

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.

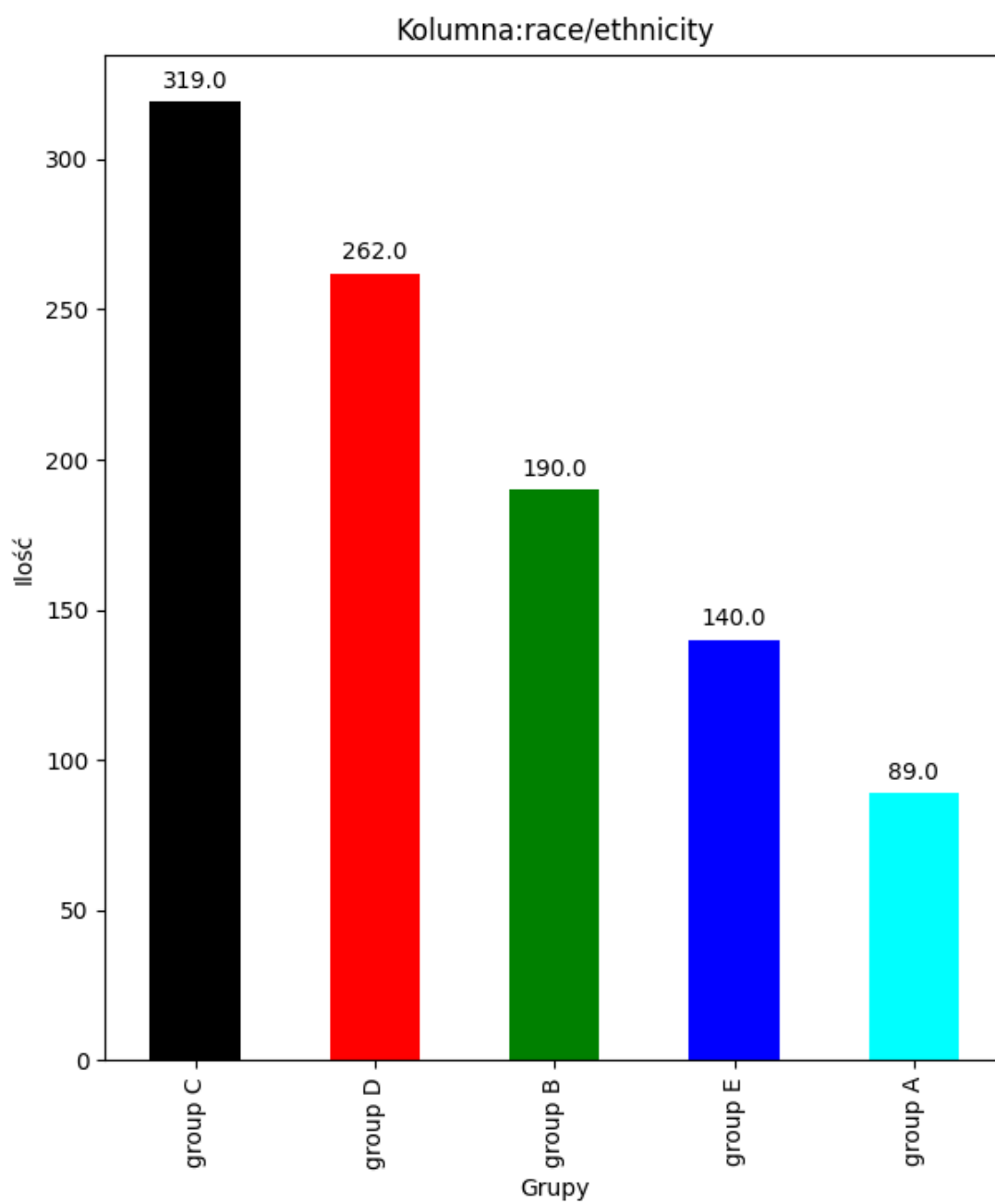
```
def encode_column_values_to_bit(df: DataFrame, column_name: str):  
    le = preprocessing.LabelEncoder()  
    le.fit(df[column_name].astype(str))  
    return le.transform(df[column_name].astype(str))
```

