Analiza danych StudentsPerformace i zbudowanie modelu w oparciu o Naiwny Klasyfikator Bayesa oraz K-NN.

Autor: Ewelina Brach

Kierunek: geoinformatyka

Numer indeksu: 400444

Kraków, 2022

1. Wstęp

Celem projektu jest analiza danych tabelarycznych dotyczących wyników testów (StudentsPerformance), jakie zdobyli uczniowie. Dane posiadają osiem wartości, czyli złożone są z ośmiu kolumn. Pięć pierwszych to wartości typu char i określają grupę osób poprzez np. płeć, grupę podczas testu, czy też przygotowanie. Kolejne to wyniki 3-etapowego testu z wartościami liczbowymi, który składał się z matematyki, czytania i pisania. Zbiór danych składa się z 1000 obserwacji.

Po wczytaniu danych, zapoznaniu się z nimi, przystąpiono do ich czyszczenia oraz do analizy.

Dane prezentują się następująco:

	gender	race	parent_lvl_edu	lunch	test_prep_course	math_score	reading_score	writing_score
1	female	group B	bachelor's degree	standard	none	72	72	74
2	female	group C	some college	standard	completed	69	90	88
3	female	group B	master's degree	standard	none	90	95	93
4	male	group A	associate's degree	free/reduced	none	47	57	44
5	male	group C	some college	standard	none	76	78	75
6	female	group B	associate's degree	standard	none	71	83	78
7	female	group B	some college	standard	completed	88	95	92
8	male	group B	some college	free/reduced	none	40	43	39

Rys. 1. Taela zawierająca analizowane dane StudentsPerformance.

2. Czyszczenie danych i analiza

Ten etap rozpoczęto od sprawdzenia, czy w zbiorze danych występują wartości NA, korzystając z poniższej funkcji:

```
> sapply(data, function(x)(sum(is.na(x)))) # NA counts
          gender             race   parent_lvl_edu            lunch
               0                0                0                0
test_prep_course       math_score    reading_score    writing_score
               0                0                0                0
>
```

Rys. 2. Funkcja na podstawie której sprawdzono obecność wartości NA.

Pozwoliło to na stwierdzenie, że w danych nie ma wartości NA.

Następnie wyświetlono statystki dla analizowanego zbioru.

```
> summary(data)
    gender             race           parent_lvl_edu       lunch
 Length:1000        Length:1000        Length:1000        Length:1000
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


 test_prep_course     math_score      reading_score    writing_score
 Length:1000        Min.   :  0.00   Min.   : 17.00   Min.   : 10.00
 Class :character   1st Qu.: 57.00   1st Qu.: 59.00   1st Qu.: 57.75
 Mode  :character   Median : 66.00   Median : 70.00   Median : 69.00
                    Mean   : 66.09   Mean   : 69.17   Mean   : 68.05
                    3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00
                    Max.   :100.00   Max.   :100.00   Max.   :100.00
```

Rys. 3. Wyświetlenie statystyk opisowych dla zbioru danych StudentsPerformance.

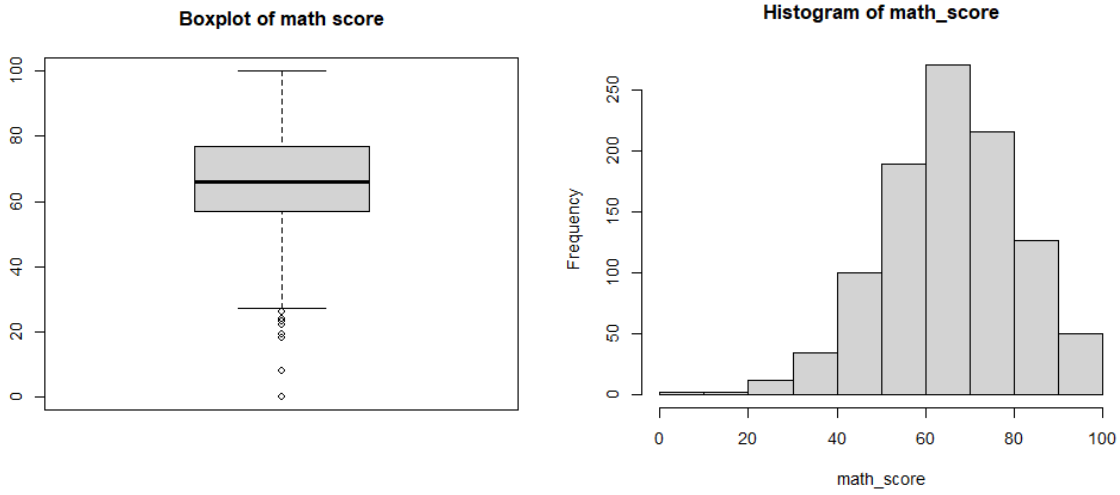
Średnia wyników z testu matematycznego jest niższa, niż z testu czytelniczego i pisemnego.

Na podstawie otrzymanych wartości min, max nie tylko dowiadujemy się o najniższym, jak i najwyższym wyniku z poszczególnego testu, ale także możemy określić poprawność danych. Widzimy, że w wynikach testowych nie znajdują się wartości ujemne, co jest dobrą wiadomością, gdyż obecność taki wartości w wynikach może sugerować błędy grube w danych. W naszym przypadku, natomiast nie zaobserwowaliśmy takich wartości, dlatego możemy stwierdzić, że w danych nie ma błędów grubych.

Kolejno przystąpiono do analizy kolumn zawierających dane o wynikach z poszczególnych testów.

a) math score

```
> shapiro.test(math_score)

        Shapiro-Wilk normality test

data:  math_score
W = 0.99315, p-value = 0.0001455
```

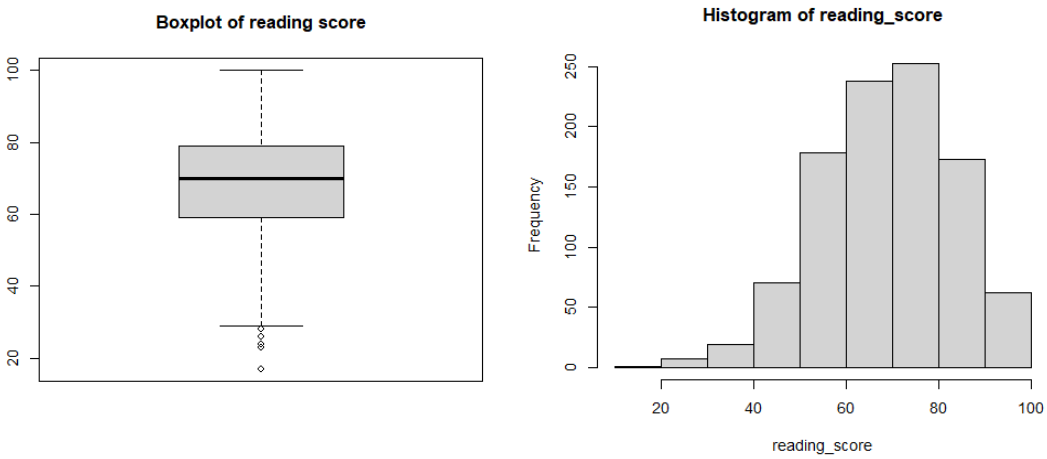


Rys. 4,5,6. Wynik testu Shapiro-Wilk, wykres typu boxplot oraz histogram dla danych StudentsPerformance$math_score.

