodne ostatecznego modelu z wykorzystaniem grid search

Krzywa ROC ostatecznego modelu prezentuje się następująco:

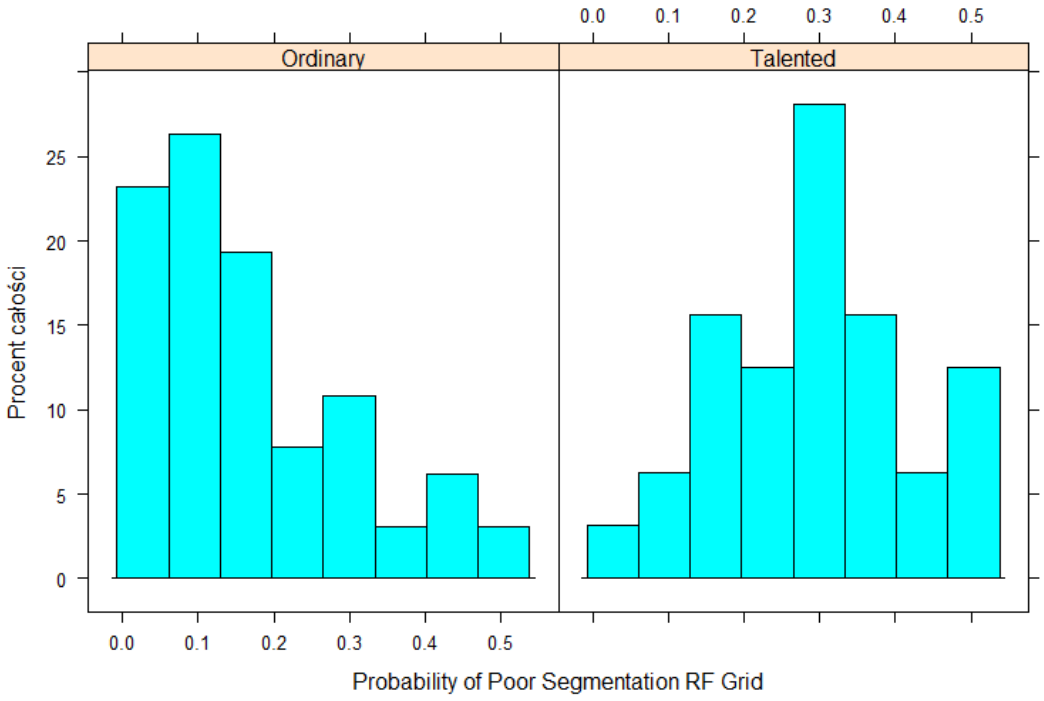


Rysunek 4- krzywa ROC ostatecznego modelu wykorzystującego grid search

AUC = 0.7793

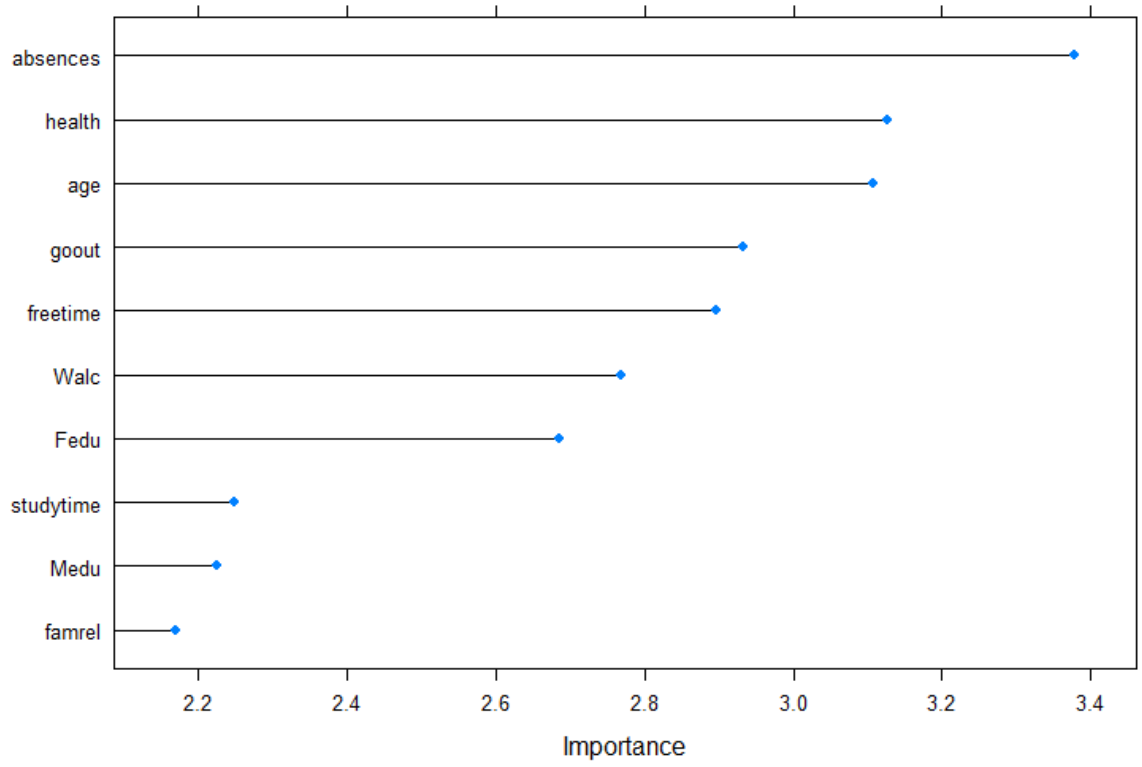
Klasyfikator odznacza się dobrą jakością.