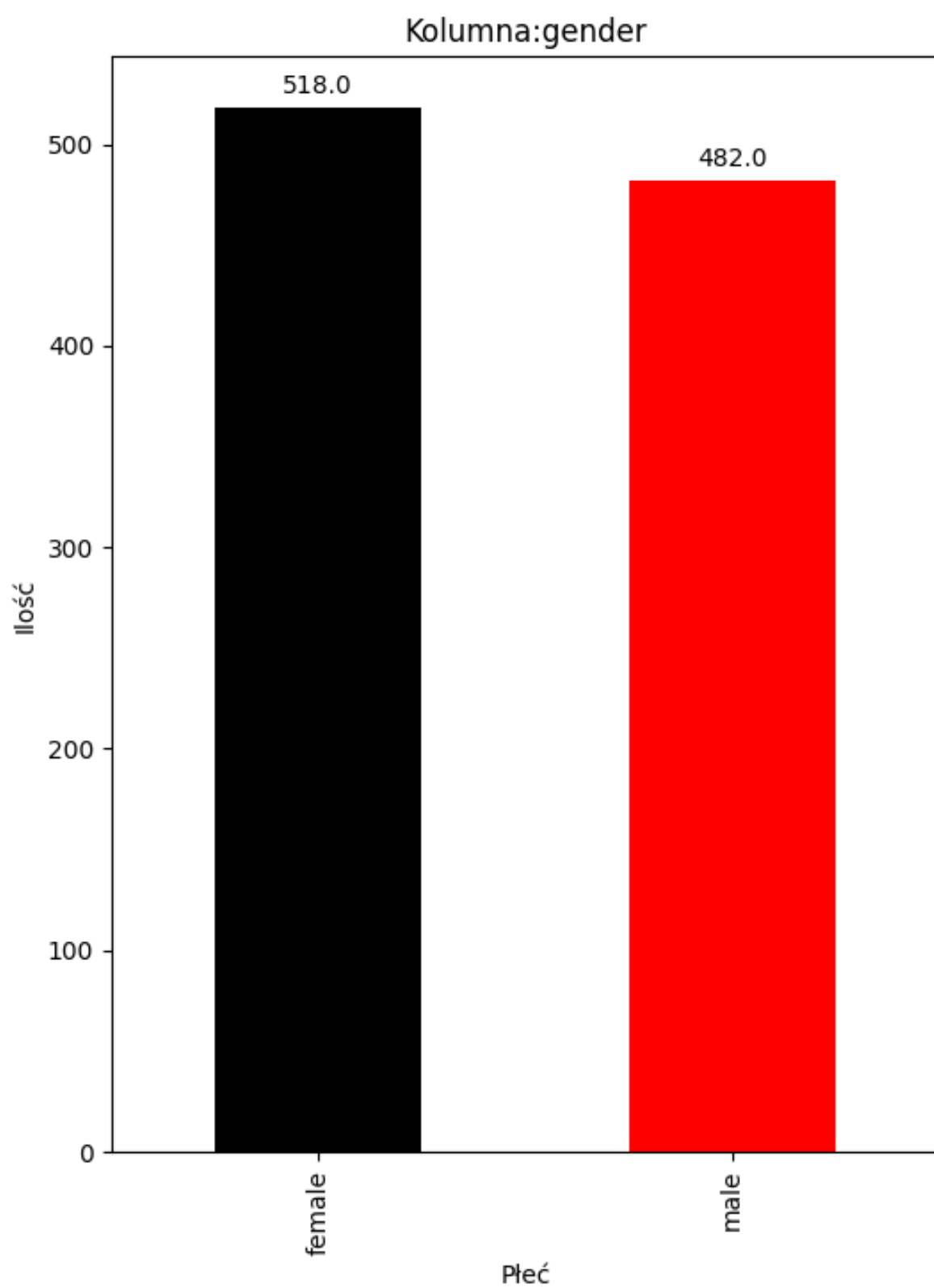
Kolumny w bazie danych:

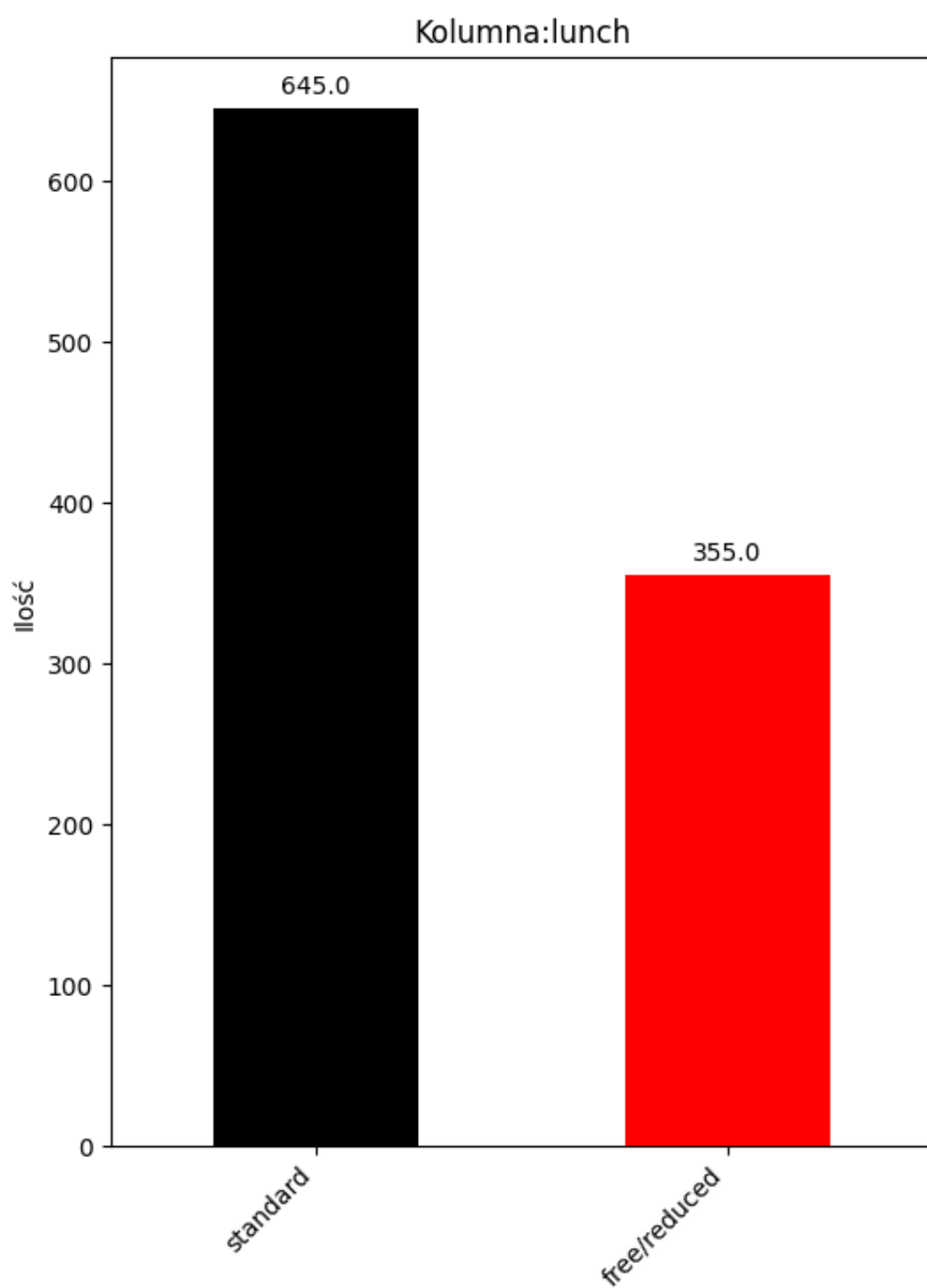
1. Gender - płeć studenta
2. race/ethnicity - rasa/grupa etniczna studenta
3. parental level of education -
4. lunch - informacja jaki student zjadł posiłek przed testem
5. test preperation couse – informacja czy student odbył kurs przygotowawczy do testu
6. math score – wyniki testu z matematyki
7.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

