Zadanie projektowe 2

Porównanie klasyfikatorów na wybranej bazie danych

Mateusz Wirzba
253970

Wybór bazy danych

Na bazę danych wybrałem bazę Student Performance ze strony:
https://www.kaggle.com/spscientist/studentsperformance-in-exams

Link do repozytorium z kodem projektu:

https://github.com/mwirzba/Projekt_2

Badanie i obróbka bazy danych

Baza danych nie zawiera żadnych nieprawidłowych informacji.

Aby móc wykorzystać wartości typu string należało je zamienić na wartości binarne:
Użyto do tego metody LabelEncoder z paczki sklearn.

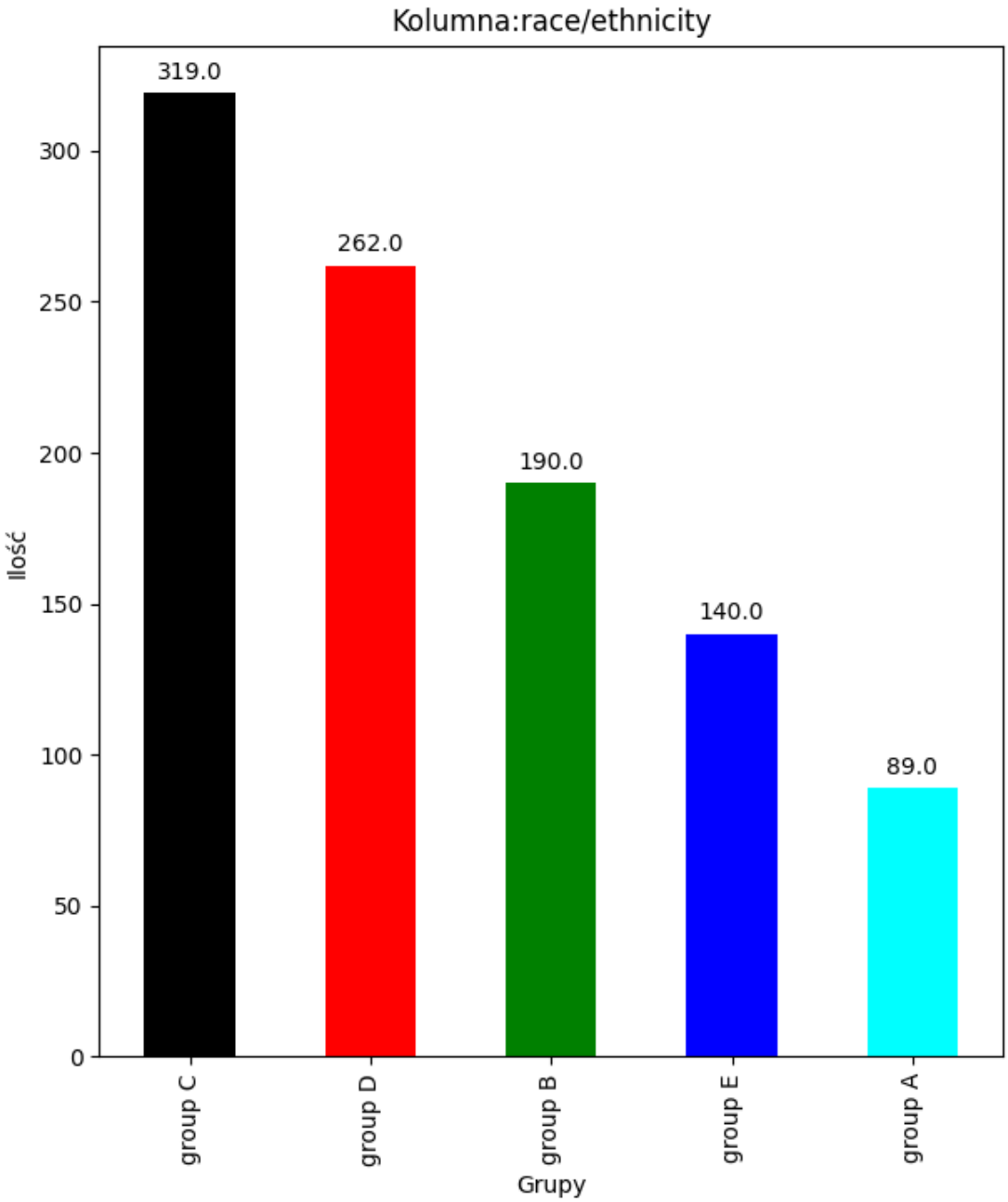
```
def encode_column_values_to_bit(df: DataFrame, column_name: str):
    le = preprocessing.LabelEncoder()
    le.fit(df[column_name].astype(str))
    return le.transform(df[column_name].astype(str))
```