Analiza pokazała, że dane nie są rozkładem normalnym, co można zauważyć na podstawie histogramu oraz wyniku testu Shapiro-Wilka (p-value <0.05). Dodatkowo występują wartości odstające.

b) reading score

```
> shapiro.test(reading_score)

        Shapiro-Wilk normality test

data:  reading_score
W = 0.99292, p-value = 0.0001055
```

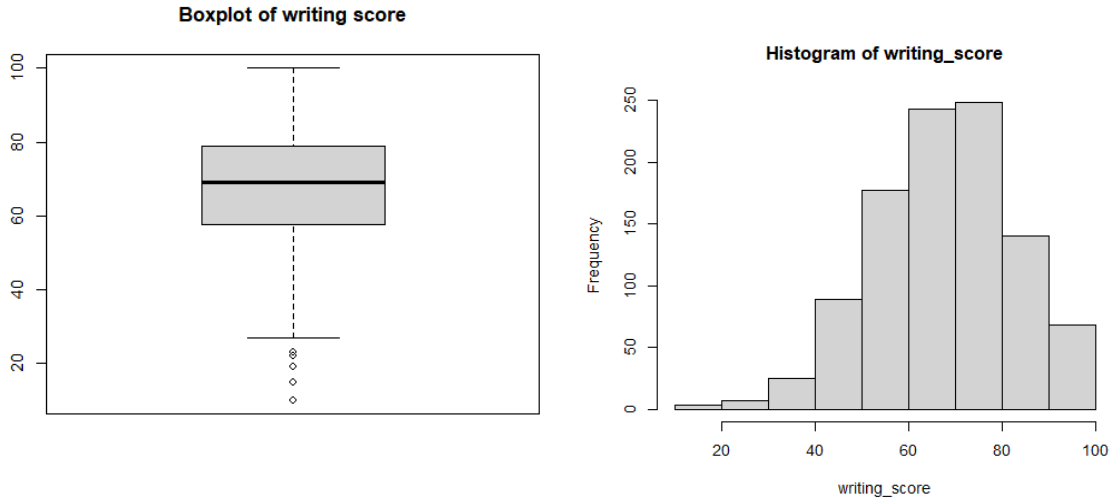


Rys. 7,8,9. Wynik testu Shapiro-Wilk, wykres typu boxplot oraz histogram dla danych StudentsPerformance$reading_score.

Dla danych reading score również można stwierdzić, że nie są one rozkładem normalnych, co jak w przypadku math score pokazał histogram oraz wynik testu Shapiro-Wilk. Także otrzymano kilka wartości odstających.

c) writing score

```
> shapiro.test(writing_score)

        Shapiro-Wilk normality test

data:  writing_score
W = 0.99196, p-value = 2.922e-05
```

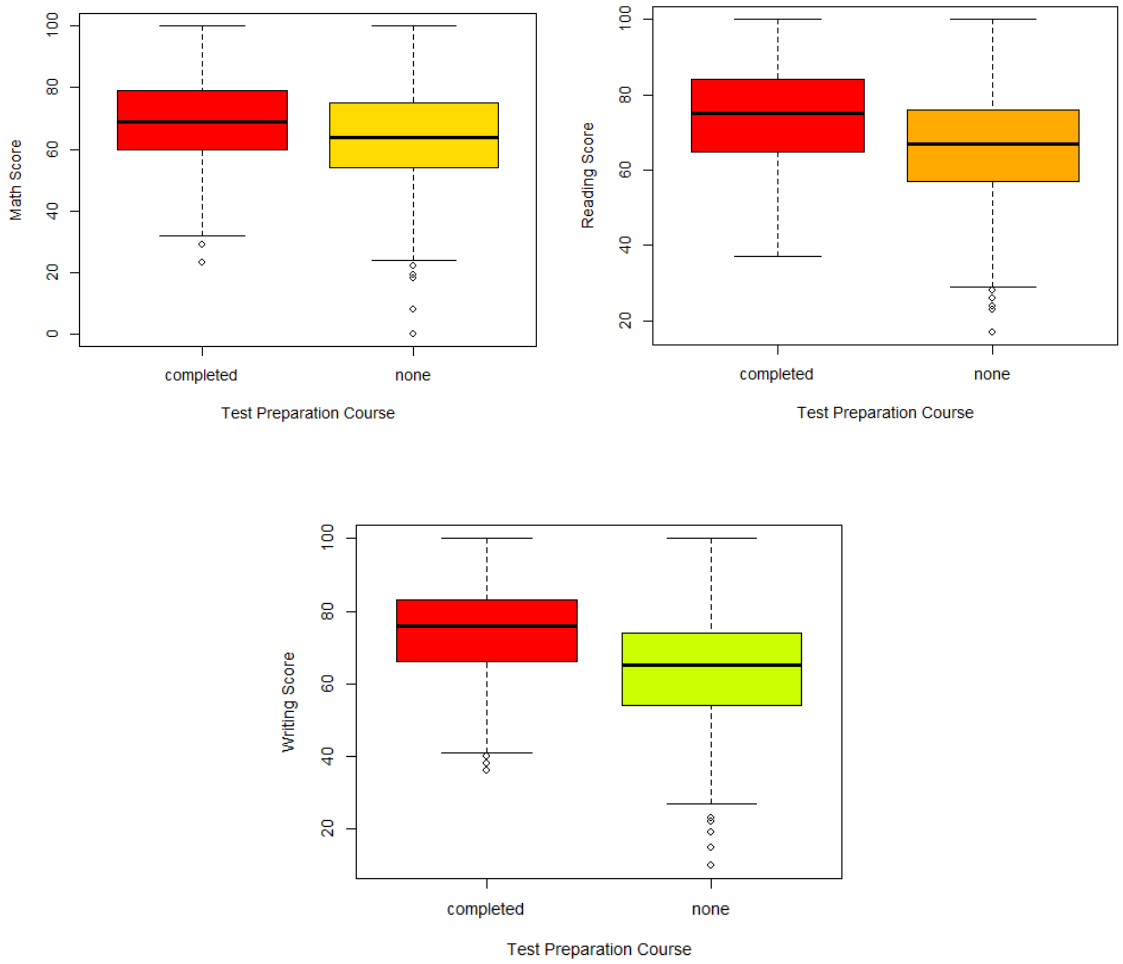


Rys. 10,11,12. Wynik testu Shapiro-Wilk, wykres typu boxplot oraz histogram dla danych StudentsPerformance$writing_score.

Dla danych writing score otrzymano podobną sekwencję, jak dla pozostałych analizowanych. Dane nie mają rozkładu normalnego, co stwierdzono na podstawie testu Shapiro-Wilk oraz histogramu. Dodatkowo także zauważono wartości odstające dla wyniku testu.

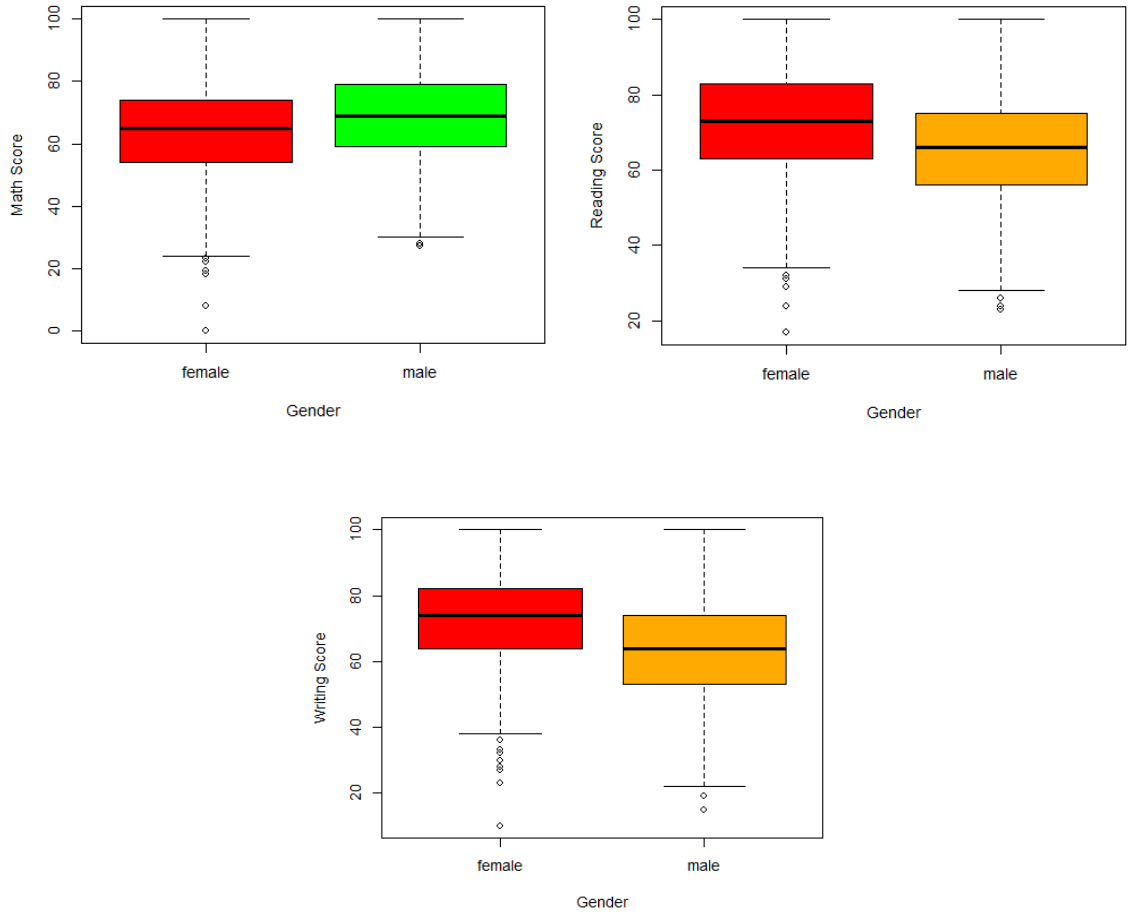
Podsumowując tę część analizy, dane nie mają rozkładu normalnego, w danych występują wartości odstające. Biorąc pod uwagę charakter danych – są to wyniki trzech testów przeprowadzonych w grupie uczniów – możliwa jest obecność różnorodnych wyników (ktoś mógł mieć 0, a ktoś 100 punktów). Dlatego zdecydowano, aby nie usuwać tych wartości.