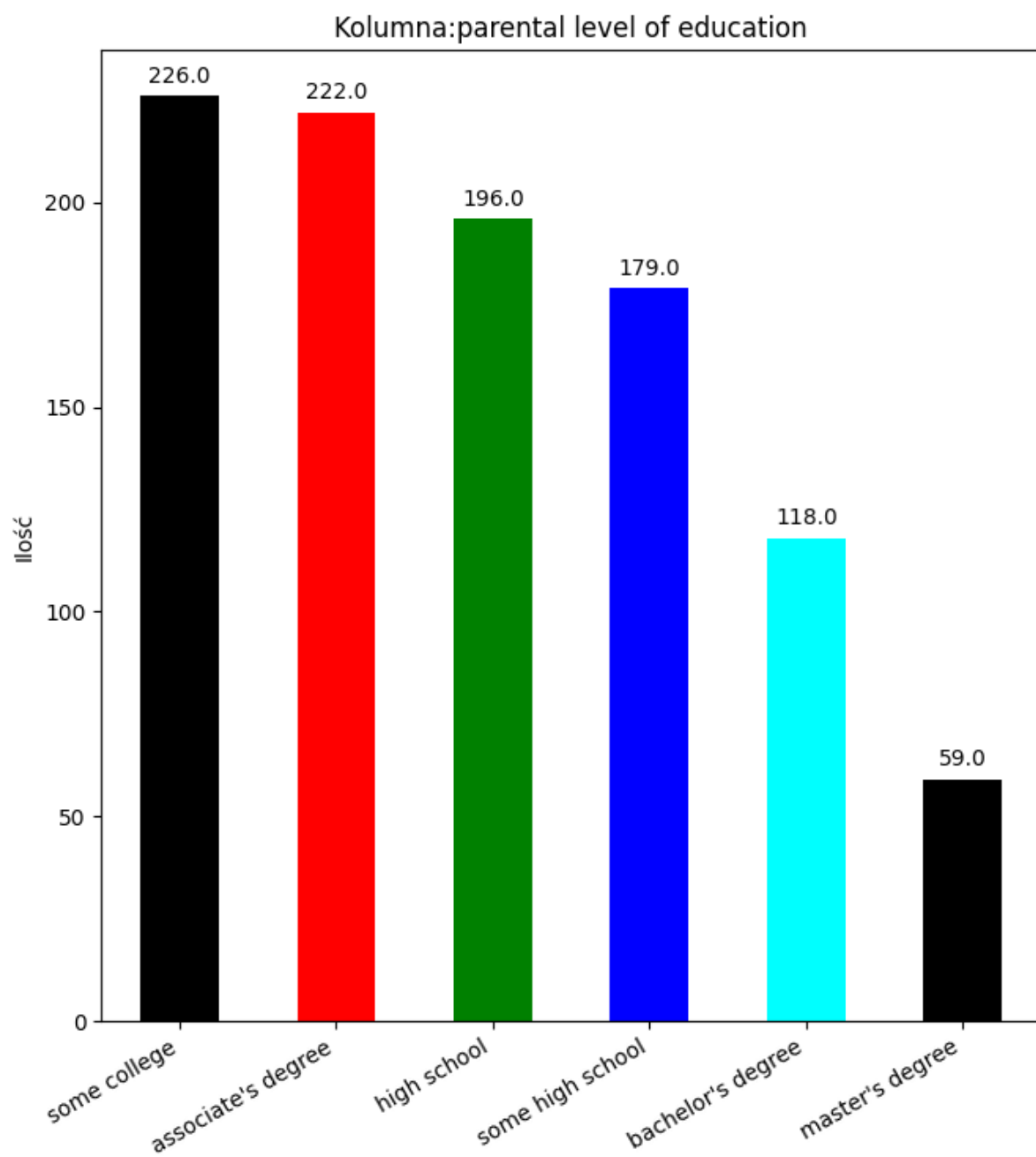
Kolumna	Min	Max	Średnia
Gender	NAN	NAN	NAN
race/ethnicity	NAN	NAN	NAN
parental level of education	NAN	NAN	NAN
lunch	NAN	NAN	NAN
test preperation couse	NAN	NAN	NAN
math score	0	100	66.089
reading score	17	100	69.169
writing score	10	100	68.023