Rozkład prawdopodobieństwa błędnej klasyfikacji



Rysunek 5 - rozkład prawdopodobieństwa błędnej klasyfikacji ostatecznego modelu wykorzystującego grid search

Ważność zmiennych w modelu *grid search* prezentuje się następująco. Z wykresu wynika, iż zmienna obrazująca liczbę nieobecności na zajęciach szkolnych jest najistotniejsza w klasyfikacji.



Rysunek 6 - Wykres ważności zmiennych w końcowym modelu z wykorzystaniem grid search

Model wykorzystujący algorytm *random search*

Optymalne hiperparametry wyrażające maksymalną liczbę zmiennych użytych do budowy drzew, minimalną wielkość liścia oraz metody podziału obserwacji uzyskano za pomocą algorytmu *random search*.

Do uzyskania wyników predykcyjnych ponownie zastosowano metodę 5-krotnej walidacji krzyżowej w trzech seriach.

Wartości optymalnych hiperparametrów wybranych do budowy modelu są następujące:

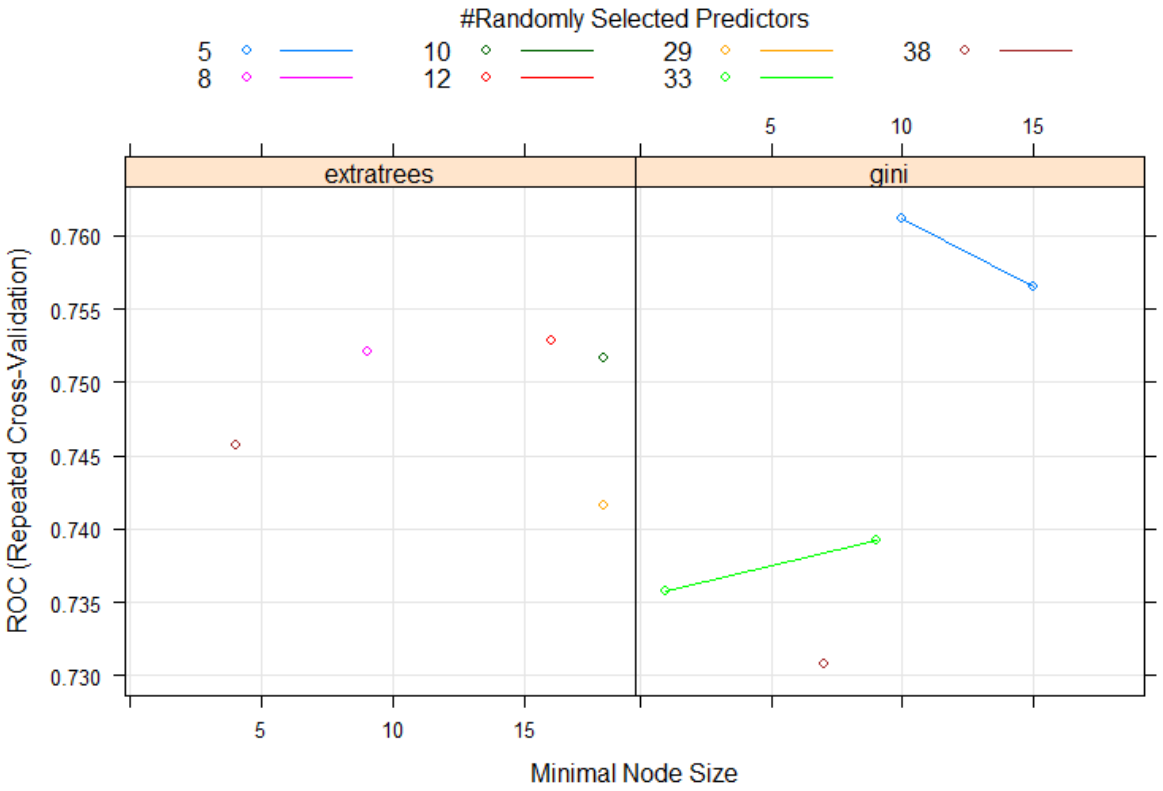
- maksymalna liczba użytych zmiennych = 5,