Kilka dodatkowych analiz:



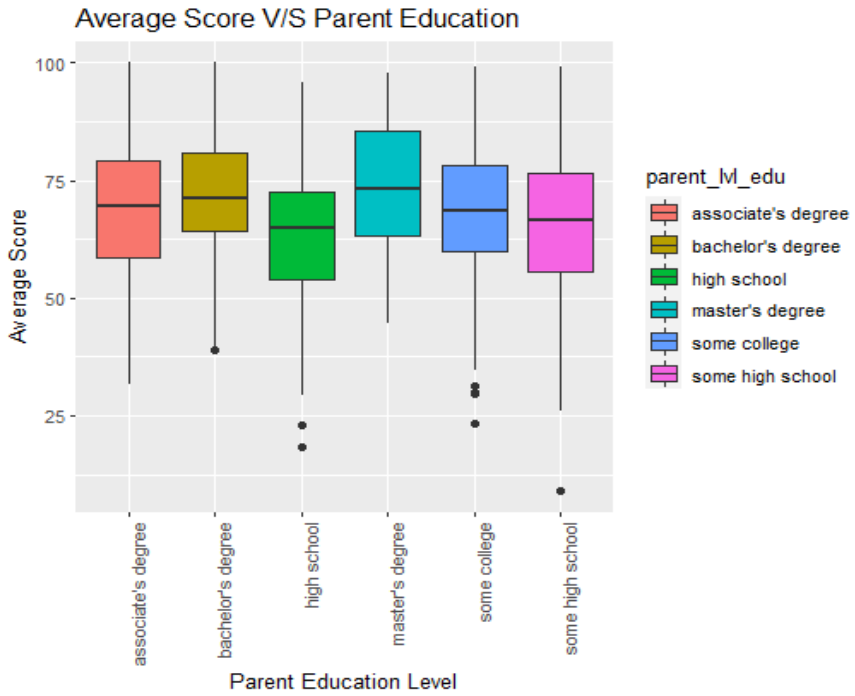
Rys. 13,14,15. Wykresy boxplot pokazujące zależność wcześniejszego przygotowania się do kursu w zależności od wyniku z matematyki, czytania i pisania.

Na podstawie otrzymanych wykresów pudełkowych można zaobserwować, że z wyjątkiem niewielkiej liczby uczniów (odstających), którzy ukończyli kurs przygotowujący do testu, mają wyższą medianą wyników z matematyki, czytania i pisania, niż uczniowie, którzy takiego kursu nie ukończyli.



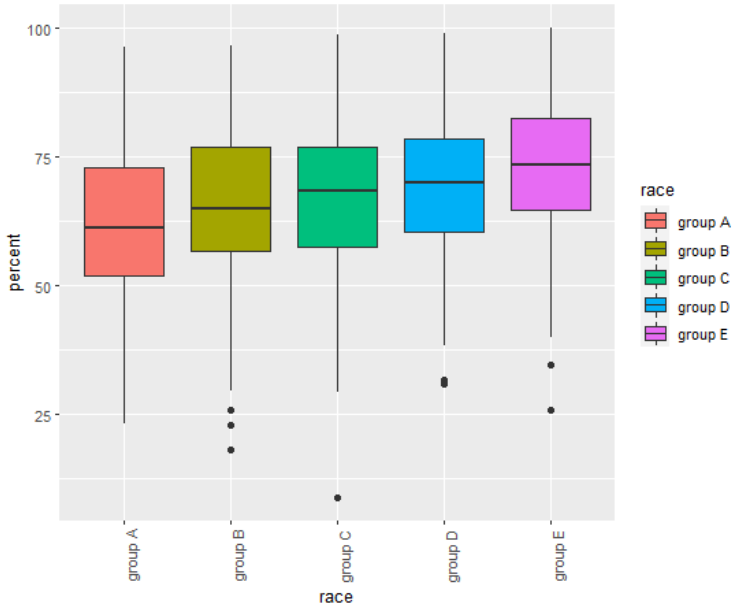
Rys. 16,17,18. Wykresy boxplot pokazujące zależność płci w zależności od wyniku z matematyki, czytania i pisania.

Biorąc pod uwagę otrzymane wykresy, możemy powiedzieć, że w teście z matematyki lepiej wypadli mężczyźni, natomiast w teście z czytania i pisania – kobiety.



Rys. 19. Wykres boxplot pokazujący zależność wykształcenia piszących w stosunku do średnich wyników testów.

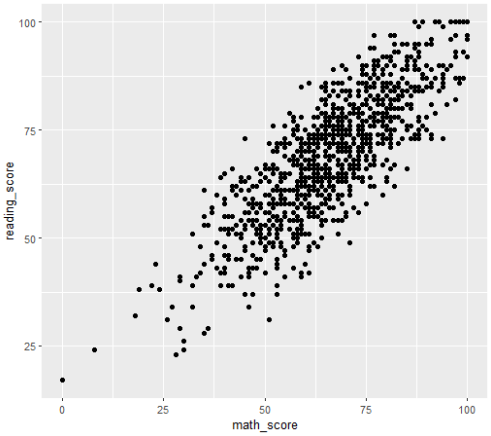
Na podstawie średniej otrzymanej ze wszystkich przeprowadzonych testów, najlepiej wypadły osoby mające tytuł magistra.



Rys. 20. Wykres boxplot pokazujący zależność między grupą w stosunku do średnich wyników testów.

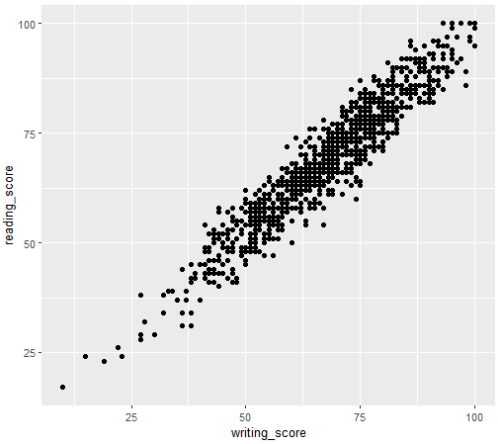
Biorąc pod uwagę powyższy wykres, możemy zauważyć, że grupa E ma najlepsze wyniki w porównaniu do reszty grup. Uczniowie z grupy A mają najniższe wyniki.

Wyliczając zależności pomiędzy wynikami poszczególnych testów, otrzymano następujące wyniki korelacji.