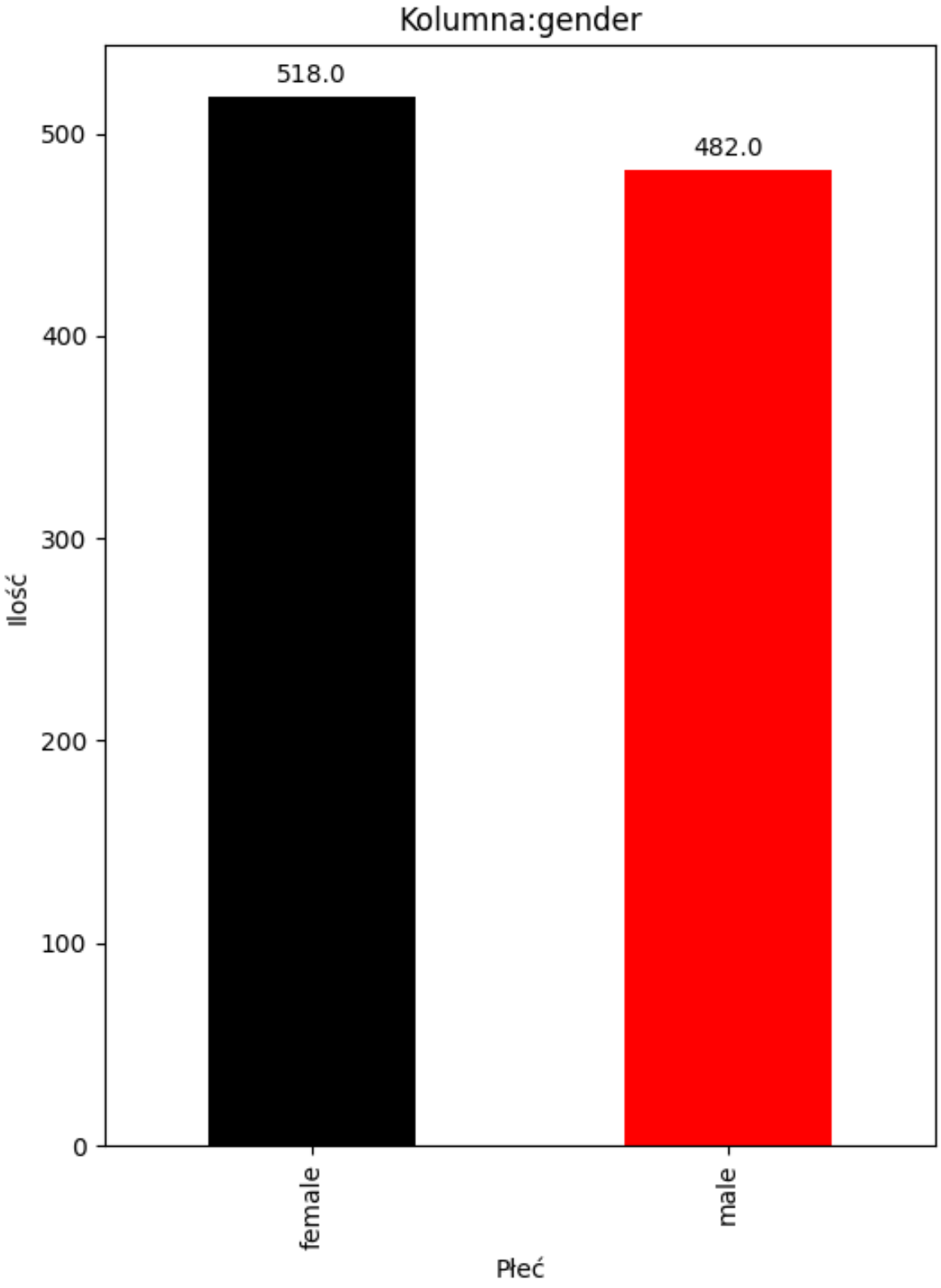
Kolumny w bazie danych:

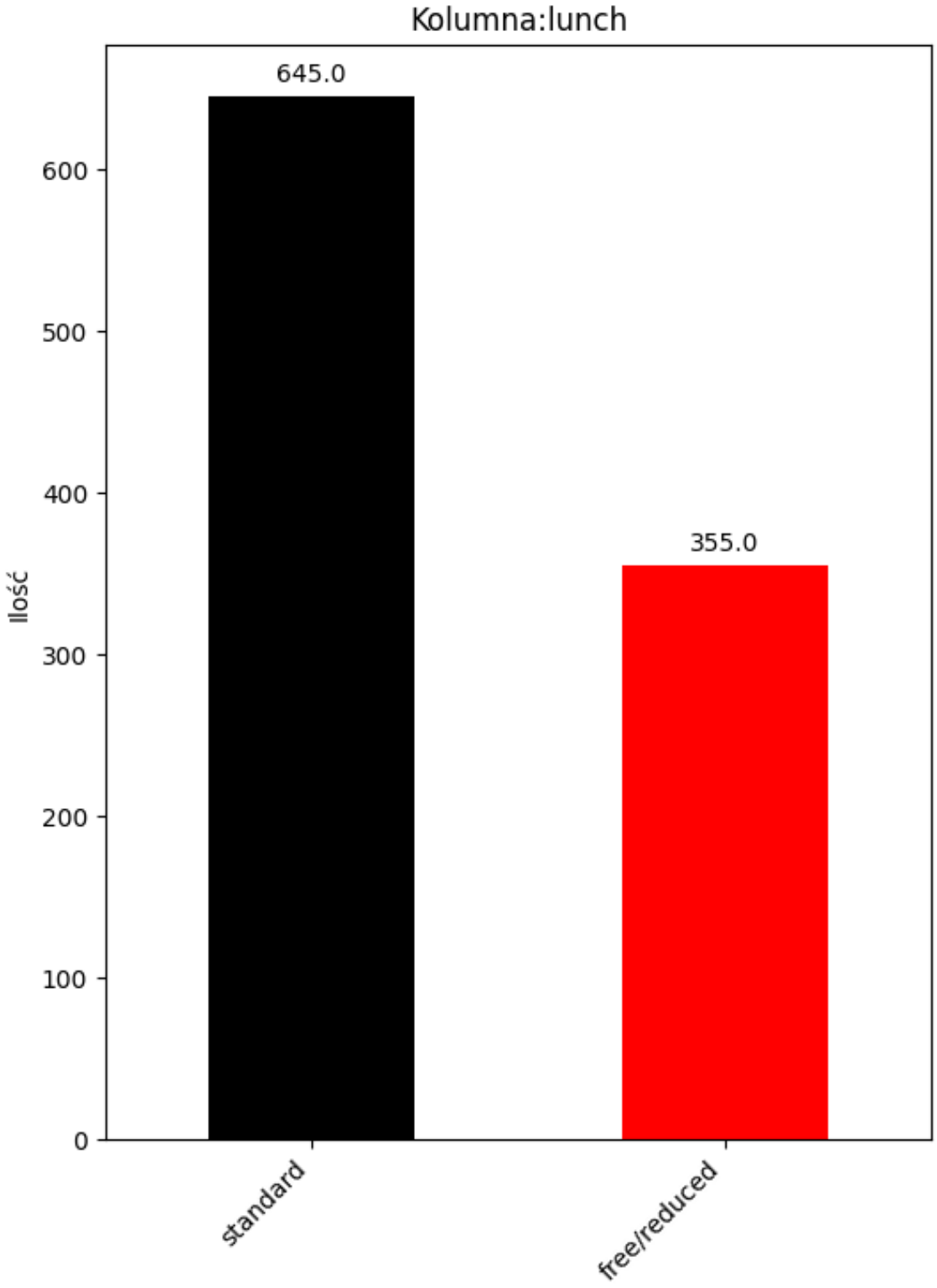
1. Gender - płeć studenta
2. race/etnicity - rasa/grupa etniczna studenta
3. parental level of education -
4. lunch - informacja jaki student zjadł posiłek przed testem
5. test preperation couse – informacja czy student odbył kurs przygotowawczy do testu
6. math score – wyniki testu z matematyki
7. reading score – wyniki testu z czytania
8. writing score – wyniki testu z pisania