- metoda podziału = *gini,*

- minimalna wielkość liścia = 10.

Wyboru optymalnego modelu dokonano na podstawie porównań krzywych ROC. Wykres nr przedstawia wszystkie modele brane pod uwagę.

Bez wątpienia metoda podziału *gini* pozwala uzyskać znacznie wyższe wartości pola dla krzywych ROC. Modelem odznaczającym się na wykresie jest ten skonstruowany z 5 zmiennych. Metoda podziału *extratrees* jest mniej efektywna.



Rysunek 7 - porównanie pola powierzchni pod krzywą ROC dla modeli wykorzystujących random search

Macierz pomyłek

Predykcja	Ordinary	Talented
Ordinary	129	32
Talented	0	0

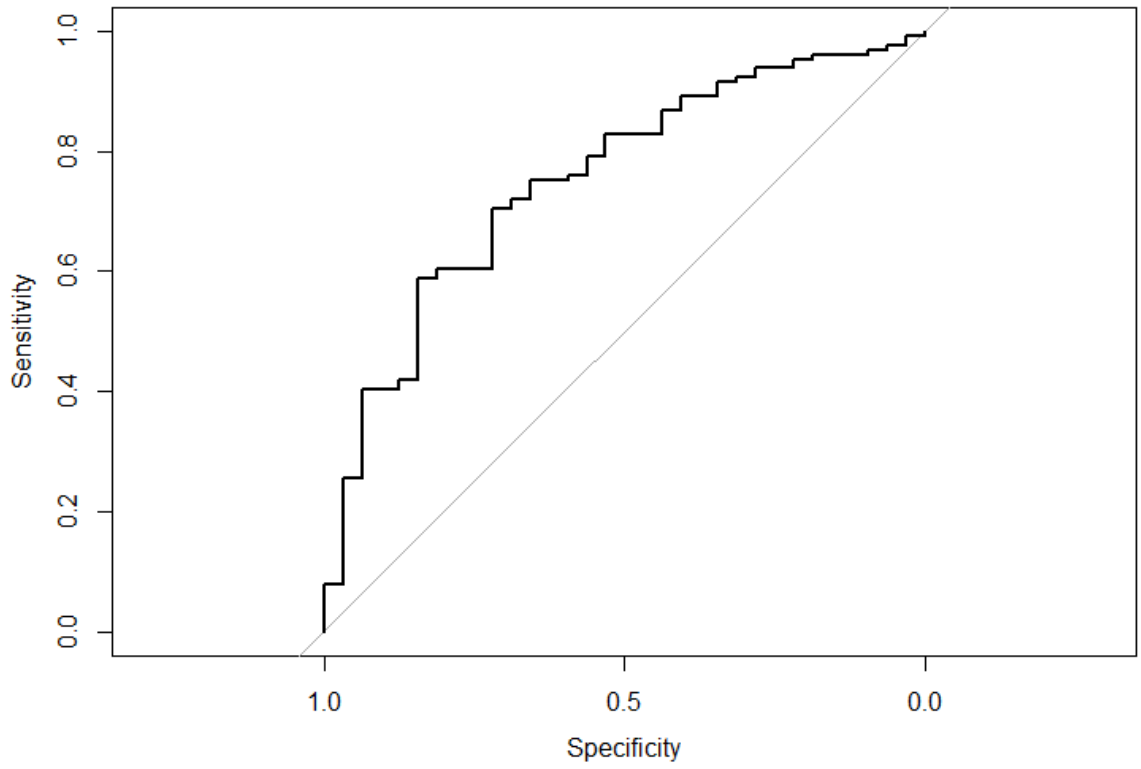
Tabela 6 - macierz błędów modelu wykorzystującego metodę "gini" i random search