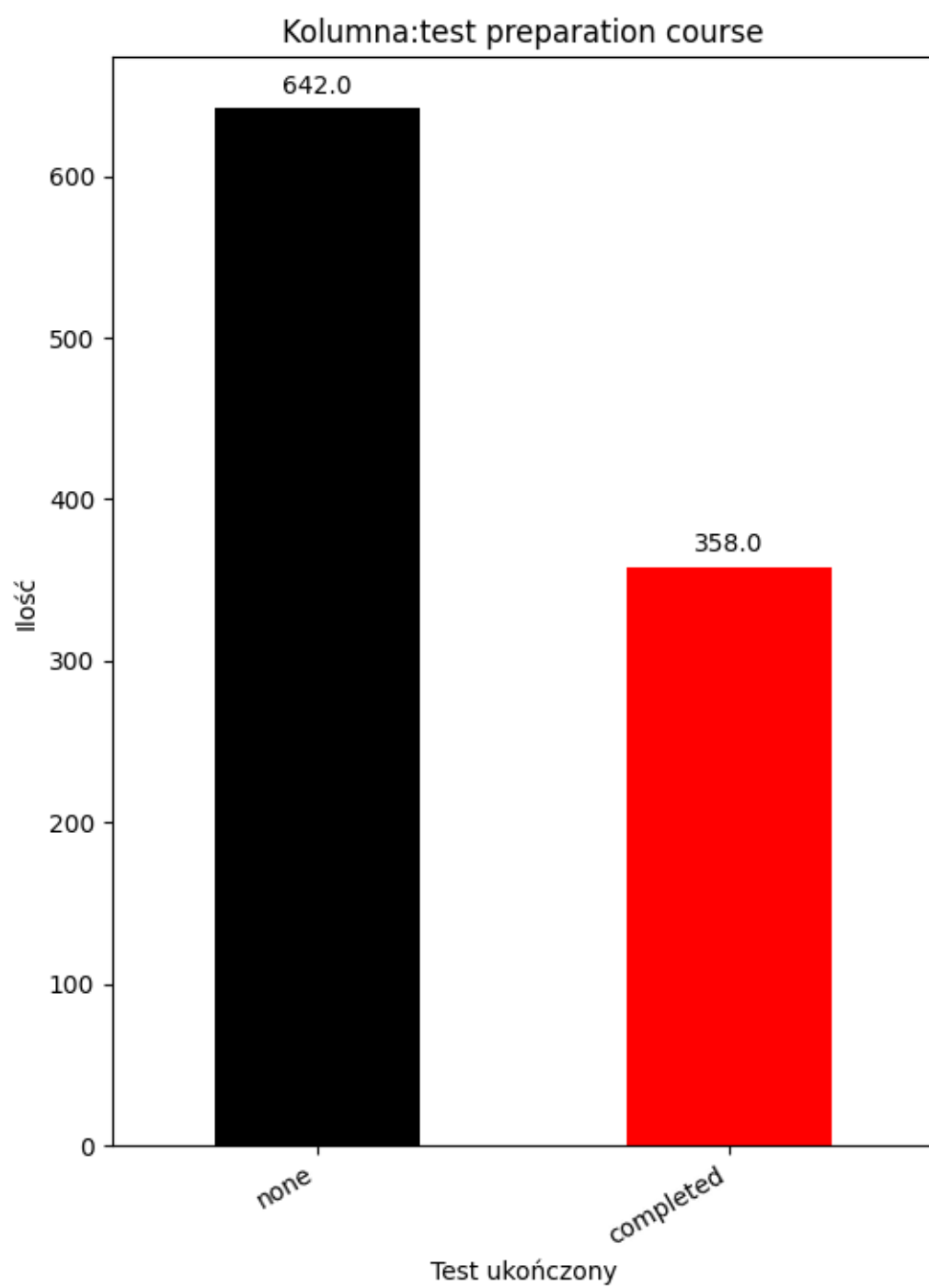
Częstotliwość występowania danych



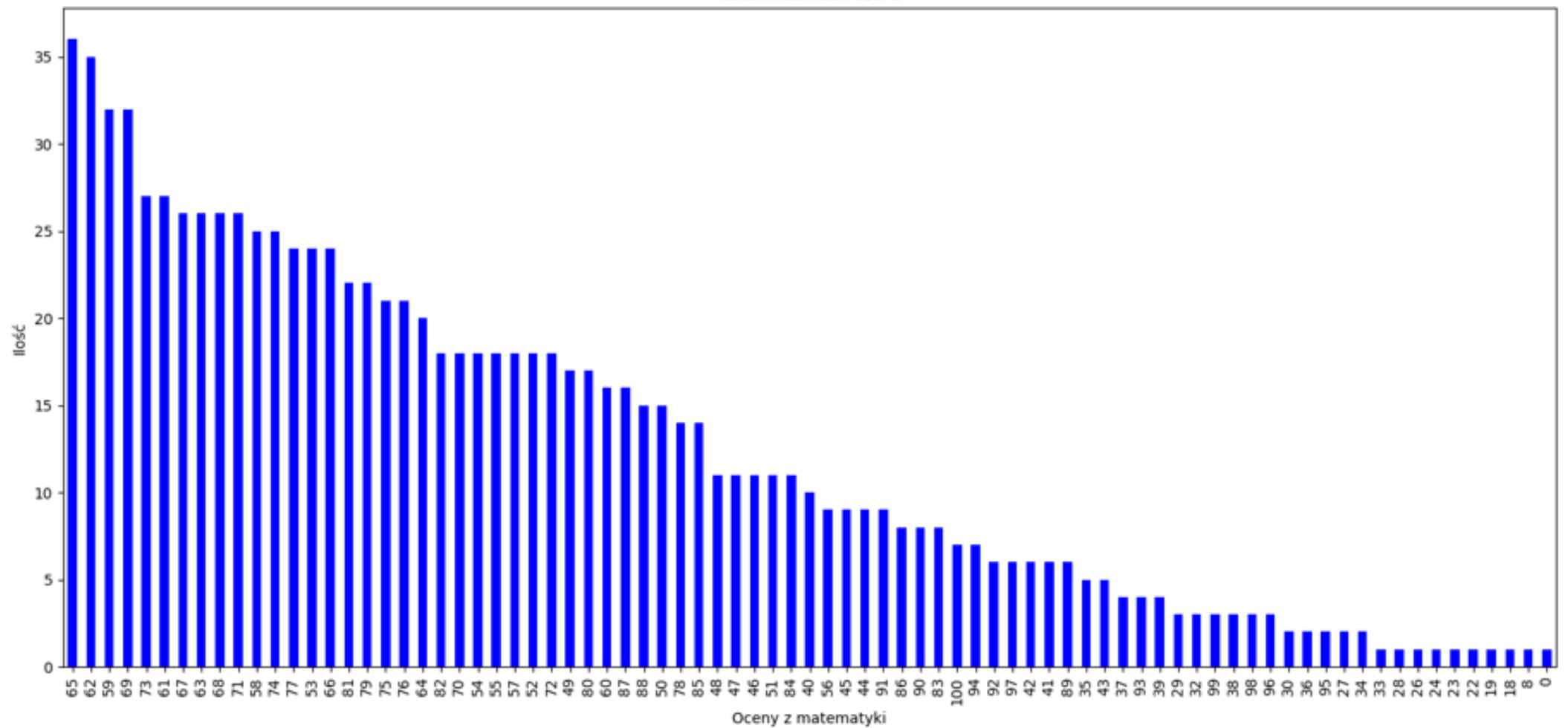