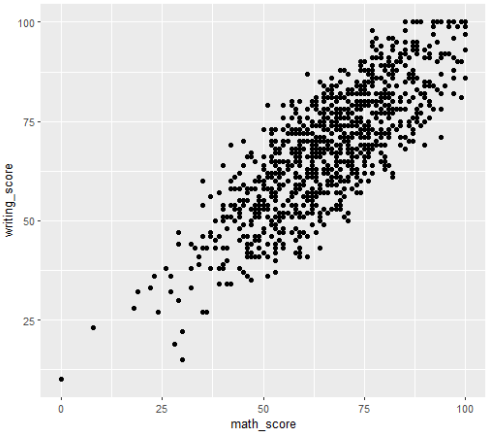

```
Call:
lm(formula = math_score ~ reading_score, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-24.3419  -6.3419  -0.0221   6.2713  24.6581

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.35759    1.33818   5.498 4.87e-08 ***
reading_score 0.84910    0.01893  44.855  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.736 on 998 degrees of freedom
Multiple R-squared:  0.6684,    Adjusted R-squared:  0.6681
F-statistic:  2012 on 1 and 998 DF,  p-value: < 2.2e-16
```

Rys. 21. Wykres korelacji pokazujący zależność pomiędzy wynikami testów z matematyki i czytania.




```
Call:
lm(formula = writing_score ~ reading_score, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9573  -2.9573   0.0363   3.1026  15.0557

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.667554   0.693792  -0.962    0.336
reading_score  0.993531   0.009814 101.233   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.529 on 998 degrees of freedom
Multiple R-squared:  0.9113,    Adjusted R-squared:  0.9112
F-statistic: 1.025e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

Rys. 22. Wykres korelacji pokazujący zależność pomiędzy wynikami testów z czytania i pisania.




```
Call:
lm(formula = math_score ~ writing_score, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-24.8467  -6.4600   0.1464   6.4356  25.5515

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.58310    1.31369   8.817   <2e-16 ***
writing_score  0.80092    0.01884  42.511   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.049 on 998 degrees of freedom
Multiple R-squared:  0.6442,    Adjusted R-squared:  0.6439
F-statistic:  1807 on 1 and 998 DF,  p-value: < 2.2e-16
```

Rys. 23. Wykres korelacji pokazujący zależność pomiędzy wynikami testów z matematyki i pisania.

Wszystkie korelacje są pozytywne.

3. Wybór modelu i uzasadnienie

Po wstępnej analizie otrzymanych danych StudentsPerformance można stwierdzić, że model Naiwnego Klasyfikatora Bayesa idealnie pasuje. Widać w nich zależność dwóch zdarzeń warunkujących się nawzajem. Przykładowe mogłoby brzmieć: Jaka jest zależność pomiędzy wynikami testów, a płcią, czy wykształceniem.

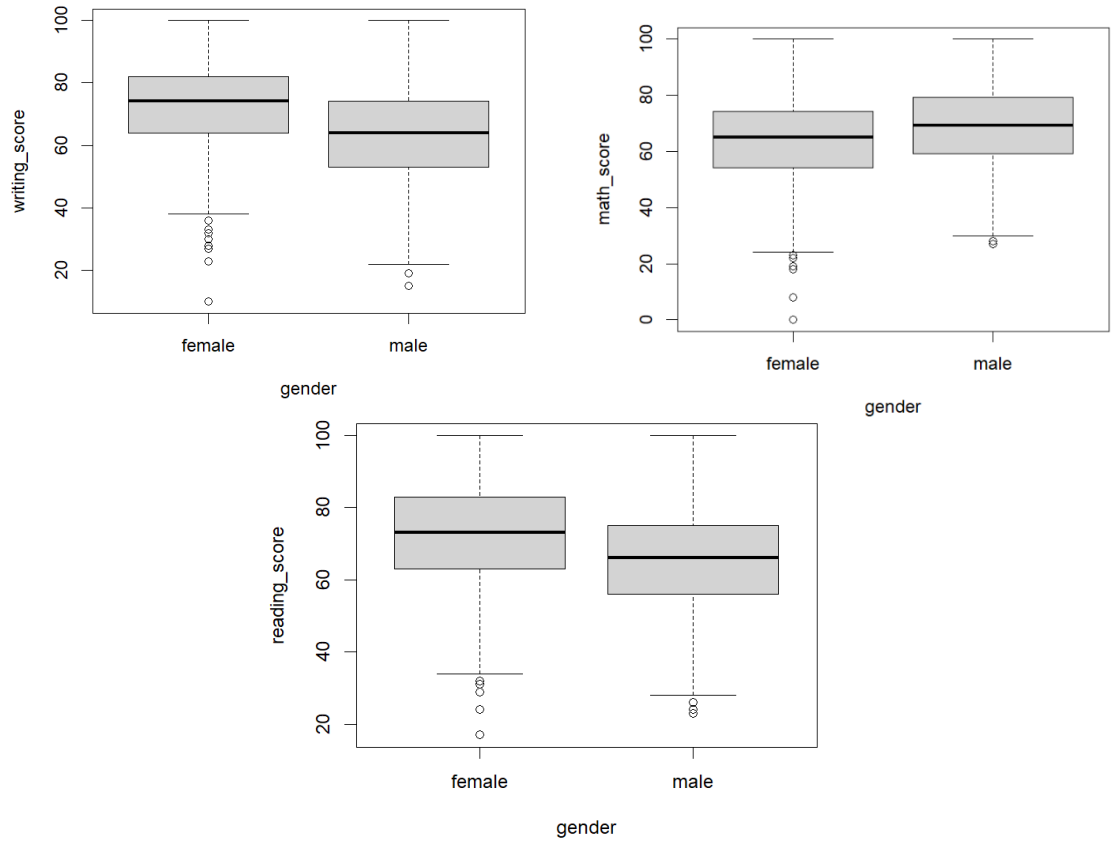
Jako drugi model postanowiono skorzystać z K-NN (k-najbliższych sąsiadów), ponieważ należy on do tej samej grupy algorytmów – klasyfikacyjnych. Zbadano jak ten prosty klasyfikator odpowiada na problem przypisania do klasy na podstawie sąsiedztwa.

4. Model Naiwnego Klasyfikatora Bayesa

Budowanie modelu przeprowadzono w następujący sposób. Zastosowano własną funkcję do wykonania standaryzacji, utworzono podzbiory, gdzie na zbioru testowego wybrano losowo 25% obserwacji, a resztę danych przypisano do zbioru treningowego. Kolejno wybrano klasy poszczególnych obserwacji z podzbiorów. Efektem końcowym był model z wyliczonymi prawdopodobieństwami apriori. Na jego postawie obliczono predykcję. Sprawdzono jakość modelu na podstawie czułości, specyficzności oraz procentu poprawnych klasyfikacji.

Z racji, że dane zawierają kilka kolumn typu char, w których widać zależność grupową, postanowiono że będą one wymodelowane w zależności od tych zmiennych. Sprawdzono jak rozkładają się wyniki poszczególnych testów, biorąc pod uwagę płeć, grupę testu, lunch, wykształcenie oraz przygotowanie.

a) dla zmiennej gender



Rys. 24. Wykresy boxplot przedstawiający zależność wyników testu matematycznego, czytania i pisania od płci.

Model:

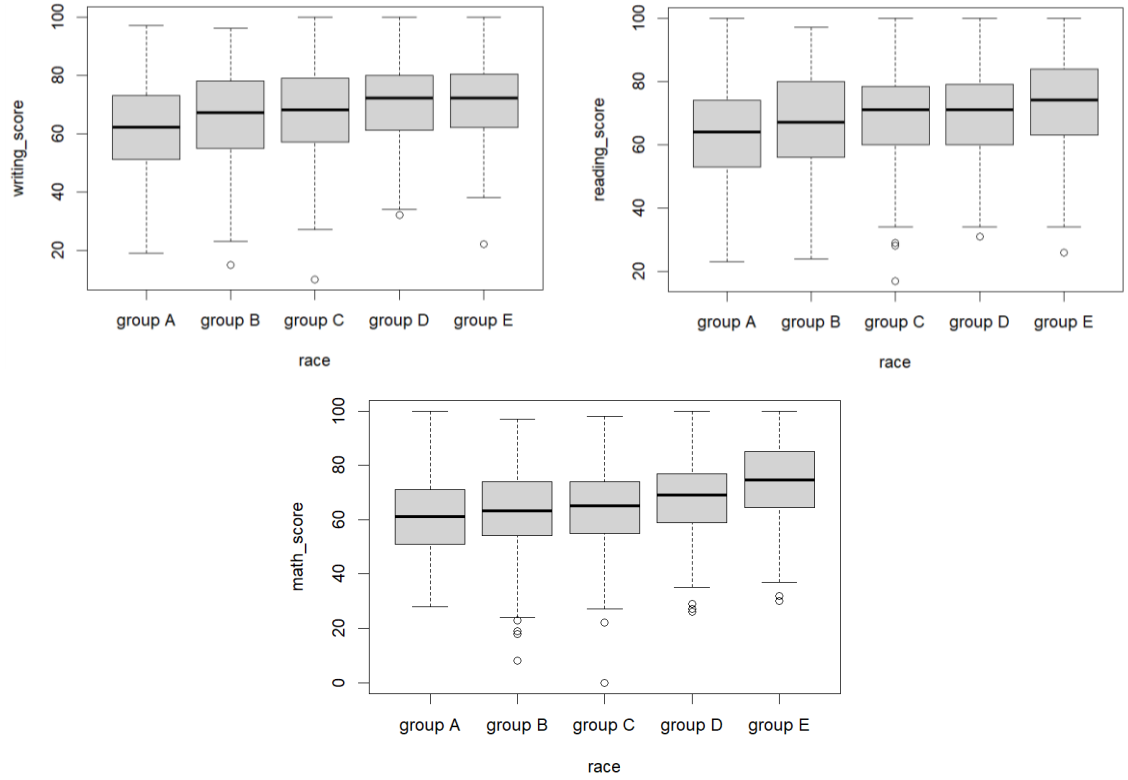
```
> nb_stud1<-naive_bayes(data.gender ~.,train_stud)
> summary(nb_stud1)