Statystyki pochodne

Statystyka	
SE	0,801
SP	0
FPR	1
FNR	0,199
ACC	0,801

Tabela 7 - macierz błędów modelu wykorzystującego metodę "gini" i random search

Krzywa ROC

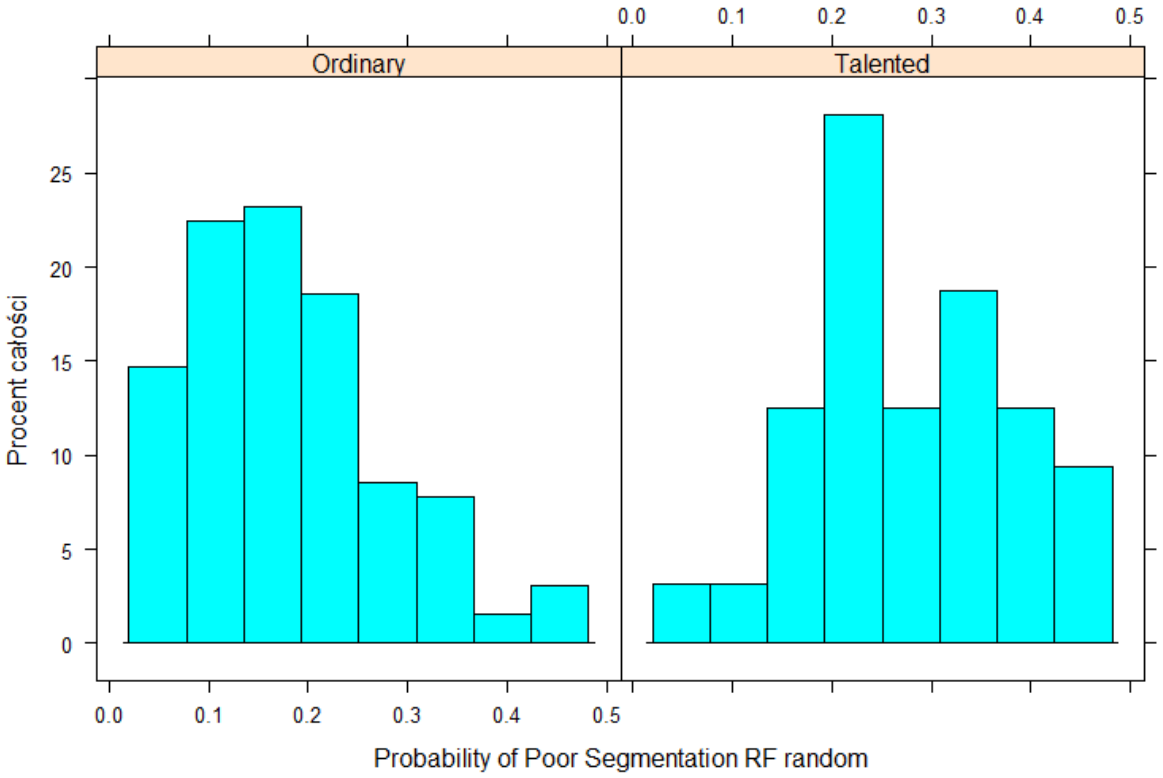


Rysunek 8 - krzywa ROC modelu wykorzystującego metodę "gini" i random search

AUC = 0.7522

Klasyfikator odznacza się dobrą jakością.

Rozkład prawdopodobieństwa błędnej klasyfikacji



Rysunek 9 - rozkład prawdopodobieństwa błędnej klasyfikacji modelu wykorzystującego metodę "gini" i random search

Tak jak w przypadku pierwszego modelu, optymalne hiperparametry uzyskane w pierwszym kroku zostały użyte do ustalenia optymalnej liczby drzew w modelu.