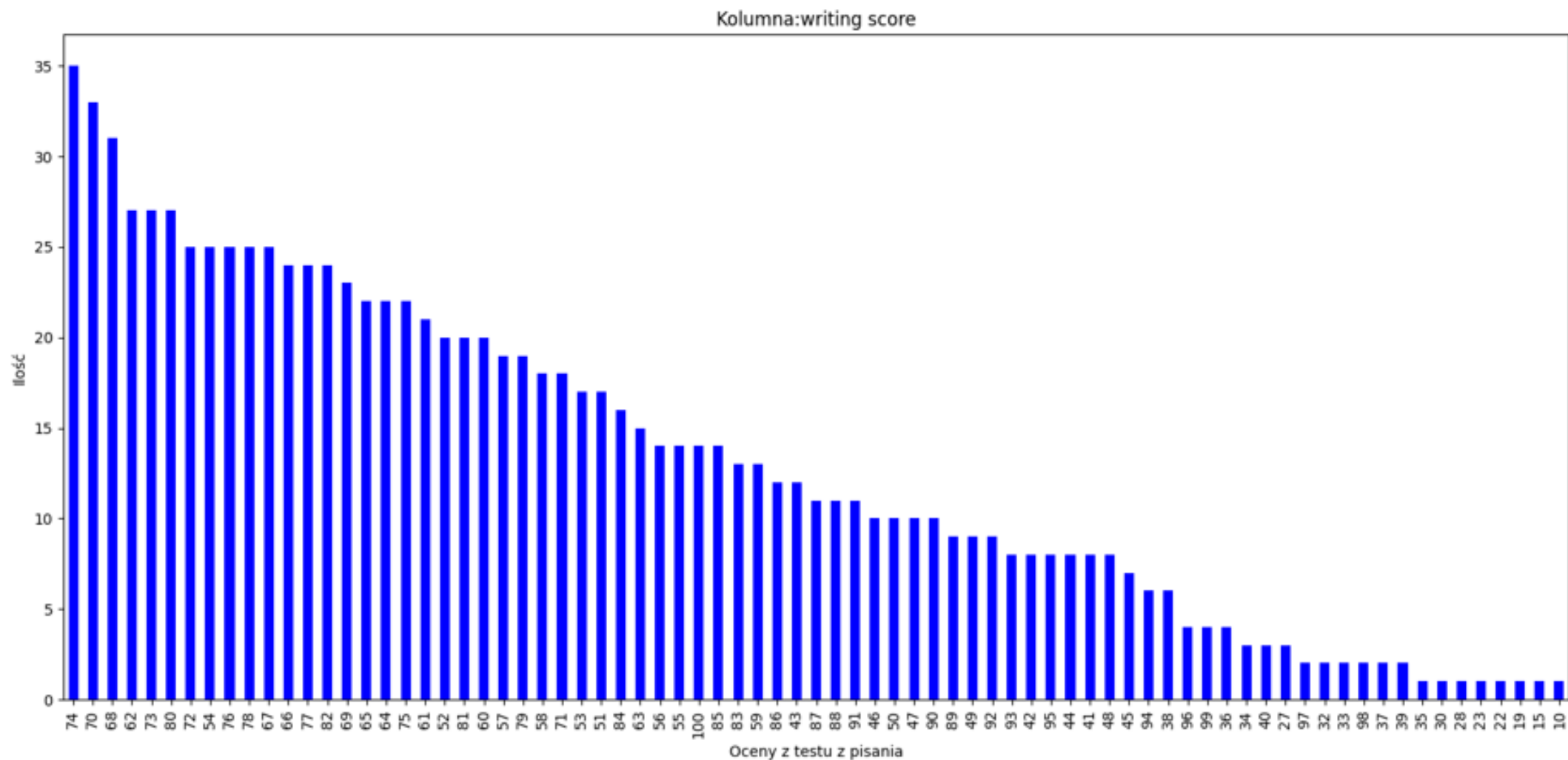




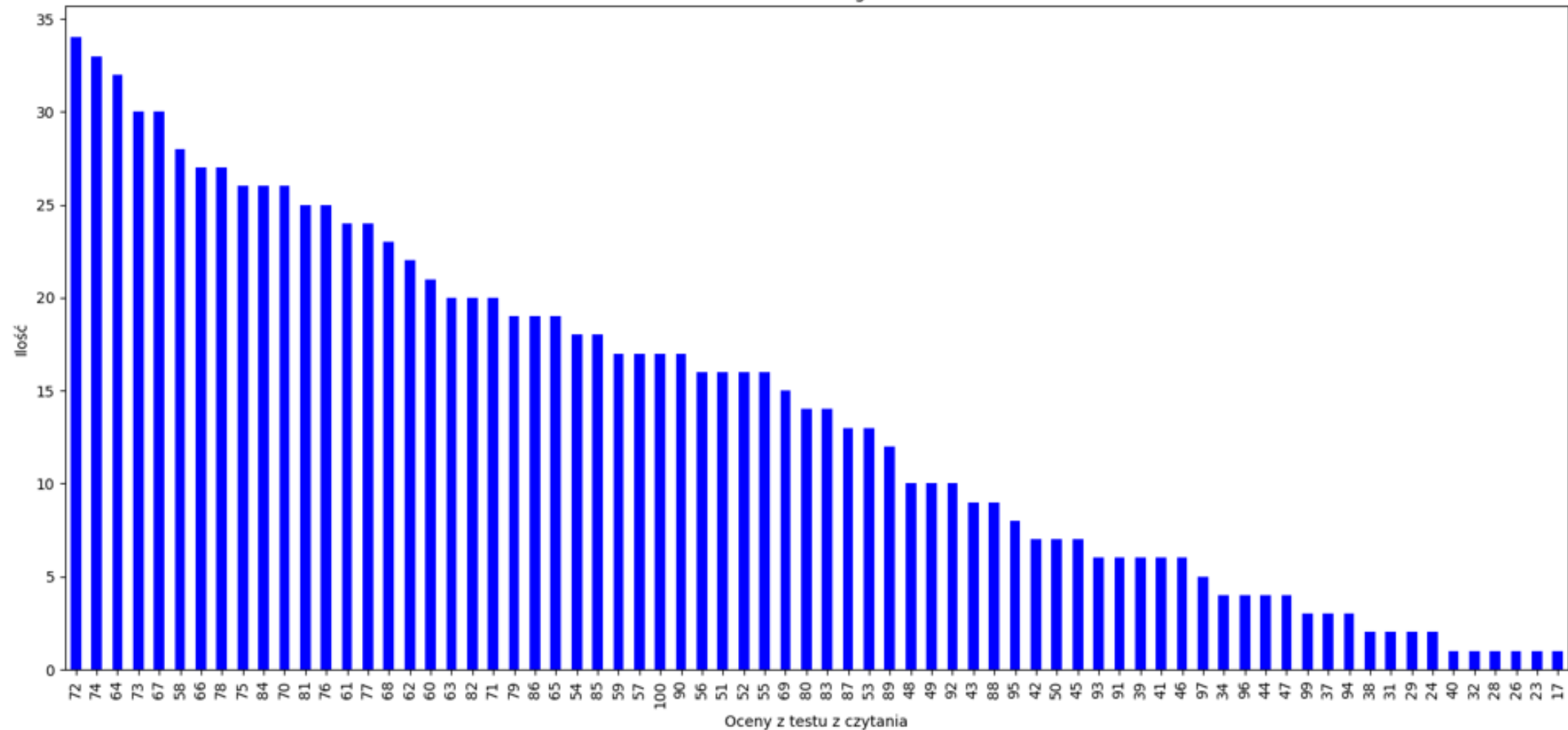


Kolumna:math score





Kolumna:reading score



Porównanie poznanych klasyfikatorów.

Jako klasę wybrano płeć.