================================== Naive Bayes ==================================

- Call: naive_bayes.formula(formula = data.gender ~ ., data = train_stud)
- Laplace: 0
- Classes: 2
- Samples: 750
- Features: 3
- Conditional distributions:
    - Gaussian: 3
- Prior probabilities:
    - female: 0.4947
    - male: 0.5053

--------------------------------------------------------------------------------
```

```
> t_nb_stud1<-table(test_stud[,4],nb_stud1_pred)
> confusionMatrix(t_nb_stud1)
Confusion Matrix and Statistics

        nb_stud1_pred
         female male
  female     93   54
  male       32   71

               Accuracy : 0.656
                 95% CI : (0.5935, 0.7147)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 4.598e-07

                  Kappa : 0.312

 Mcnemar's Test P-Value : 0.02354

            Sensitivity : 0.7440
            Specificity : 0.5680
         Pos Pred Value : 0.6327
         Neg Pred Value : 0.6893
             Prevalence : 0.5000
         Detection Rate : 0.3720
   Detection Prevalence : 0.5880
      Balanced Accuracy : 0.6560

       'Positive' Class : female
```

Rys. 25. Statystyki dla otrzymanego modelu, wyliczonego na podstawie zmiennej gender oraz macierz konfuzji.

Procent poprawnych sklasyfikowanych wartości, czyli trafność wyniosła 66%, specyficzność 0.57, a czułość 0.74. Są to dość wysokie wartości w stosunku do wielkości zbioru. Na ich podstawie możemy stwierdzić, że model w miarę dobrze poradził sobie z wykrywaniem przypadków pozytywnych jak i negatywnych. Płeć uczestników na znaczenie w uzyskanych wynikach testowych. Pozytywną klasą jest female.

b) race/ethnicity



Rys. 26. Wykresy boxplot przedstawiające zależność wyników testu matematycznego, czytania i pisania od race/ethnicity.

Model:

```
> nb_stud3<-naive_bayes(data.race ~.,train_stud)
> summary(nb_stud3)

============================================================ Naive Bayes =============

- Call: naive_bayes.formula(formula = data.race ~ ., data = train_stud)
- Laplace: 0
- Classes: 5
- Samples: 750
- Features: 3
- Conditional distributions:
    - Gaussian: 3
- Prior probabilities:
    - group A: 0.0907
    - group B: 0.1907
    - group C: 0.3133
    - group D: 0.2613
    - group E: 0.144

-------------------------------------------------------------------------------
```

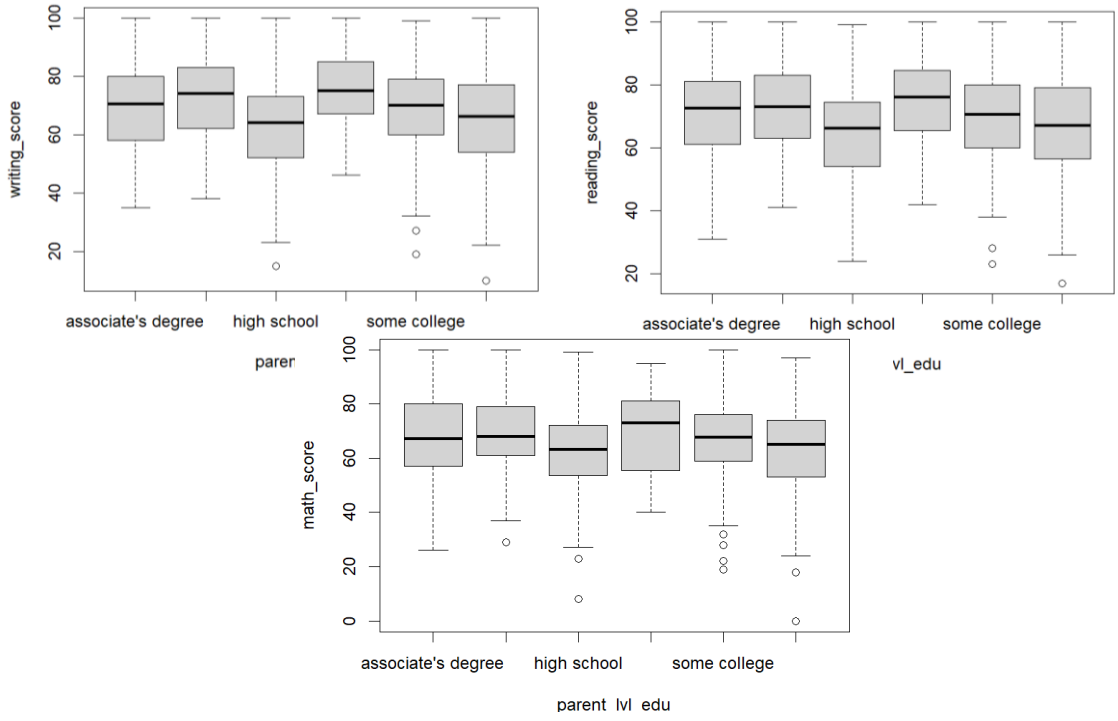
```
> confusionMatrix(t_nb_stud3)
Confusion Matrix and Statistics

          nb_stud3_pred
           group A group B group C group D group E
  group A        0       5      13       2       1
  group B        0       6      21      13       7
  group C        0       5      38      31      10
  group D        0       3      30      26       7
  group E        0       1      14      11       6

Overall Statistics

                Accuracy : 0.304
                  95% CI : (0.2476, 0.3651)
     No Information Rate : 0.464
     P-Value [Acc > NIR] : 1

                   Kappa : 0.0407

 Mcnemar's Test P-Value : 4.64e-06

Statistics by Class:

