AKADEMIA GÓRNICZO-HUTNICZA
im. Stanisława Staszica w Krakowie

WYDZIAŁ ZARZĄDZANIA

Machine Learning a prognozowanie. Sygnały biologiczne ciała, a palenie wyrobów tytoniowych

Rafał Nojek, Rafał Zając
Rok akademicki 2022/2023
Informatyka i Ekonometria

Spis treści

Informacje o projekcie

Według Światowej Organizacji Zdrowia palenie jest pojedynczą, najbardziej możliwą do zapobieżenia przyczyną wczesnej śmierci. Palenie papierosów powoduje około dwudziestu procent wszystkich zgonów w Stanach Zjednoczonych każdego roku, a także zwiększa szanse na wystąpienie wielu poważnych chorób. Choroby związane z paleniem w Stanach Zjednoczonych kosztują 300 miliardów dolarów każdego roku, wliczając w to bezpośrednią opiekę medyczną i utraconą produktywność z powodu chorób generowanych przez palenie. Ogromne konsekwencje dla zdrowia publicznego wynikające z palenia papierosów ilustrują potrzebę leczenia pomagającego ludziom rzucić palenie.

Wstępnym krokiem w wielu programach rzucania palenia jest charakterystyka wzorców palenia w czasie. Można je określić za pomocą min. oznaczenia we krwi markerów palenia (swoistych i nieswoistych), bądź przeprowadzenia odpowiedniego badania przedmiotowego. Celem naszej pracy jest, na podstawie danych opisujących pacjentów, zbudowanie optymalnego modelu do orzekania, czy dany pacjent jest palaczem czy nie.

Zbiór danych

Analizowane dane dotyczą sygnałów biologicznych naszego ciała, które mogą wskazywać na to, czy dana osoba pali papierosy. Zostały pobrane ze strony:

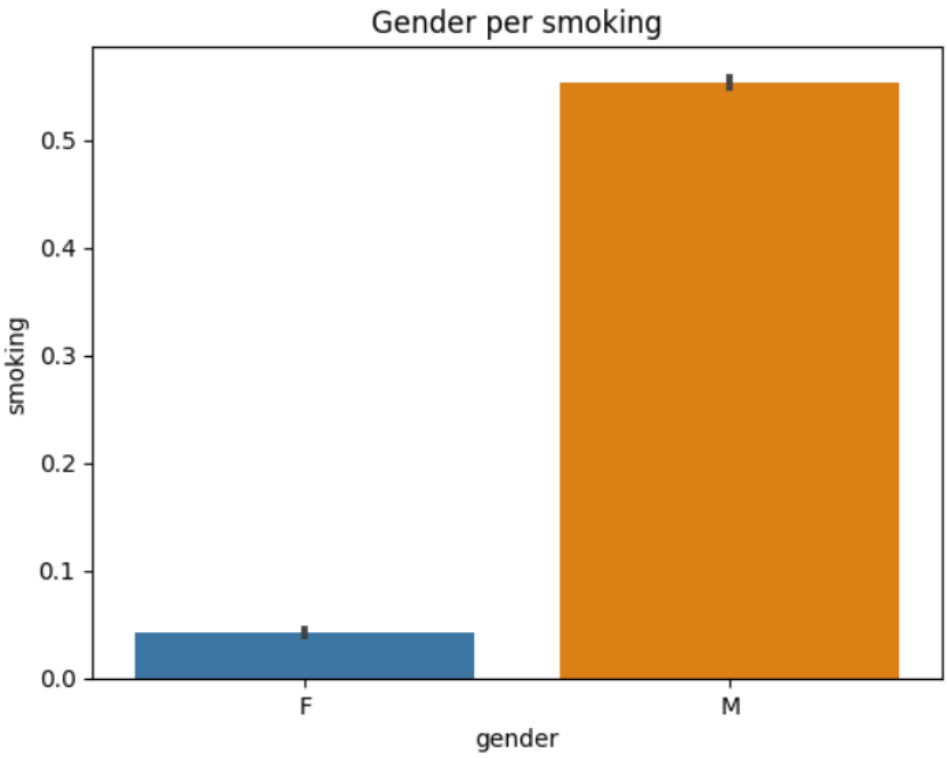
https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking

Każda obserwacja składa sie ze zmiennych:

- **Gender** – płeć
- **Age** – wiek
- **Height(cm)** – wzrost w cm
- **Weight(kg)** – waga w kg
- **Waist(cm)** – szerokość talii w cm
- **Eyesight(left)**– wartość wady oka lewego
- **Eyesight(right)** – wartość wady oka prawego
- **Hearing(left)** – czy osoba słyszy na lewe ucho
- **Hearing(right)** – czy osoba słyszy na prawe ucho
- **Systolic** – skurczowe ciśnienie krwi
- **Relaxation** – ciśnienie krwi podczas relaksu
- **Fasting blood sugar**– poziom cukru we krwi na czczo
- **Cholsterol** – poziom cholesterolu

- **Trygliceride** – poziom trójglicerydów
- **HDL** – poziom cholsterolu HDL (określany “dobrym” cholsterolem)
- **LDL** – poziom cholesterolu LDL (określany “złym” cholsterolem)
- **Hemoglobin** – poziom hemoglobiny
- **Urine protein** - białko w osoczu
- **Serum creatinine** – kreatyna w surowicy
- **AST** – enzym aminotransferaza asparaginianowa
- **ALT** – enzym aminotransferaza alaninowa
- **GTP** – enzym gamma-glutamylotranspeptydaza
- **Dental caries** - próchnica
- **Tartar** - kamień nazębny
- **Smoking** – czy dana osoba pali czy nie (0 – nie pali, 1 - pali)

Zmienną objaśnianą w danym badaniu będzie zmienna Smoking. Pozostałe zmienne są objaśniające. Zmienne “ID” oraz “oral” nie są wykorzystywane w badaniu. Zmienne “Eyesight(left/right)” oraz “Hearing(left/right)”, w późniejszym etapie, połączono w odpowiedni sposób i zamieniono na “Eyesight” oraz “Hearing”.



Rysunek 1 Płeć a palenie

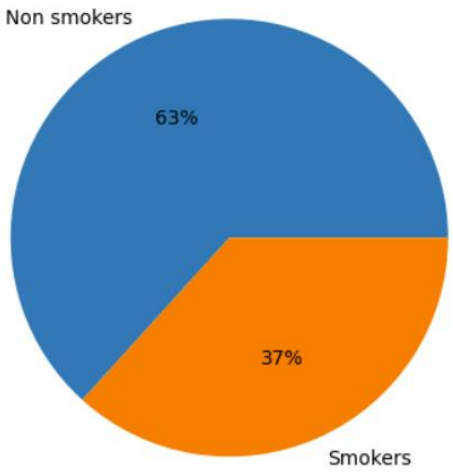
Na przykładzie analizowanych danych widać, że więcej palaczy to mężczyźni.

Statystyki opisowe i charakterystyka zmiennych

Aby przeprowadzić poprawną analizę, należy przygotować odpowiednio dane. W tym celu, obliczone zostały statystki opisowe, współczynnik zmienności oraz korelacje między zmiennymi.

	age	height(cm)	weight(kg)	dental_caries	smoking
count	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000
mean	44.182917	164.649321	65.864936	0.213334	0.367288
std	12.071418	9.194597	12.820306	0.409665	0.482070
min	20.000000	130.000000	30.000000	0.000000	0.000000
25%	40.000000	160.000000	55.000000	0.000000	0.000000
50%	40.000000	165.000000	65.000000	0.000000	0.000000
75%	55.000000	170.000000	75.000000	0.000000	1.000000
max	85.000000	190.000000	135.000000	1.000000	1.000000

Rysunek 2 Statystyki opisowe I



Rysunek 3 Rozkład zmiennej objaśnianej

Średni wiek analizowanych osób to około 44 lata, najmłodsza osoba miała 20 lat, a najstarsza 85. Minimalna waga osoby do 30 kg, co jest bardzo niskim wynikiem, natomiast maksymalna waga to 135 kg. Średnia zmiennej objaśnianej to 0,37. W przypadku tego badania oznacza to, że zbiór danych posiada większą liczbę osób niepalących niż palących. Będzie to mieć wpływ na dobór stosowanych przez nas metryk do oceny zbudowanych modeli. Chcąc osiągnąć cel naszego projektu, w ocenie modeli powinniśmy skupiać się na głównie na ocenie czułości (ang. True Positive Rate) i maksymalizować jej wynik manipulując progiem odcięcia danego modelu, jednakże dla

uproszczenia zdecydowaliśmy się w niego nie ingerować i zbierać znane nam metryki, aby porównać ze sobą modele.