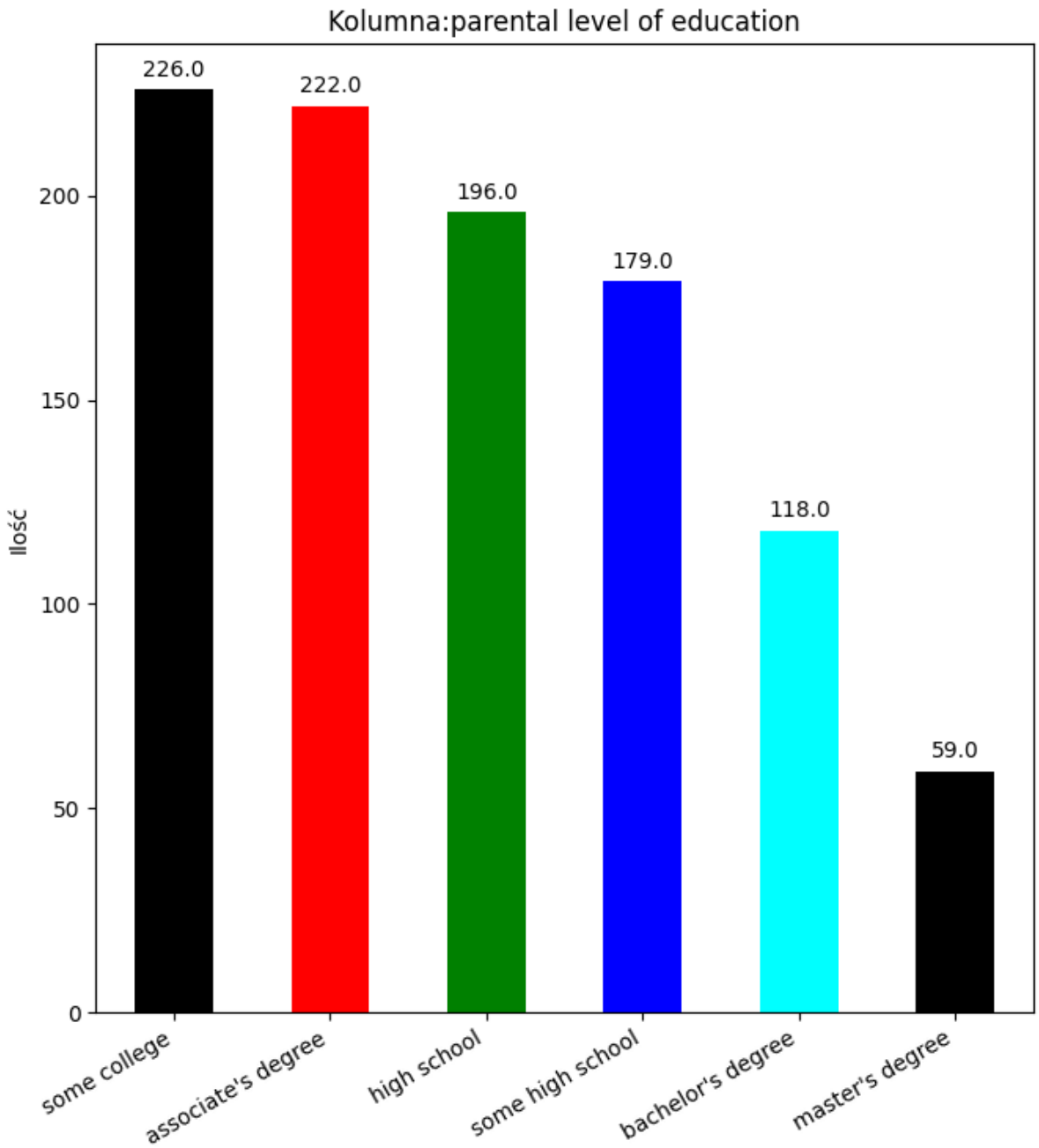
Kolumna	**Min**	**Max**	**Średnia**
Gender	NAN	NAN	NAN
race/etnicity	NAN	NAN	NAN
parental level of education	NAN	NAN	NAN
lunch	NAN	NAN	NAN
test preperation couse	NAN	NAN	NAN
math score	0	100	66.089
reading score	17	100	69.169
writing score	10	100	68.023