                     Class: group A Class: group B Class: group C Class: group D Class: group E
Sensitivity                     NA         0.3000         0.3276         0.3133         0.1935
Specificity                  0.916         0.8217         0.6567         0.7605         0.8813
Pos Pred Value                  NA         0.1277         0.4524         0.3939         0.1875
Neg Pred Value                  NA         0.9310         0.5301         0.6902         0.8853
Prevalence                   0.000         0.0800         0.4640         0.3320         0.1240
Detection Rate               0.000         0.0240         0.1520         0.1040         0.0240
Detection Prevalence         0.084         0.1880         0.3360         0.2640         0.1280
Balanced Accuracy               NA         0.5609         0.4922         0.5369         0.5374
>
```

Rys. 27. Statystyki dla otrzymanego modelu, wyliczonego na podstawie zmiennej race/ethicity oraz macierz konfuzji.

W przypadku modelowania na podstawie zmiennej race/ethnicity, trafność zmalała w stosunku do zmiennej gender i wyniosła 30%. Czułość i specyficzność z kolei została przedstawiona dla konkretnych klas. Czułość najwyższa jest dla grupy C i wynosi 0.33, a najniższa dla grupy E – w tym, dla grupy A czułość nie została obliczona i zwrócono wynik NA. Specyficzność najwyższa jest grupy A, a najniższa dla grupy C. We wszystkich grupach najlepiej wychwycone zostały przypadki negatywne.

Biorąc pod uwagę obie te wartości, można zauważyć że suma czułości i specyficzności oscyluje w granicach 1. Oznacza to, że test nie ma wartości i nie ma związku z tym, do jakiej grupy należały osoby wykonujące testy sprawdzające.

c) parent level education



Rys. 28. Wykresy boxplot przedstawiające zależność wyników testu matematycznego, czytania i pisania od parent_level_education.

Model:

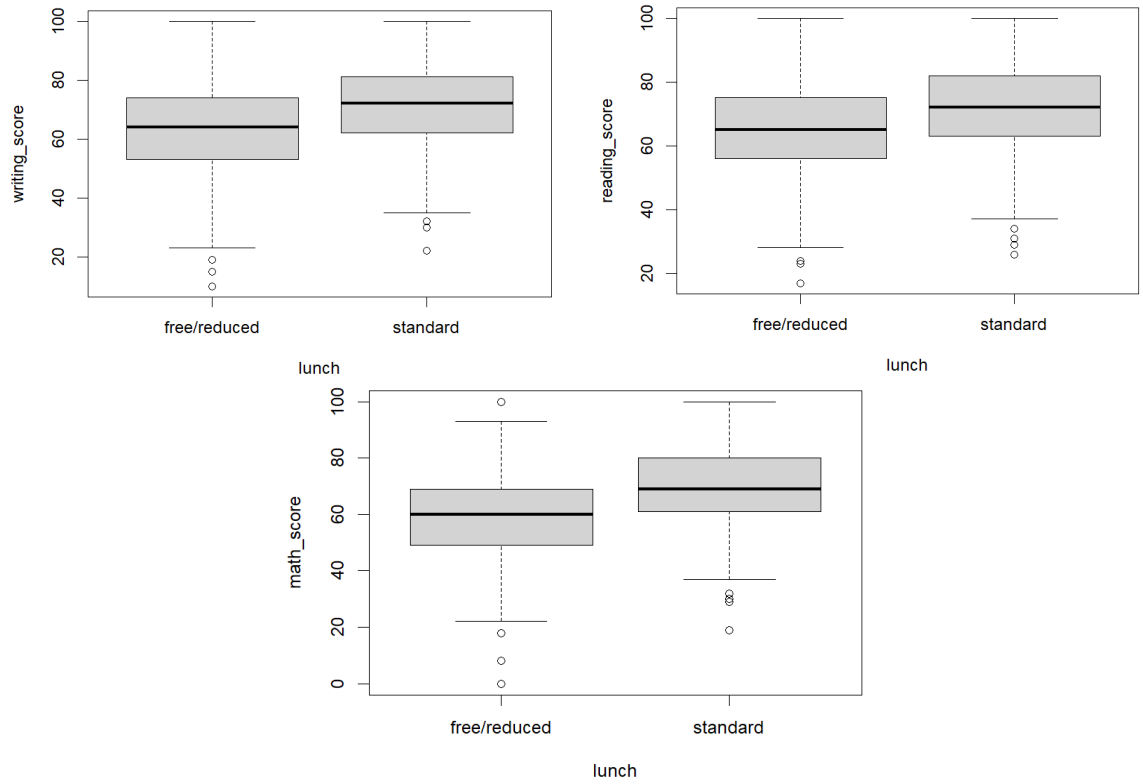
```
========================================================= Naive Bayes ==================

- Call: naive_bayes.formula(formula = data.parent_lvl_edu ~ ., data = train_stud)
- Laplace: 0
- Classes: 6
- Samples: 750
- Features: 3
- Conditional distributions:
    - Gaussian: 3
- Prior probabilities:
    - associate's degree: 0.216
    - bachelor's degree: 0.1227
    - high school: 0.2
    - master's degree: 0.052
    - some college: 0.228
    - some high school: 0.1813

-------------------------------------------------------------------------------------
```

```
> confusionMatrix(t_nb_stud5)
Confusion Matrix and Statistics

                    nb_stud5_pred
                     associate's degree bachelor's degree high school master's degree some college some high school
  associate's degree                 19                 9          21                8            1                0
  bachelor's degree                   9                 4           6                2            0                2
  high school                        26                 2          17                0            0                2
  master's degree                     5                 2           6                1            0                0
  some college                       29                 7          18                0            0                1
  some high school                   26                 8          19                0            0                0

Overall Statistics

               Accuracy : 0.164
                 95% CI : (0.1203, 0.2158)
    No Information Rate : 0.456
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0328

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: associate's degree Class: bachelor's degree Class: high school Class: master's degree
Sensitivity                             0.1667                   0.1250              0.1954              0.09091
Specificity                             0.7132                   0.9128              0.8160              0.94561
Pos Pred Value                          0.3276                   0.1739              0.3617              0.07143
Neg Pred Value                          0.5052                   0.8767              0.6552              0.95763
Prevalence                              0.4560                   0.1280              0.3480              0.04400
Detection Rate                          0.0760                   0.0160              0.0680              0.00400
Detection Prevalence                    0.2320                   0.0920              0.1880              0.05600
Balanced Accuracy                       0.4400                   0.5189              0.5057              0.51826
                     Class: some college Class: some high school
Sensitivity                       0.0000                  0.0000
Specificity                       0.7791                  0.7837
Pos Pred Value                    0.0000                  0.0000
Neg Pred Value                    0.9949                  0.9746
Prevalence                        0.0040                  0.0200
Detection Rate                    0.0000                  0.0000
Detection Prevalence              0.2200                  0.2120
Balanced Accuracy                 0.3896                  0.3918
> |
```

Rys. 29. Statystyki dla otrzymanego modelu, wyliczonego na podstawie zmiennej parent level education oraz macierz konfuzji.

Na podstawie otrzymanych wyników dla zmiennej parent level education otrzymano trafność wynoszącą ok. 20%, a wyniki czułości i specyficzności zostały podzielone na grupy. Najwyższy wynik czułości osiągnęła klasa o wykształceniu bachelor's degree, a najniższa dla some college i high school o wartości 0. Specyficzność najwyższą wartość osiągnęła dla tytułu master's degree, a najniższa dla associate's degree.

d) lunch



Rys. 30. Wykresy boxplot przedstawiające zależność wyników testu matematycznego, czytania i pisania od lunch.

Model:

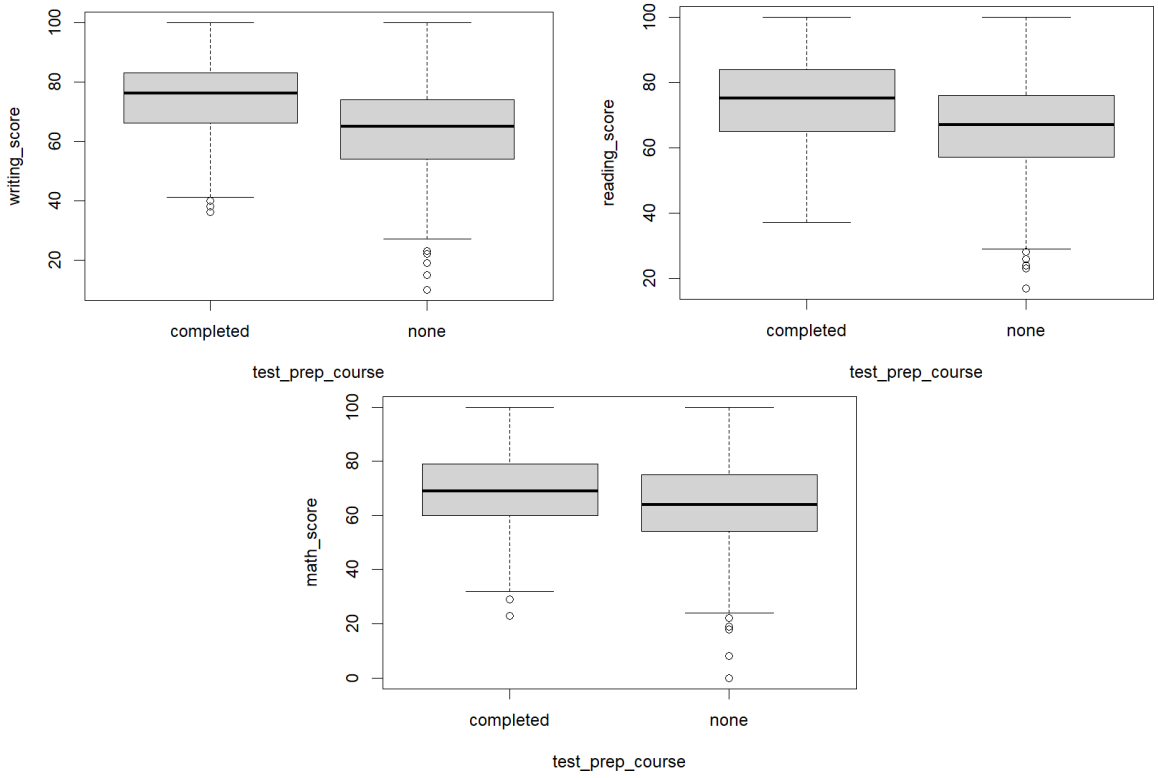
```
================================== Naive Bayes ==============================