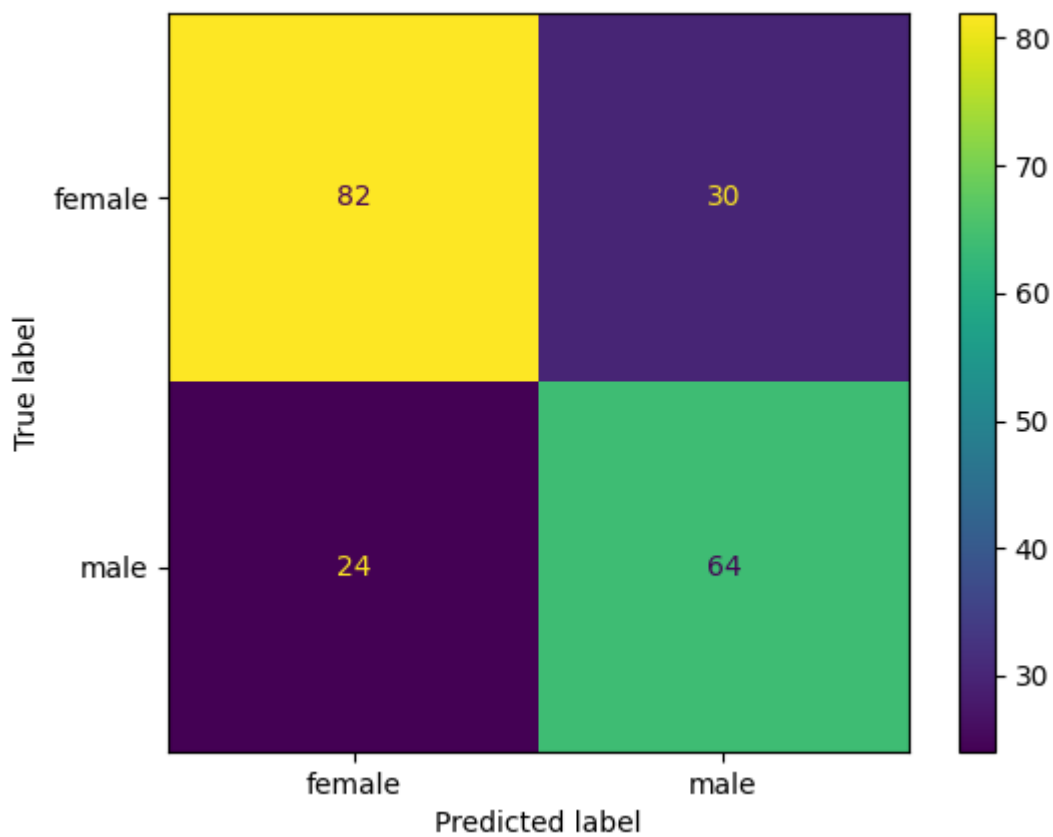
Będziemy próbowali zgadnąć płeć studenta na podstawie danych.

Baza danych posiada tylko 1000 rekordów.

Z tego powodu przeznaczyłem 80% na zbiory treningowe i 20% na testowe.