	waist(cm)	eyesight(left)	eyesight(right)	hearing(left)	hearing(right)	systolic	relaxation	fasting_blood_sugar	Cholesterol
count	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000
mean	82.046418	1.012623	1.007443	1.025587	1.026144	121.494218	76.004830	99.312325	196.901422
std	9.274223	0.486873	0.485964	0.157902	0.159564	13.675989	9.679278	20.795591	36.297940
min	51.000000	0.100000	0.100000	1.000000	1.000000	71.000000	40.000000	46.000000	55.000000
25%	76.000000	0.800000	0.800000	1.000000	1.000000	112.000000	70.000000	89.000000	172.000000
50%	82.000000	1.000000	1.000000	1.000000	1.000000	120.000000	76.000000	96.000000	195.000000
75%	88.000000	1.200000	1.200000	1.000000	1.000000	130.000000	82.000000	104.000000	220.000000
max	129.000000	9.900000	9.900000	2.000000	2.000000	240.000000	146.000000	505.000000	445.000000

Rysunek 4 Statystyki opisowe II

Dla zmiennej Cholesterol występuje wysoka wartość odchylenia standardowego. Maximum tej zmiennej równe 445 to krytyczna wartość pomiarowa dla człowieka. Widać duże rozbieżności między wartością minimalną oraz maksymalną dla systolic, relaxation oraz fasting_blood_sugar.

	triglyceride	HDL	LDL	hemoglobin	Urine protein	serum_creatinine	AST	ALT	Gtp
count	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000
mean	126.665697	57.290347	114.964501	14.622592	1.087212	0.885738	26.182935	27.036037	39.952201
std	71.639817	14.738963	40.926476	1.564498	0.404882	0.221524	19.355460	30.947853	50.290539
min	8.000000	4.000000	1.000000	4.900000	1.000000	0.100000	6.000000	1.000000	1.000000
25%	74.000000	47.000000	92.000000	13.600000	1.000000	0.800000	19.000000	15.000000	17.000000
50%	108.000000	55.000000	113.000000	14.800000	1.000000	0.900000	23.000000	21.000000	25.000000
75%	160.000000	66.000000	136.000000	15.800000	1.000000	1.000000	28.000000	31.000000	43.000000
max	999.000000	618.000000	1860.000000	21.100000	6.000000	11.600000	1311.000000	2914.000000	999.000000

Rysunek 5 Statystyki opisowe III

Tak jak na poprzednim obrazku, widać bardzo duże różnice między wartością minimalną oraz maksymalną dla Gtp, ALT, AST, LDL, HDL, triglyceride. U różnych badanych te dane mogą przyjmować znacznie różniące się między sobą wartości. Zmienne trygliceride, Gtp oraz LDL posiadają wysoką wartość odchylenia standardowego.