- Call: naive_bayes.formula(formula = data.lunch ~ ., data = train_stud)
- Laplace: 0
- Classes: 2
- Samples: 750
- Features: 3
- Conditional distributions:
    - Gaussian: 3
- Prior probabilities:
    - free/reduced: 0.3493
    - standard: 0.6507
```

```
> confusionMatrix(t_nb_stud7)
Confusion Matrix and Statistics

              nb_stud7_pred
               free/reduced standard
  free/reduced           47       46
  standard               35      122

               Accuracy : 0.676
                 95% CI : (0.6142, 0.7336)
    No Information Rate : 0.672
    P-Value [Acc > NIR] : 0.4762

                  Kappa : 0.2894

 Mcnemar's Test P-Value : 0.2665

            Sensitivity : 0.5732
            Specificity : 0.7262
         Pos Pred Value : 0.5054
         Neg Pred Value : 0.7771
             Prevalence : 0.3280
         Detection Rate : 0.1880
   Detection Prevalence : 0.3720
      Balanced Accuracy : 0.6497

       'Positive' Class : free/reduced
```

Rys. 31. Statystyki dla otrzymanego modelu, wyliczonego na podstawie zmiennej lunch oraz macierz konfuzji.

Dla zmiennej lunch procent popranie sklasyfikowanych wartości wyniósł ok. 70%, specyficzność ok. 0.73, a czułość ok. 0.6, co daje zadawalające wyniki. Oznacza to, że zmienna lunch ma wpływ na wyniki przeprowadzonych testów sprawdzających, a pozytywną klasą jest free/reduced.

e) test preparation course



Rys. 32. Wykresy boxplot przedstawiające zależność wyników testu matematycznego, czytania i pisania od przygotowania.

Model:

```
================================== Naive Bayes ==================================

- Call: naive_bayes.formula(formula = data.test_prep_course ~ ., data = train_stud)
- Laplace: 0
- Classes: 2
- Samples: 750
- Features: 3
- Conditional distributions:
    - Gaussian: 3
- Prior probabilities:
    - completed: 0.3467
    - none: 0.6533

---------------------------------------------------------------------------------
```

```
> confusionMatrix(t_nb_stud9)
Confusion Matrix and Statistics

          nb_stud9_pred
           completed none
completed         53   45
none              48  104

               Accuracy : 0.628
                 95% CI : (0.5648, 0.6881)
    No Information Rate : 0.596
    P-Value [Acc > NIR] : 0.1669

                  Kappa : 0.2238

 Mcnemar's Test P-Value : 0.8357

            Sensitivity : 0.5248
            Specificity : 0.6980
         Pos Pred Value : 0.5408
         Neg Pred Value : 0.6842
             Prevalence : 0.4040
         Detection Rate : 0.2120
   Detection Prevalence : 0.3920
      Balanced Accuracy : 0.6114

       'Positive' Class : completed
```

Rys. 33. Statystyki dla otrzymanego modelu, wyliczonego na podstawie zmiennej test preparation course oraz macierz konfuzji.

W przypadku zmiennej test preparation course wynik trafności to ok. 60%, czułość ok. 50%, a specyficzność ok. 70%, co również daje zadawalające wyniki dla uzyskanego modelu. Przygotowanie do testu jak najbardziej ma wpływ na wynik przeprowadzonych testów. Pozytywna klasa to completed.

6. Podsumowanie oceny modelu

Najlepszy wyniki poprawnie sklasyfikowanych wartości uzyskany został dla modelu wykonanego na podstawie zmiennych gender, lunch oraz test preparation course. Dodatkowo suma czułości i specyficzności, czyli suma zdolności modelu do wychwytywania przypadków pozytywnych i zdolności modelu do wykrywania przypadków negatywnych dla tych zmiennych była różna niż jeden. Na podstawie tych stwierdzeń, możemy powiedzieć że płeć, lunch oraz przygotowanie do kursu ma wpływ na wyniki przeprowadzonych testów.

7. Model K-NN

Pierwszym krokiem analizy było skorzystanie z własnej funkcji do wykonywania standaryzacji. Wykonano podzbiory, gdzie dla zbioru testowego wybrano 25% wszystkich obserwacji, wybieranych losowo. Reszta została przypisana do zbioru treningowego. Kolejno wybrano odpowiednie klasy poszczególnych obserwacji w podzbiorach. Następnym krokiem było wyliczenie modelu, sprawdzono jego jakość, czułość i specyficzność oraz otrzymano macierz pomyłek.

Podobnie, jak w przypadku klasyfikatora Bayesa, postanowiono że modelowane odbędzie się w zależności od zmiennych typu char. Sprawdzono jak rozkładają się wyniki poszczególnych testów, biorąc pod uwagę płeć, grupę testu, wykształcenie oraz przygotowanie.

Przy budowaniu modelu sprawdzono, dla którego sąsiedztwa procent poprawnie sklasyfikowanych wartości będzie najwyższy, a następnie wybrano najbardziej optymistyczny model.

a) gender

Po przetestowaniu liczby sąsiedztw, wybrano to dla którego procent poprawnych klasyfikacji był najwyższy, czyli k=8.

```
> t_stud_g3<-table(test_stud_class,model_stud_g3)
> confusionMatrix(t_stud_g3)
Confusion Matrix and Statistics

               model_stud_g3
test_stud_class female male
         female    128   13
         male       15   94

               Accuracy : 0.888
                 95% CI : (0.8422, 0.9243)
    No Information Rate : 0.572
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7718

 Mcnemar's Test P-Value : 0.8501

            Sensitivity : 0.8951
            Specificity : 0.8785
         Pos Pred Value : 0.9078
         Neg Pred Value : 0.8624
             Prevalence : 0.5720
         Detection Rate : 0.5120
   Detection Prevalence : 0.5640
      Balanced Accuracy : 0.8868

       'Positive' Class : female
```

Rys. 34. Macierz konfuzji wyliczona na podstawie zmiennej gender.

Model ma wysoki procent poprawnie sklasyfikowanych wartości: ok. 90%, wysoką czułość: ok. 0.9 oraz specyficzność: ok. 0.9. Oznacza to, że dobrze poradził sobie z wychwyceniem przypadków pozytywnych, jak i negatywnych. Na podstawie uzyskanego modelu, widzimy że płeć ma wpływ na wyniki testów, a pozytywną klasą jest female.

b) race/enthicity

Najlepszy wynik dla k=9.