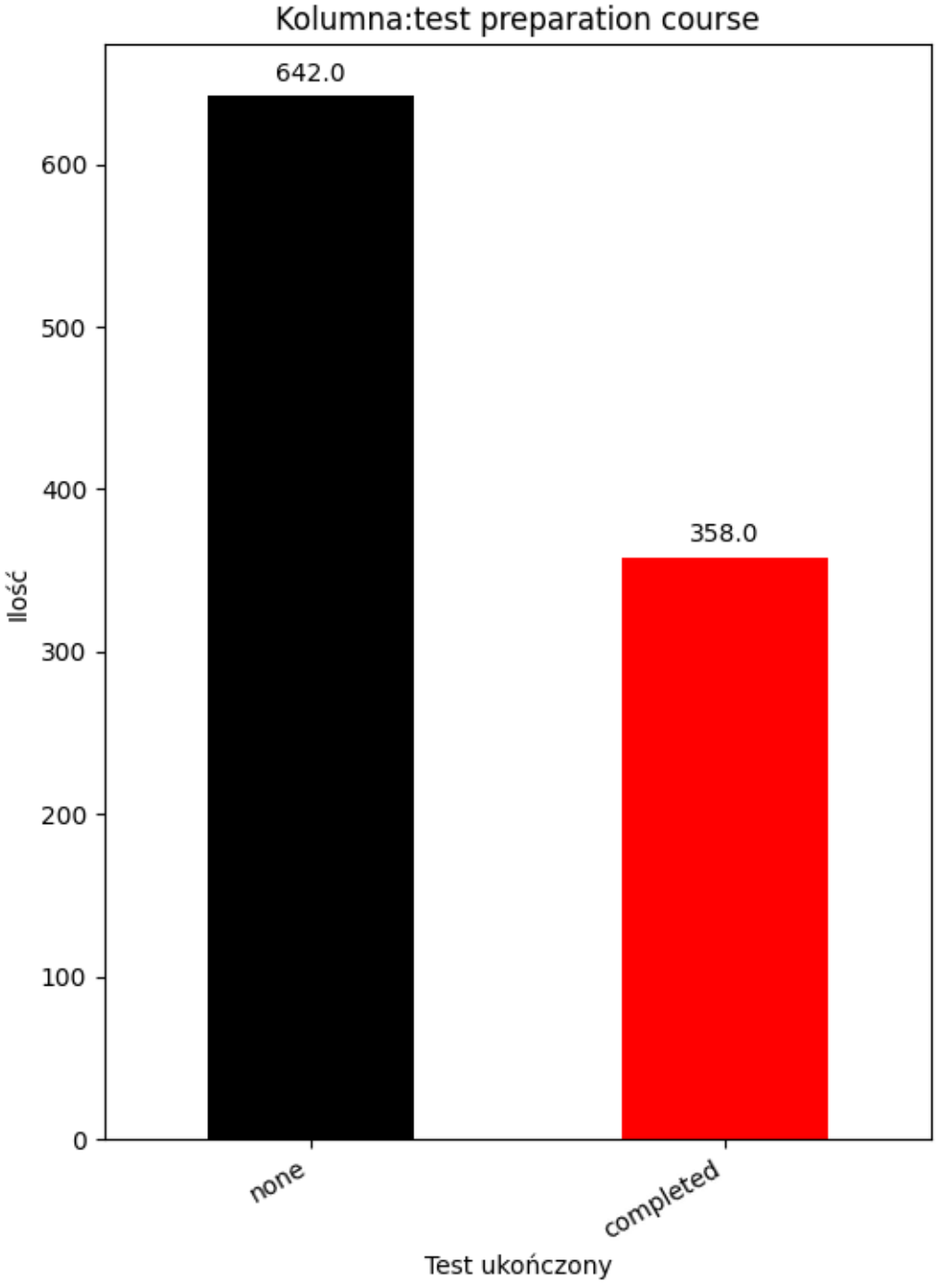
Częstotliwość występowania danych

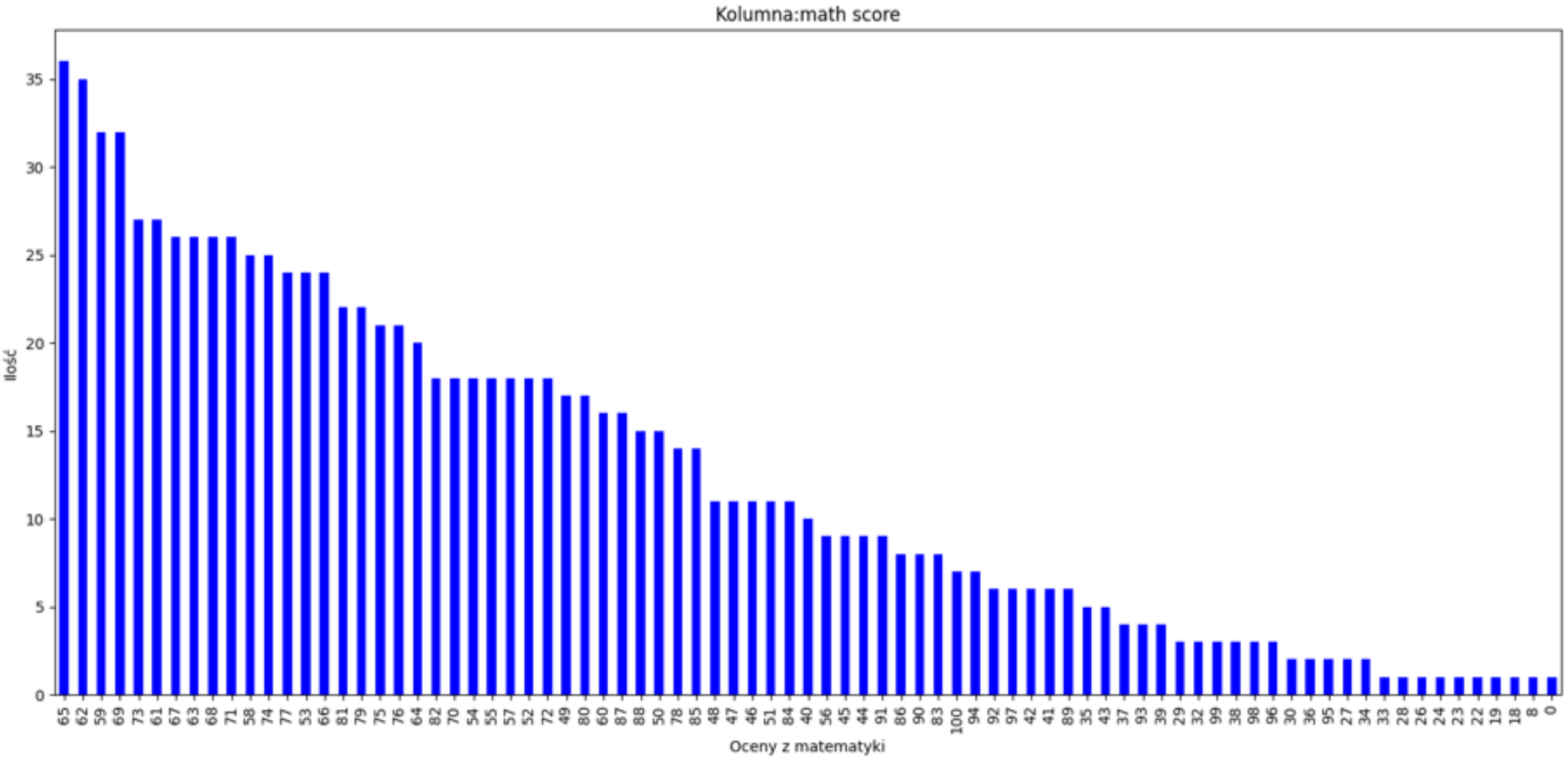


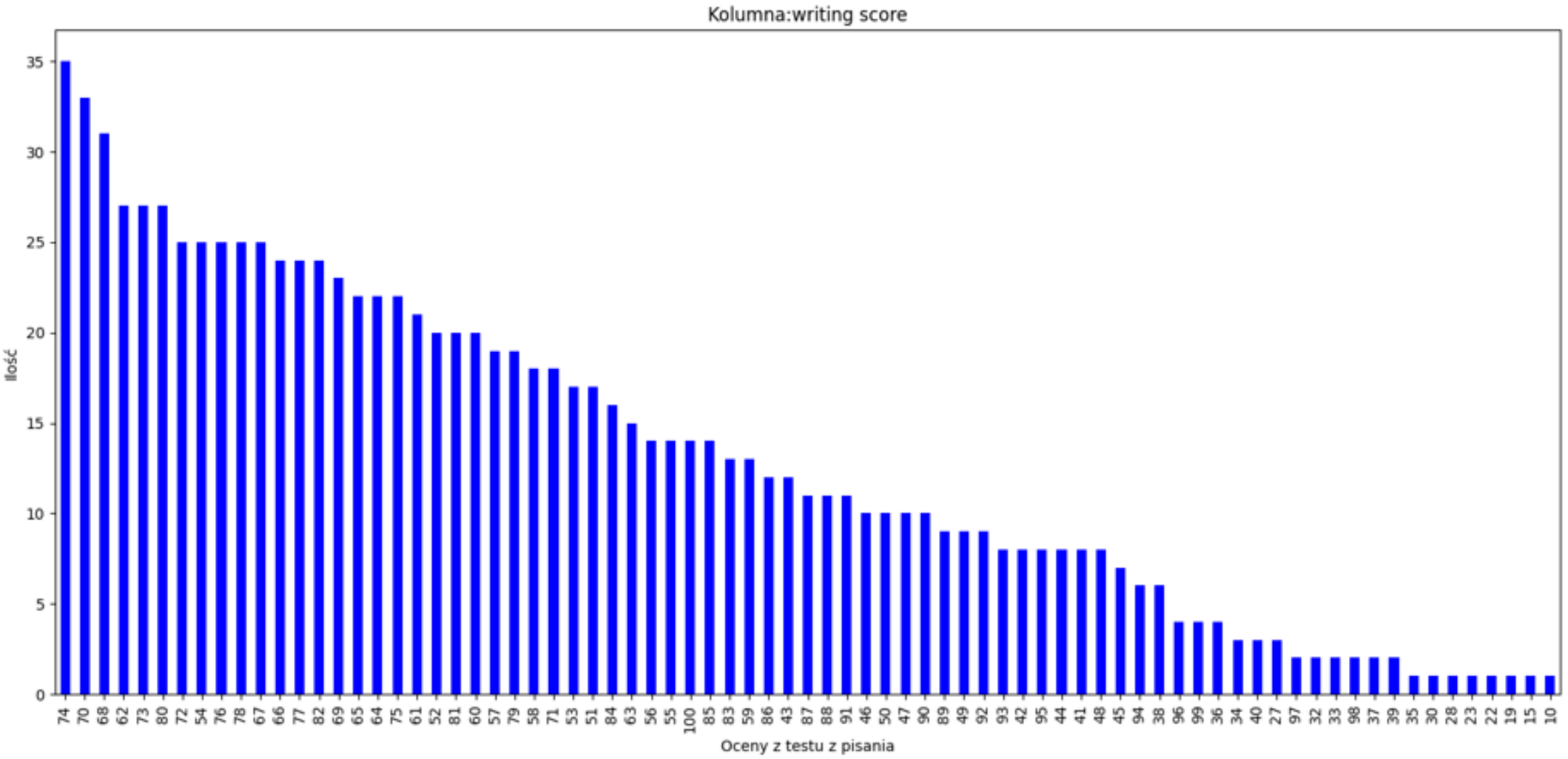

Kolumna:gender
518.0
482.0
500
400
300
200
100
0
Ilość
female
male
Płeć

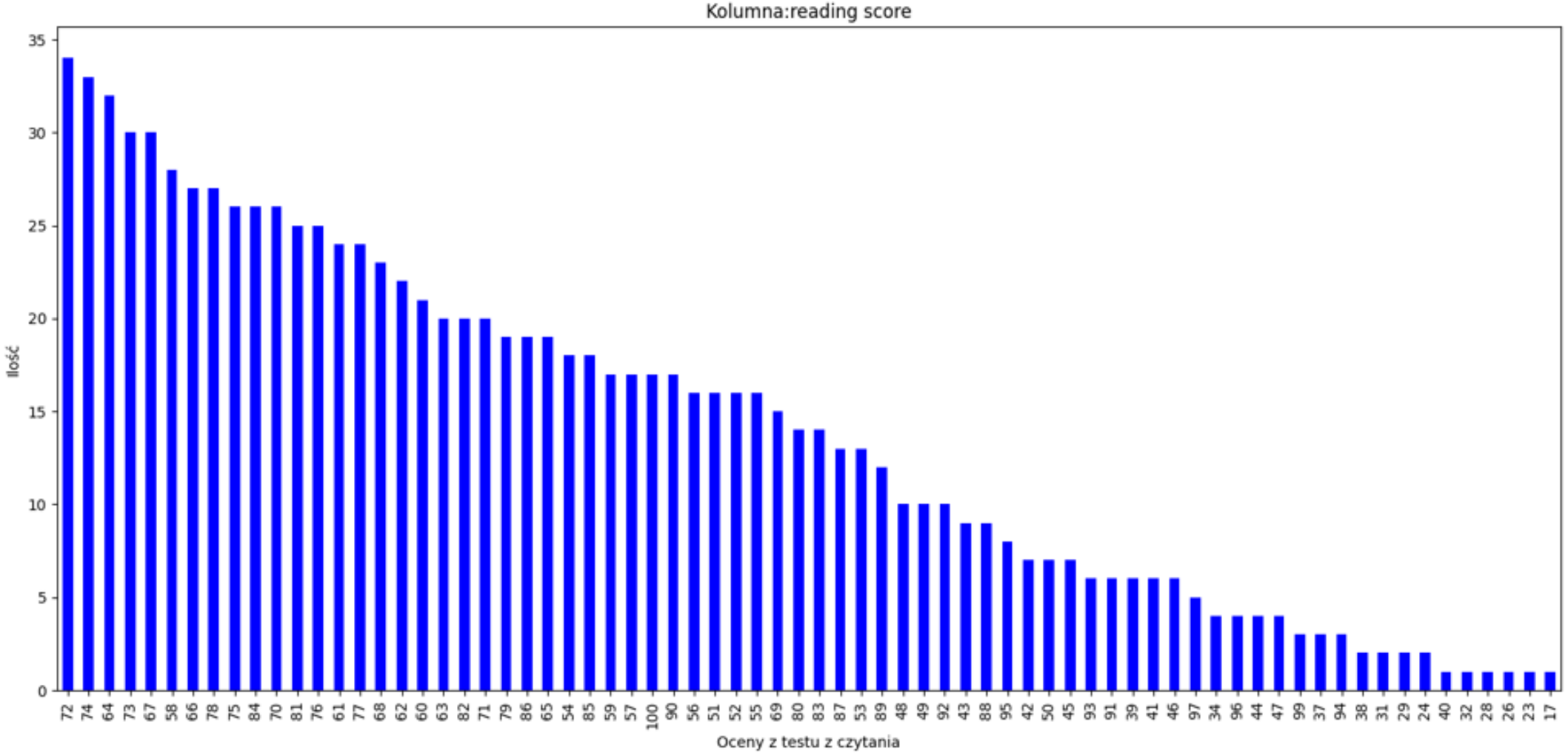

Kolumna:lunch
645.0
355.0
600
500
400
300
200
100
0
Ilość
standard
free/reduced


Kolumna:parental level of education
Ilość
226.0
222.0
196.0
179.0
118.0
59.0
0
50
100
150
200
some college
associate's degree
high school
some high school
bachelor's degree
master's degree


Kolumna:test preparation course
642.0
358.0
600
500
400
300
200
100
0
Ilość
none
completed
Test ukończony


Kolumna:math score
Ilość
0
5
10
15
20
25
30
35
65 62 59 69 73 61 67 63 68 71 58 74 77 53 66 81 79 75 76 64 82 70 54 55 57 52 72 49 80 60 87 88 50 78 85 48 47 46 51 84 40 56 45 44 91 86 90 83 100 94 92 97 42 41 89 35 43 37 93 39 29 32 99 38 98 96 30 36 95 27 34 33 28 26 24 23 22 19 18 8 0
Oceny z matematyki


Kolumna:writing score
Ilość
0
5
10
15
20
25
30
35
74 70 68 62 73 80 72 54 76 78 67 66 77 82 69 65 64 75 61 52 81 60 57 79 58 71 53 51 84 63 56 55 100 85 83 59 86 43 87 88 91 46 50 47 90 89 49 92 93 42 95 44 41 48 45 94 38 96 99 36 34 40 27 97 32 33 98 37 39 35 30 28 23 22 19 15 10
Oceny z testu z pisania