Uśrednione miary jakości pięciu modeli o różnej liczbie drzew porównano w tabeli.

Liczba drzew	500	1000	1500	2000	2500
AUC	0.7577598	0.7429487	0.7435897	0.7500000	0.7833333
Sensitivity	0.9871795	0.9871795	1.0000000	0.9871795	0.9871795
Specificity	0.03350877	0.02017544	0.02368421	0.01684211	0.04035088

Tabela 8 - porównanie lasów losowych dla modelu wykorzystującego random search

Model zawierający 2500 drzew został wybrany jako najbardziej optymalny w oparciu o miarę najlepszej średniej precyzji.

Model końcowy z optymalnymi parametrami

Uzyskawszy zestaw optymalnych hiperparametrów, zbudowano finalny model.

Optymalne parametry:

- zmienne: 5,

- metoda podziału obserwacji: *gini*,

- minimalna wielkość liścia: 10,

- liczba drzew: 2500.

Macierz błędów

Predykcja	Ordinary	Talented
Ordinary	129	32
Talented	0	0

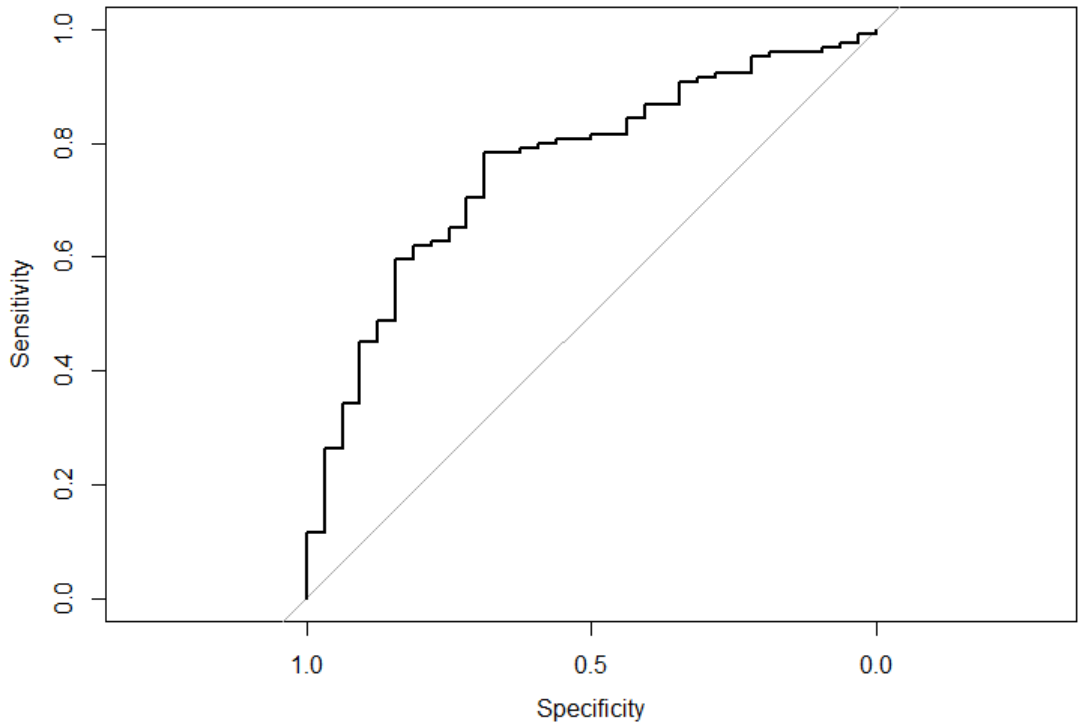
Tabela 9 macierz błędów ostatecznego modelu z wykorzystaniem random search

Statystyki pochodne

Statystyka	
SE	0,801
SP	0
FPR	1
FNR	0,199
ACC	0,801

Tabela 10 - statystyki pochodne ostatecznego modelu z wykorzystaniem random search

