WUM Projekt 1

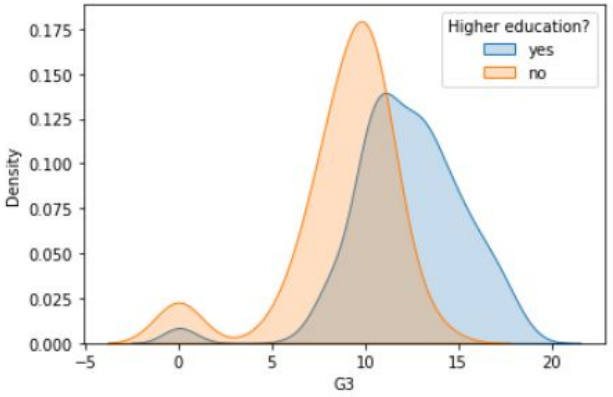
Mikołaj Spytek
Artur Żółkowski

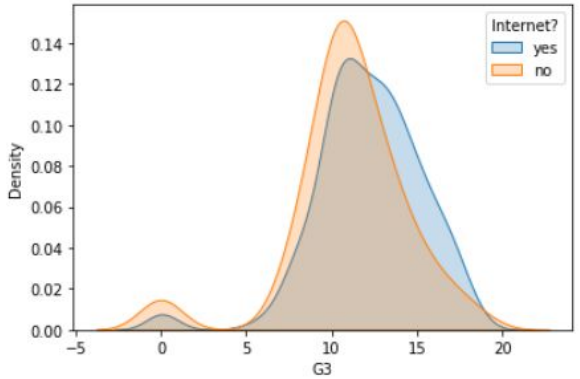
Zbiór danych

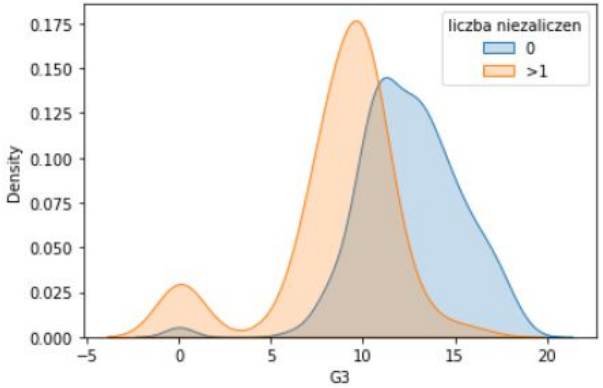
```
RangeIndex: 649 entries, 0 to 648
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      649 non-null    object
 1   sex         649 non-null    object
 2   age         649 non-null    int64
 3   address     649 non-null    object
 4   famsize     649 non-null    object
 5   Pstatus     649 non-null    object
 6   Medu        649 non-null    int64
 7   Fedu        649 non-null    int64
 8   Mjob        649 non-null    object
 9   Fjob        649 non-null    object
 10  reason      649 non-null    object
 11  guardian    649 non-null    object
 12  traveltime  649 non-null    int64
 13  studytime   649 non-null    int64
 14  failures    649 non-null    int64
 15  schoolsup   649 non-null    object
 16  famsup      649 non-null    object
 17  paid        649 non-null    object
 18  activities  649 non-null    object
 19  nursery     649 non-null    object
 20  higher      649 non-null    object
 21  internet    649 non-null    object
 22  romantic    649 non-null    object
 23  famrel      649 non-null    int64
 24  freetime    649 non-null    int64
 25  goout       649 non-null    int64
 26  Dalc        649 non-null    int64
 27  Walc        649 non-null    int64
 28  health      649 non-null    int64
 29  absences    649 non-null    int64
 30  G1          649 non-null    int64
 31  G2          649 non-null    int64
 32  G3          649 non-null    int64
```