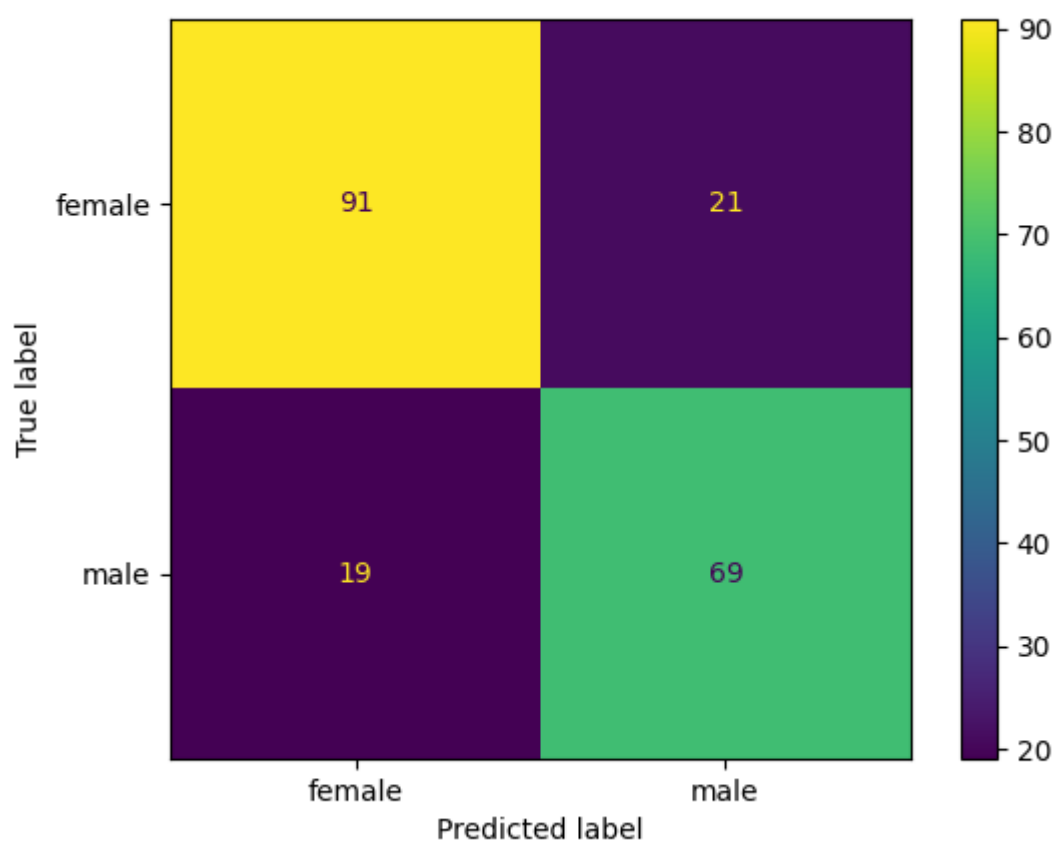
Naive Bayes

Procent poprawnych odpowiedzi: 73%



Drzewa decyzyjne

Procent poprawnych odpowiedzi: 80.5%

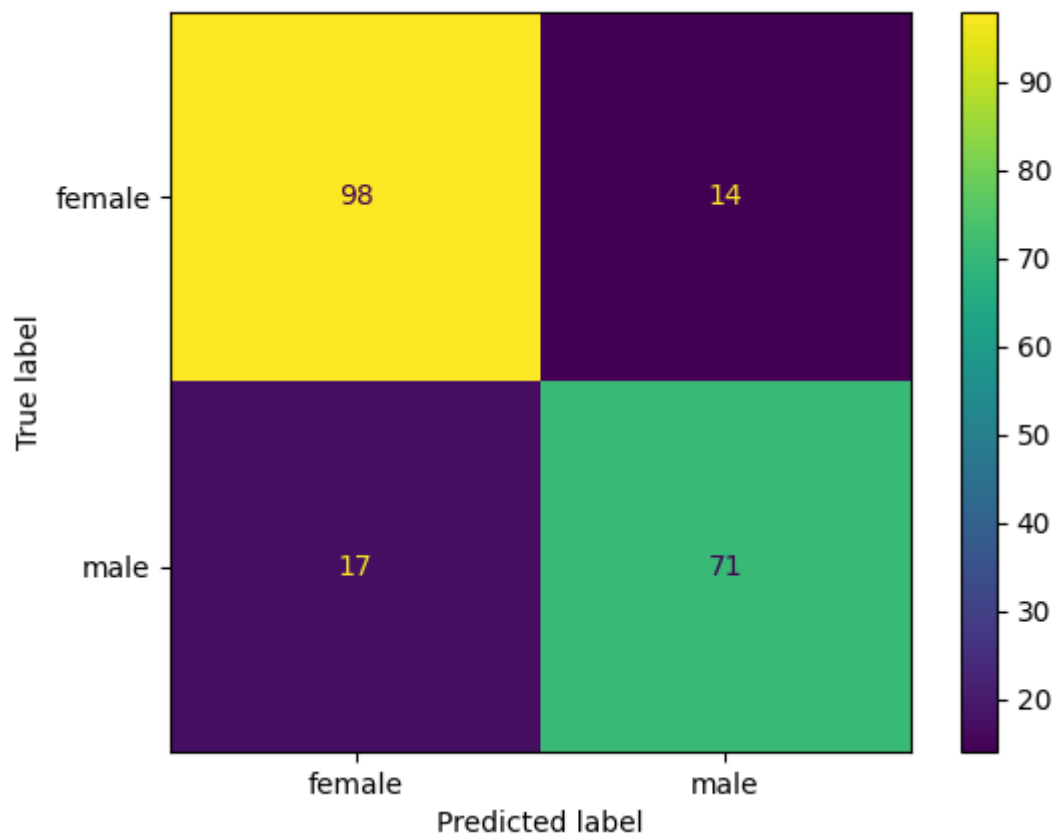


K najbliższych sąsiadów

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

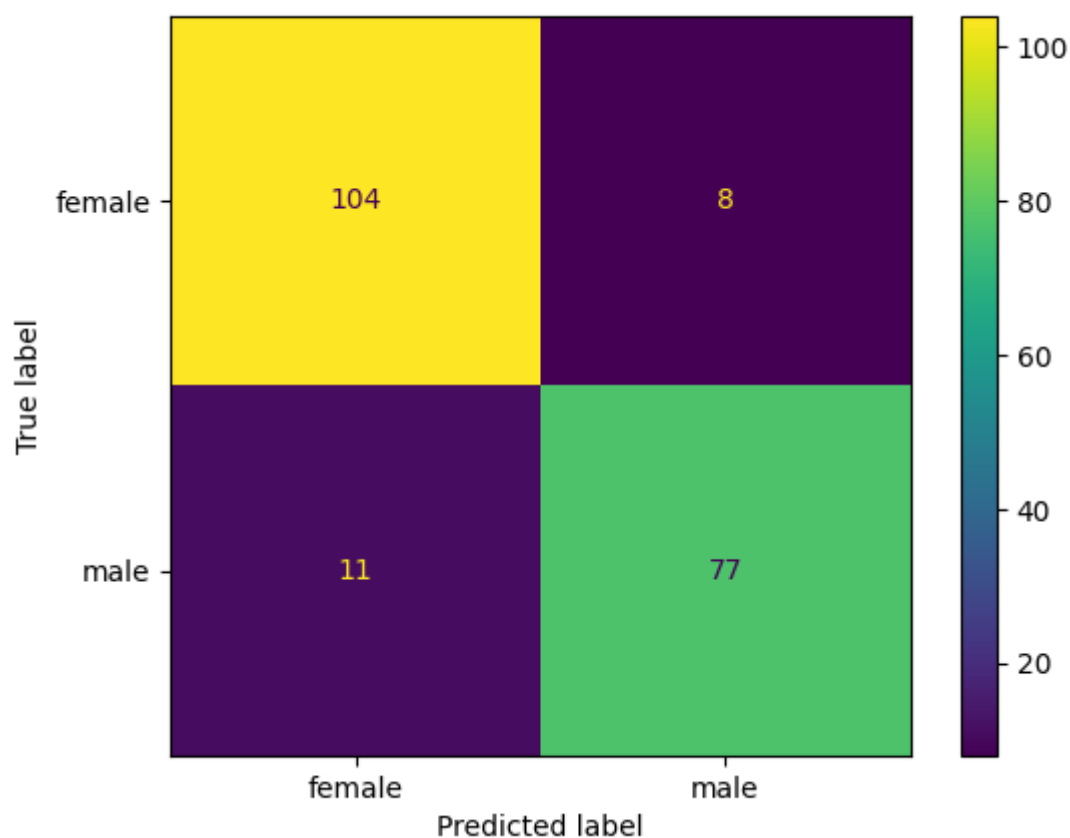
Wniosek: Najlepsze K najbliższych sąsiadów to 70

Macierz błędu dla 70 najbliższych sąsiadów.



Sieci neuronowe

Procent poprawnych odpowiedzi: 90.5%



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.