Krzywa ROC modelu końcowego

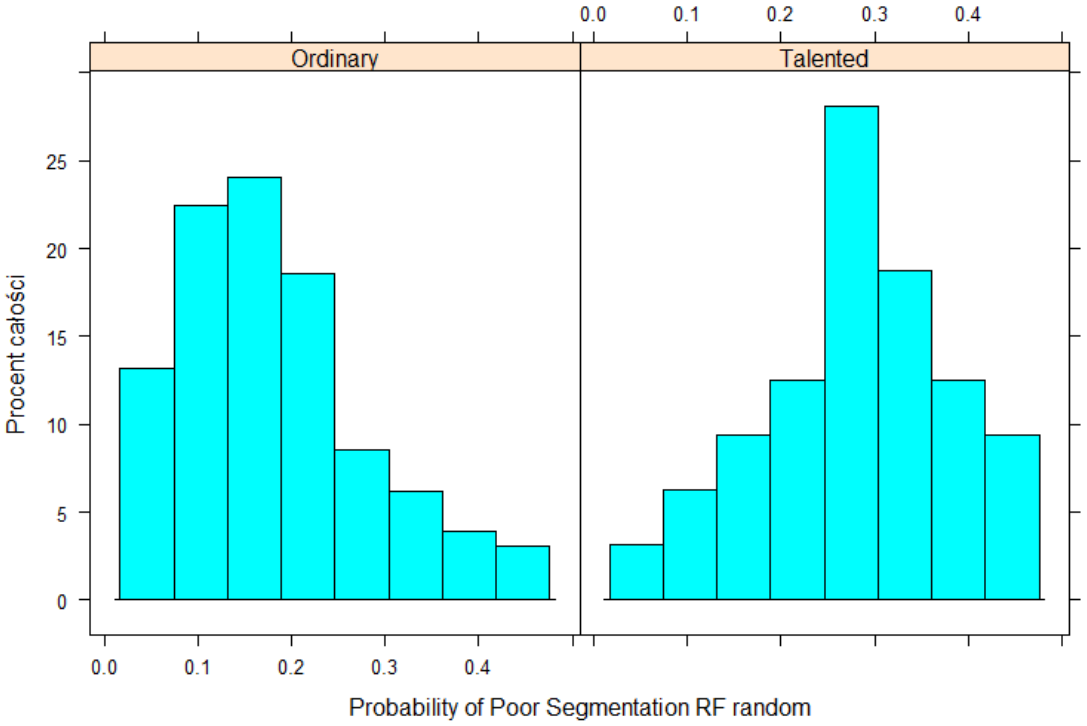


Rysunek 10 krzywa ROC ostatecznego modelu wykorzystującego random search

AUC = 0.7587

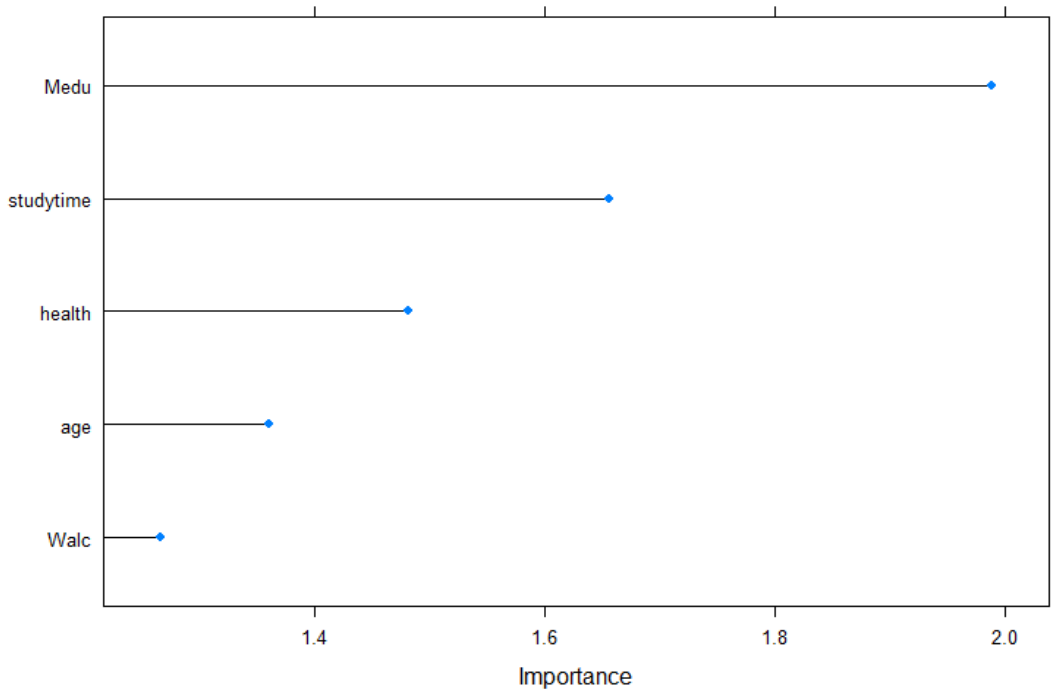
Jakość klasyfikatora jest dobra.