kolumny z ramki danych

Rozkłady wybranych zmiennych







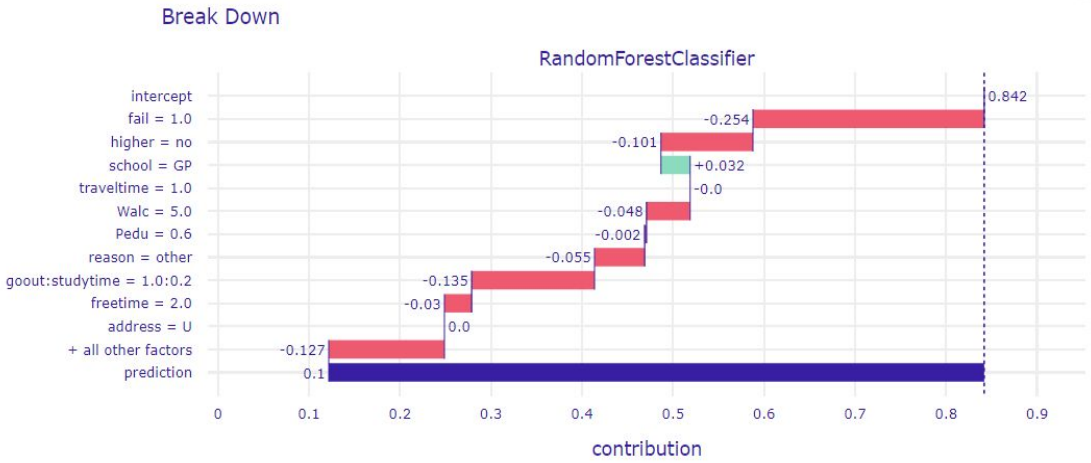
Podejście klasyfikacyjne

```
count    649.000000
mean      11.906009
std        3.230656
min        0.000000
25%       10.000000
50%       12.000000
75%       14.000000
max       19.000000
Name: G3, dtype: float64
```

statystyki opisujące zmienną celu

G3	n
0	15
1	1
5	1
6	3
7	10
8	35
9	35
10	97
11	104
12	72
13	82
14	63
15	49
16	36
17	29
18	15
19	2

liczba wystąpień poszczególnych ocen



Wyjaśnienia rezultatów jednego ze stworzonych modeli

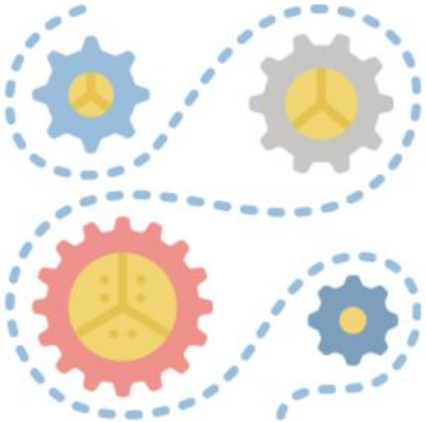
Preprocessing i automatyczne wybieranie zmiennych

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}.$$

mutual information

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

statystyka chi^2, podstawa wyboru jednego z selectorów



selectKBest z selectorem chi2	
failures	school_MS Fjob_health
address_U failures	Pstatus_T failures
Mjob_at_home Fjob_health	Mjob_services failures
Fjob_health reason_home	Fjob_services failures
reason_other failures	guardian_mother failures
internet_yes failures	romantic_yes failures

selectKBest z selectorem mutual information	
failures	school_MS goout
Pstatus_T failures	Fjob_health Walc
higher_yes studytime	higher_yes Dalc
age failures	failures^2
failures Fedu	failures studytime
failures Dalc	failures Walc

recursive feature selection	
failures	absences
school_MS Pstatus_T	school_MS guardian_father
schoolsup_yes nursery_yes	nursery_yes failures
higher_yes Fedu	higher_yes famrel
internet_yes failures	Medu famrel
failures Fedu	Fjob_health reason_home

l1 based feature selection	
age	failures
guardian_mother failures	higher_yes Medu
failures^2	failures Medu
failures traveltime	failures studytime
failures famrel	failures Dalc
failures Walc	traveltime Dalc

ręcznie wybrane zmienne	
fail	reason
higher	school
age	goout
Pedu	Gender Relations
address	internet
Mjob	studytime

Wybór modeli i hiperparametrów

- Logistic Regression
- SVR
- Random Forest
- Gradient Boosting

```
GradientBoostingRegressor(learning_rate=0.045, n_estimators=100, criterion='mse', random_state=0)
RandomForestRegressor(n_estimators=20, max_features=0.5, min_samples_split=3, n_jobs=-1, random_state=0)
LogisticRegression(max_iter=1000)
= SVR(C=1.5)
```

hiperparametry modeli

Wyniki

	Linear Regression	SVR	Random Forest	Gradient Boosting
SelectKBest (chi2)	3.139358	2.884703	2.926594	2.875707
SelectKBest (mutual information)	2.647433	2.655581	2.603689	2.632258
RFE	2.949454	2.758008	2.716849	2.791437
L1 Based Model Selection	2.801049	2.813967	2.905404	2.834317
Hand-prepared features	2.639725	2.680089	2.770634	2.591747

RMSE baseline'u: `3.070836358822292`

Wyjaśnienie modelu