```
> confusionMatrix(t_stud_r3)
Confusion Matrix and Statistics

               model_stud_r3
test_stud_class group A group B group C group D group E
        group A       1       3      11       8       1
        group B       0       6      26       9       3
        group C       1      16      39      20      11
        group D       0       9      30      16       3
        group E       1       6      11      10       9

Overall Statistics

               Accuracy : 0.284
                 95% CI : (0.229, 0.3442)
    No Information Rate : 0.468
    P-Value [Acc > NIR] : 1.00000

                  Kappa : 0.0237

 Mcnemar's Test P-Value : 0.00151

Statistics by Class:

                     Class: group A Class: group B Class: group C Class: group D Class: group E
Sensitivity              0.33333        0.1500         0.3333         0.2540         0.3333
Specificity              0.90688        0.8190         0.6391         0.7754         0.8744
Pos Pred Value           0.04167        0.1364         0.4483         0.2759         0.2432
Neg Pred Value           0.99115        0.8350         0.5215         0.7552         0.9155
Prevalence               0.01200        0.1600         0.4680         0.2520         0.1080
Detection Rate           0.00400        0.0240         0.1560         0.0640         0.0360
Detection Prevalence     0.09600        0.1760         0.3480         0.2320         0.1480
Balanced Accuracy        0.62011        0.4845         0.4862         0.5147         0.6039
```

Rys. 35. Macierz konfuzji wyliczona na podstawie zmiennej race/enthicity.

W przypadku modelowania na podstawie zmiennej race/ethnicity, trafność zmalała w stosunku do zmiennej gender i wynosi ok. 30%. Czułość i specyficzność z kolei została przedstawiona dla konkretnych klas. Czułość najwyższa jest dla grupy A, C oraz E i wynosi 0.33, a najniższa dla grupy B. Specyficzność najwyższa jest grupy A, a najniższa dla grupy C. We wszystkich grupach najlepiej wychwycone zostały przypadki negatywne.

Biorąc pod uwagę obie te wartości, można zauważyć że suma czułości i specyficzności oscyluje w granicach 1 dla większości grup. Może wskazywać na to, że test nie ma wartości, a przynależność do grupy nie ma wpływu na wynik testu

c) parent level education

Dla k=14

```
> confusionMatrix(t_stud_p3)
Confusion Matrix and Statistics

                    model_stud_p3
test_stud_class      associate's degree bachelor's degree high school master's degree some college
  associate's degree                 19                 3           9                0           14
  bachelor's degree                  12                 3           7                1           10
  high school                        16                 1          11                0           12
  master's degree                     3                 3           2                0            4
  some college                       13                 4          11                0           18
  some high school                    9                 1          18                1            9
                    model_stud_p3
test_stud_class      some high school
  associate's degree               10
  bachelor's degree                 4
  high school                       7
  master's degree                   3
  some college                      6
  some high school                  6

Overall Statistics

               Accuracy : 0.228
                 95% CI : (0.1775, 0.2851)
    No Information Rate : 0.288
    P-Value [Acc > NIR] : 0.986335

                  Kappa : 0.0381

 Mcnemar's Test P-Value : 0.004995

Statistics by Class:

                     Class: associate's degree Class: bachelor's degree Class: high school
Sensitivity                             0.2639                  0.20000             0.1897
Specificity                             0.7978                  0.85532             0.8125
Pos Pred Value                          0.3455                  0.08108             0.2340
Neg Pred Value                          0.7282                  0.94366             0.7685
Prevalence                              0.2880                  0.06000             0.2320
Detection Rate                          0.0760                  0.01200             0.0440
Detection Prevalence                    0.2200                  0.14800             0.1880
Balanced Accuracy                       0.5308                  0.52766             0.5011
                     Class: master's degree Class: some college Class: some high school
Sensitivity                          0.0000              0.2687                 0.1667
Specificity                          0.9395              0.8142                 0.8224
Pos Pred Value                       0.0000              0.3462                 0.1364
Neg Pred Value                       0.9915              0.7525                 0.8544
Prevalence                           0.0080              0.2680                 0.1440
Detection Rate                       0.0000              0.0720                 0.0240
Detection Prevalence                 0.0600              0.2080                 0.1760
Balanced Accuracy                    0.4698              0.5414                 0.4945
>
```

Rys. 36. Macierz konfuzji wyliczona na podstawie zmiennej parent level education.

Na podstawie otrzymanych wyników dla zmiennej parent level education otrzymano trafność wynoszącą ok. 20%, a wyniki czułości i specyficzności zostały podzielone na grupy. Najwyższy wynik czułości osiągnęła klasa o wykształceniu associate's degree oraz some college, a najniższa dla master's degree o wartości 0. Specyficzność najwyższą wartość osiągnęła dla tytułu master's degree, a najniższa dla associate's degree.

d) lunch

Najlepszy wynik gdy k=5

```
> t_stud_l3<-table(test_stud_class,model_stud_l3)
> confusionMatrix(t_stud_l3)
Confusion Matrix and Statistics

               model_stud_l3
test_stud_class free/reduced standard
   free/reduced           36       44
   standard               26      144

               Accuracy : 0.72
                 95% CI : (0.6599, 0.7747)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : 0.89229

                  Kappa : 0.3159

 Mcnemar's Test P-Value : 0.04216

            Sensitivity : 0.5806
            Specificity : 0.7660
         Pos Pred Value : 0.4500
         Neg Pred Value : 0.8471
             Prevalence : 0.2480
         Detection Rate : 0.1440
   Detection Prevalence : 0.3200
      Balanced Accuracy : 0.6733

       'Positive' Class : free/reduced
```

Rys. 37. Macierz konfuzji wyliczona na podstawie zmiennej lunch.

Dla zmiennej lunch procent popranie sklasyfikowanych wartości wyniósł ok. 70%, specyficzność ok. 0.77, a czułość ok. 0.6, co daje zadawalające wyniki. Oznacza to, że zmienna lunch ma wpływ na wyniki przeprowadzonych testów sprawdzających, a pozytywną klasą jest free/reduced.

e) test_prep_course

```
> t_stud_t3<-table(test_stud_class,model_stud_t3)
> confusionMatrix(t_stud_t3)
Confusion Matrix and Statistics

               model_stud_t3
test_stud_class completed none
      completed        33   62
      none             36  119

               Accuracy : 0.608
                 95% CI : (0.5445, 0.6689)
    No Information Rate : 0.724
    P-Value [Acc > NIR] : 0.99997

                  Kappa : 0.1215

 Mcnemar's Test P-Value : 0.01156

            Sensitivity : 0.4783
            Specificity : 0.6575
         Pos Pred Value : 0.3474
         Neg Pred Value : 0.7677
             Prevalence : 0.2760
         Detection Rate : 0.1320
   Detection Prevalence : 0.3800
      Balanced Accuracy : 0.5679

       'Positive' Class : completed
```

Rys. 38. Macierz konfuzji wyliczona na podstawie zmiennej test preparation course.

W przypadku zmiennej test preparation course wynik trafności to ok. 60%, czułość ok. 50%, a specyficzność ok. 66%, co również daje zadawalające wyniki dla uzyskanego modelu. Przygotowanie do testu jak najbardziej ma wpływ na wynik przeprowadzonych testów. Pozytywna klasa to completed.

8. Podsumowanie oceny modelu K-NN

Najlepszy wyniki poprawnie sklasyfikowanych wartości uzyskany został dla modelu wykonanego na podstawie zmiennych gender, lunch oraz test preparation course. Dodatkowo suma czułości i specyficzności, czyli suma zdolności modelu do wychwytywania przypadków pozytywnych i zdolności modelu do wykrywania przypadków negatywnych dla tych zmiennych była różna niż jeden. Na podstawie tych stwierdzeń, możemy powiedzieć że płeć, lunch oraz przygotowanie do kursu ma wpływ na wyniki przeprowadzonych testów.

9. Ogólne wnioski

Porównując otrzymane modele dla klasyfikatora Bayesa oraz dla K-NN, najwyższa liczba poprawnie sklasyfikowanych wartości, a co za tym idzie wysokie wartości dla specyficzności i czułości uzyskano dla tych samych zmiennych tj. gender, lunch oraz test preaparation course. Oznacza to, że model K-NN potwierdził stwierdzenie uzyskane na podstawie modelu Naiwnego Klasyfikatora Bayesa. Wyniki testów sprawdzających zależą od płci, lunchu oraz przygotowania. Dodatkowo także dla wymienionych zmiennych zdolność modelu do wychwytywania przypadków pozytywnych jak i negatywnych była największa.

Najniższa skuteczność dla obu klasyfikatorów uzyskały modele zbudowane na podstawie zmiennych parent level education oraz race/enthicity. W kilku przypadkach zaobserwowano, że suma wartości czułości i skuteczności jest równa jeden, dlatego można powiedzieć, że testy na podstawie tych zmiennych mają słabą wartość i nie mają wpływu na wyniki analizowanych testów sprawdzających.