Rozkład prawdopodobieństwa błędnej klasyfikacji



Rysunek 11 Rysunek 5 - rozkład prawdopodobieństwa błędnej klasyfikacji ostatecznego modelu wykorzystującego random search

Ważność zmiennych w modelu *random search* prezentuje się następująco. Z wykresu wynika, iż zmienna obrazująca wykształcenie matki jest najistotniejsza w klasyfikacji.



Rysunek 12 - Wykres ważności zmiennych w modelu końcowym wykorzystującym random search

4. Walidacja

W niniejszym raporcie opracowane zostały dwa modele oparte na algorytmie lasów losowych. W pierwszej kolejności oszacowano hiperparametry modelu przy użyciu algorytmu *grid search,* a następnie – modelu z algorytmem *random search.*

Na koniec jakość obydwu modeli zostanie zweryfikowana przy użyciu zupełnie nowego zestawu danych dotyczących nauki matematyki. Podobnie jak w przypadku nauki języka portugalskiego, jako klasę pozytywną zdefiniowano wyniki znajdujące się w pierwszym kwartylu najwyższych wyników zmiennej prognozowanej G3, a jako klasę negatywną - wszystkie wyniki poniżej tego kwartyla.

Wyniki macierzy błędów dla obydwu modeli są następujące:

Model *grid search*		
Predykcja	Ordinary	Talented
Ordinary	293	45
Talented	29	28

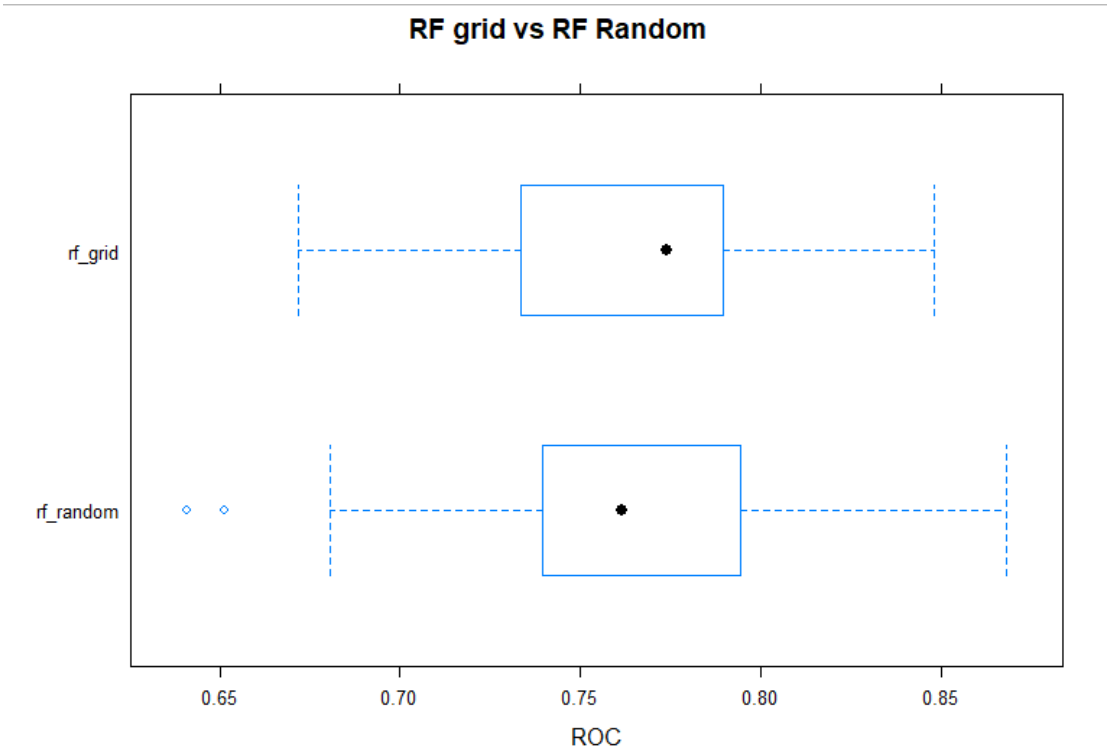
Model *random search*		
Predykcja	Ordinary	Talented
Ordinary	306	57
Talented	16	16