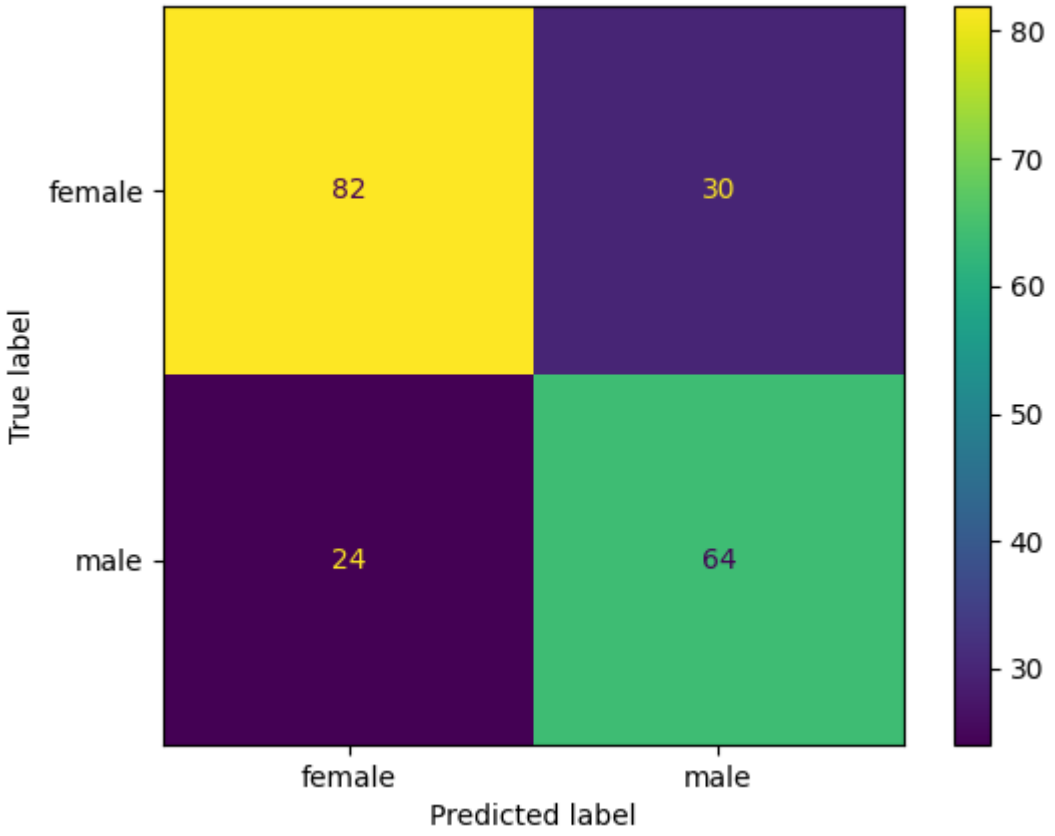

Kolumna:reading score
Ilość
0
5
10
15
20
25
30
35
72 74 64 73 67 58 66 78 75 84 70 81 76 61 77 68 62 60 63 82 71 79 86 65 54 85 59 57 100 90 56 51 52 55 69 80 83 87 53 89 48 49 92 43 88 95 42 50 45 93 91 39 41 46 97 34 96 44 47 99 37 94 38 31 29 24 40 32 28 26 23 17
Oceny z testu z czytania

Porównanie poznanych klasyfikatorów.

Jako klasę wybrano płeć.
Będziemy próbowali zgadnąć płeć studenta na podstawie danych.
Baza danych posiada tylko 1000 rekordów.
Z tego powodu przeznaczyłem 80% na zbiory treningowe i 20% na testowe.