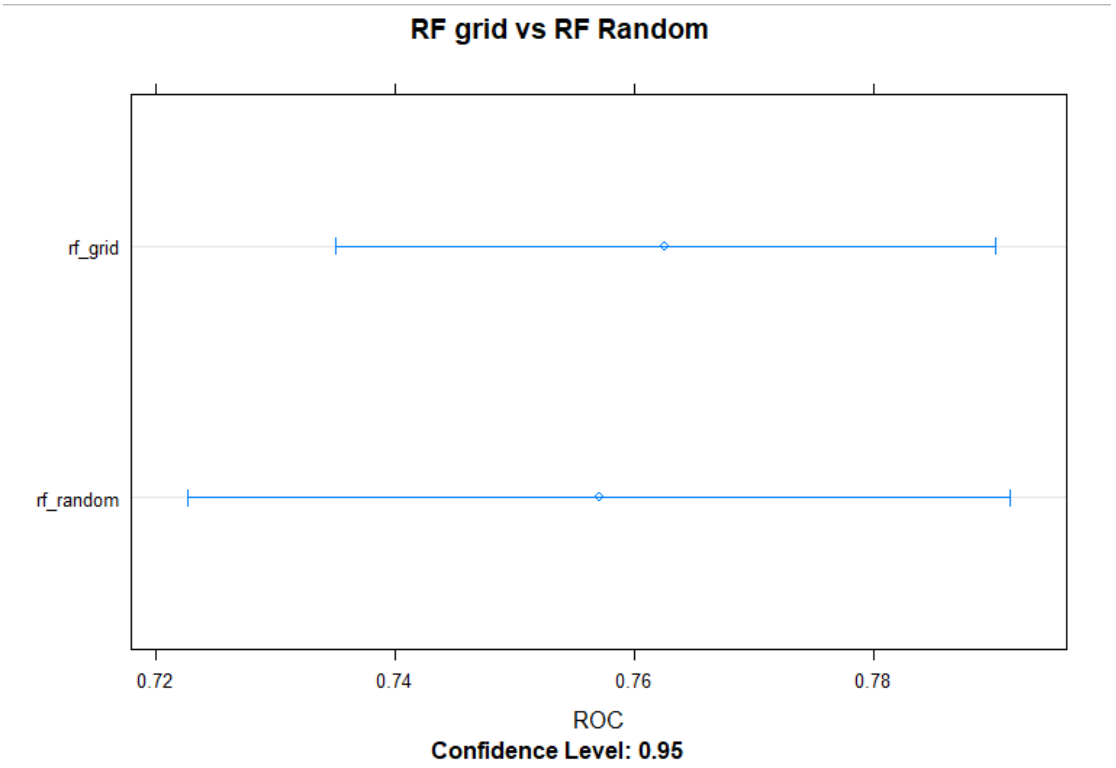
Statystyki pochodne

Model *grid search*	
SE	0,8669
SP	0,4912
FPR	0,5088
FNR	0,1331
ACC	0,8127
Model *grid search*	
AUC	0.7625734

Model *random search*	
SE	0,843
SP	0,5
FPR	0,5
FNR	0,157
ACC	0,8152
Model *random search*	
AUC	0.7570249

Uśredniona wartość AUC dla modelu *grid search* jest nieznacznie wyższa niż dla modelu *random search,* co świadczy o jego lepszej jakości.





Ze względu na zastosowaną metodę nie podejmujemy się interpretacji wyników – możemy powiedzieć jedynie o tym, co decyduje o klasyfikacji w ramach modelu, jednakże interpretowanie wyników jako pewien ciąg przyczynowo-skutkowy może być w tym przypadku wyjątkowo myląca.