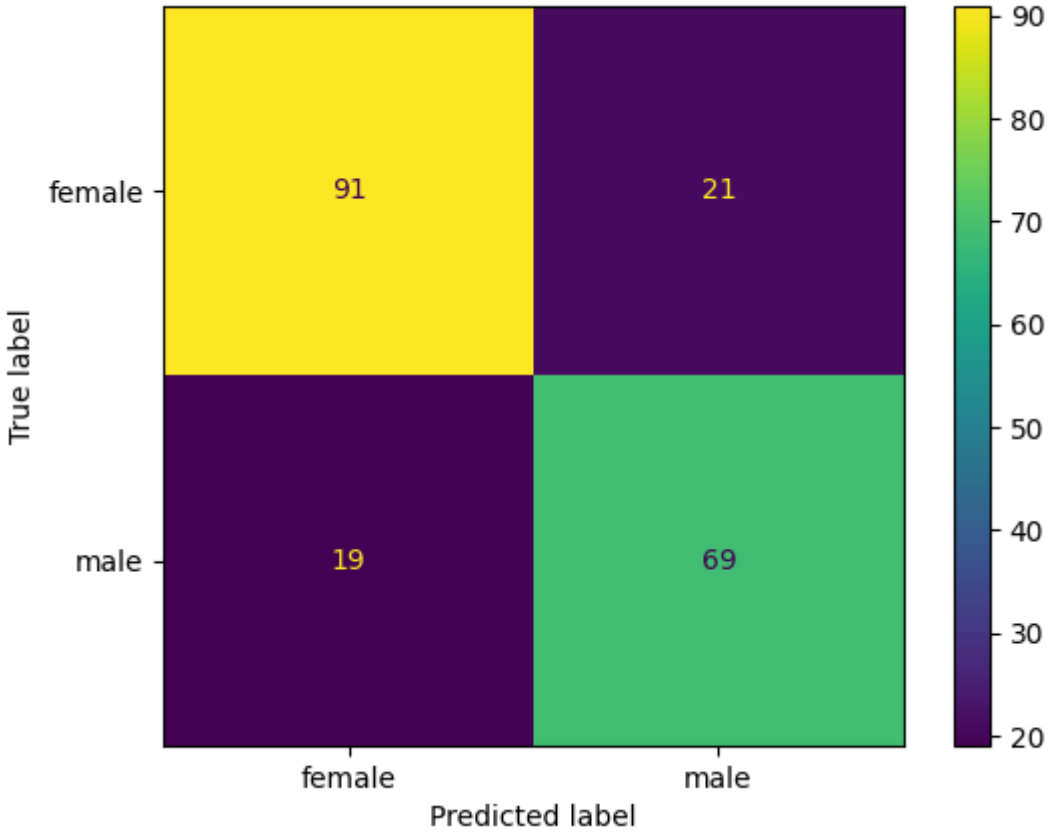
Naive Bayes

Procent poprawnych odpowiedzi: 73%



Drzewa decyzyjne

Procent poprawnych odpowiedzi: 80.5%

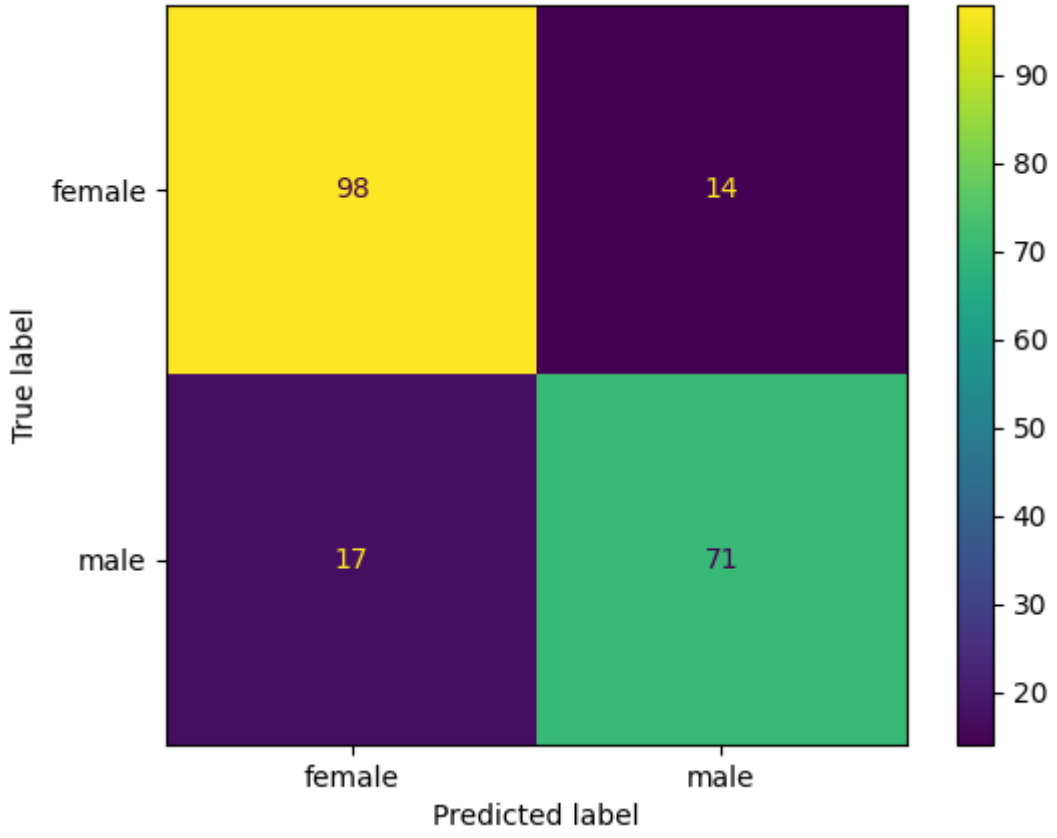


K najbliższych sąsiadów

Ilość sąsiadów	Procent poprawnych wyników
10	0.835
20	0.84
30	0.84
40	0.835
50	0.83
60	0.83
70	**0.845**
80	0.82
90	0.815
100	0.82
150	0.78
200	0.775

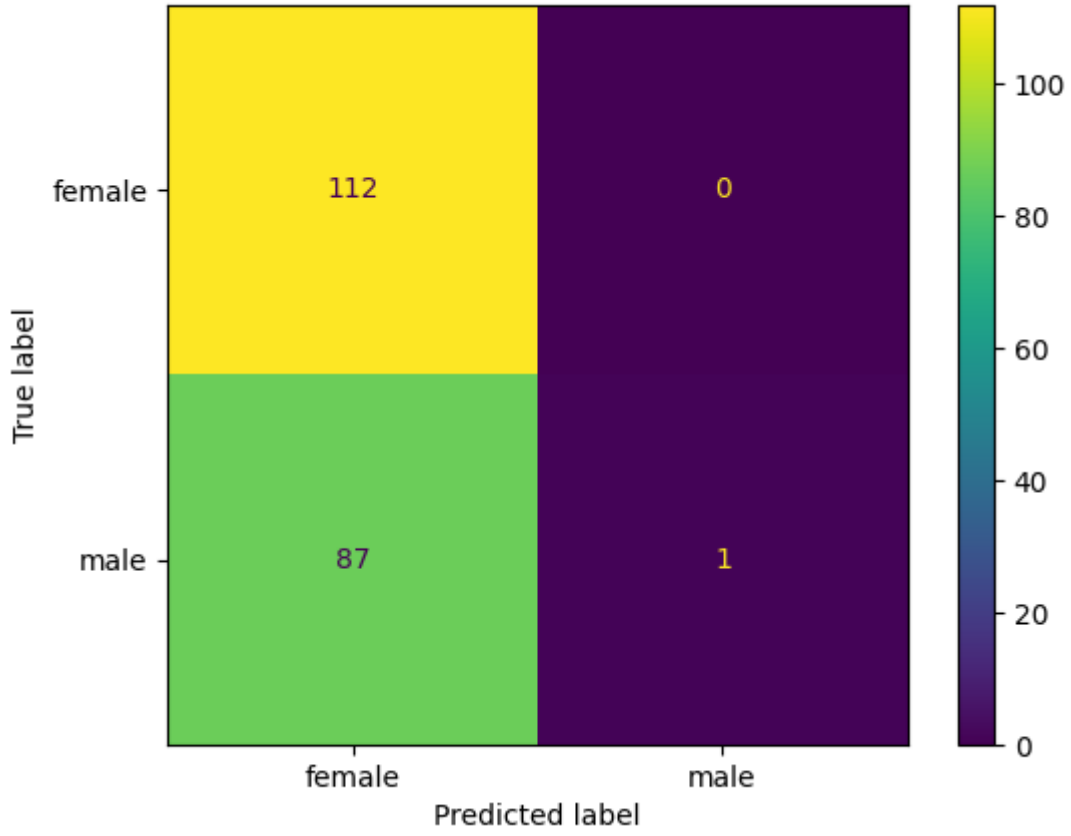
Wniosek: Najlepsze K najbliższych sąsiadów to 70

Macierz błędu dla 70 najbliższych sąsiadów.



Sieci neuronowe

Procent poprawnych odpowiedzi: 90%



Podsumowanie

Klasyfikator	**Procent poprawnych odpowiedzi**
Naive Bayes	73
Drzewa decyzyjne	80
K najbliższych sąsiadów	84.5
Sieci neuronowe	**90.5**

Najlepszym klasyfikatorem okazał się klasyfikator sieci neuronowych.