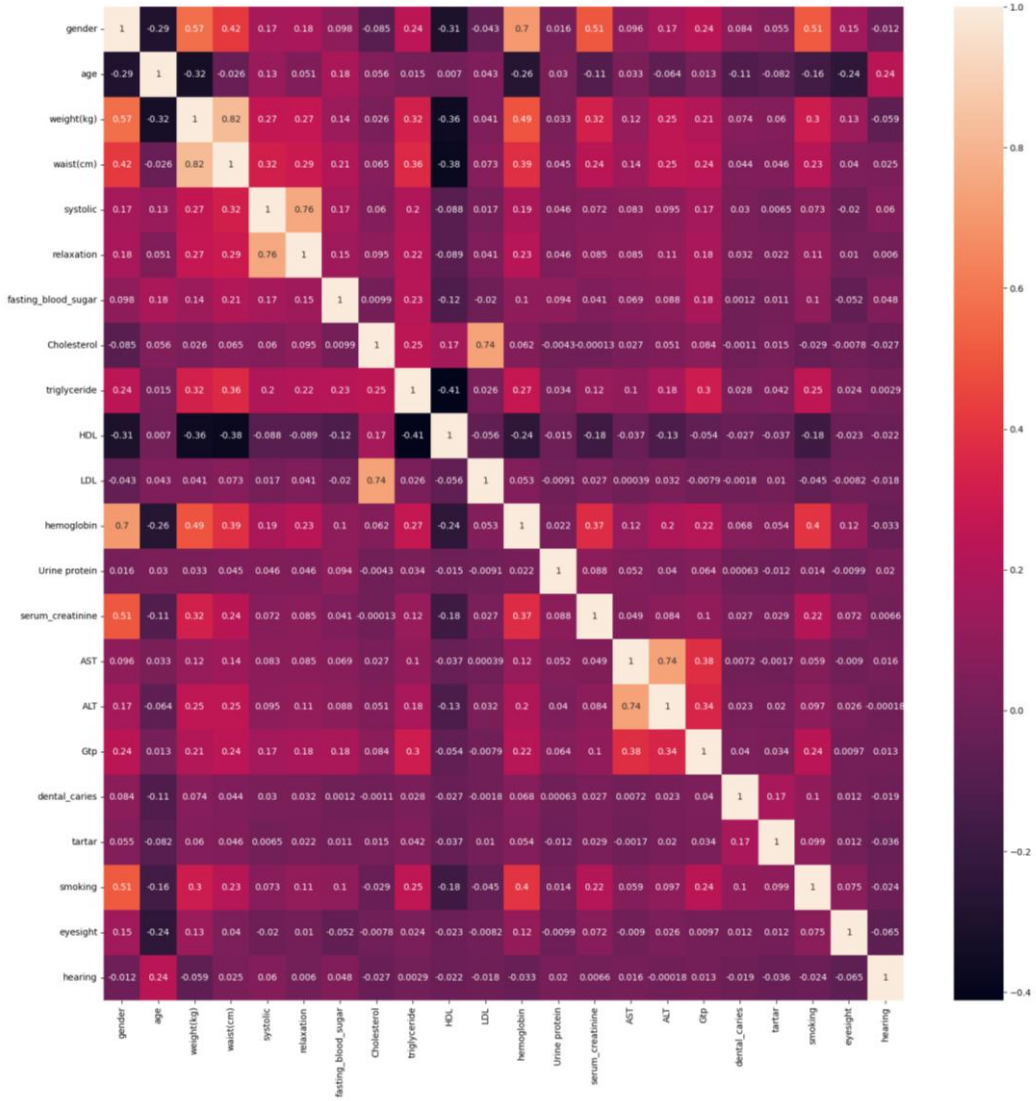
Współczynnik zmienności

age	0.273212
height(cm)	0.055843
weight(kg)	0.194644
waist(cm)	0.113035
systolic	0.112564
relaxation	0.127350
fasting_blood_sugar	0.209394
Cholesterol	0.184344
triglyceride	0.565577
HDL	0.257266
LDL	0.355989
hemoglobin	0.106991
Urine protein	0.372401
serum_creatinine	0.250099
AST	0.739233
ALT	1.144679
Gtp	1.258756
dental_caries	1.920282
tartar	0.894427
smoking	1.312501
eyesight	0.396331
hearing	0.134450

Rysunek 6 Współczynnik zmienności

Wartość współczynnika zmienności jest mniejsza od 0,1 dla zmiennej “hieght(cm)”. Nie będzie ona brana pod uwagę w dalszej analizie.

Korelacje



Rysunek 7 Macierz korelacji

Żadne z korelacji nie osiągnęła krytycznej wartości, wszystkie zmienne poddane są dalszej analizie.

Metoda K najbliższych sąsiadów

Metoda ta wyznacza k najbliższych sąsiadów obiektu (punktów o najmniejszej odległości według zadanej metryki), a następnie wyznacza wynik w oparciu o głos większości tych obiektów. Najważniejszym problemem tej metody jest wybór właściwej wartości k. Wynik zależy od przyjętej definicji odległości (miary podobieństwa). W naszym przypadku zastosowaliśmy klasyczną miarę odległości - euklidesową. Sprzyjał

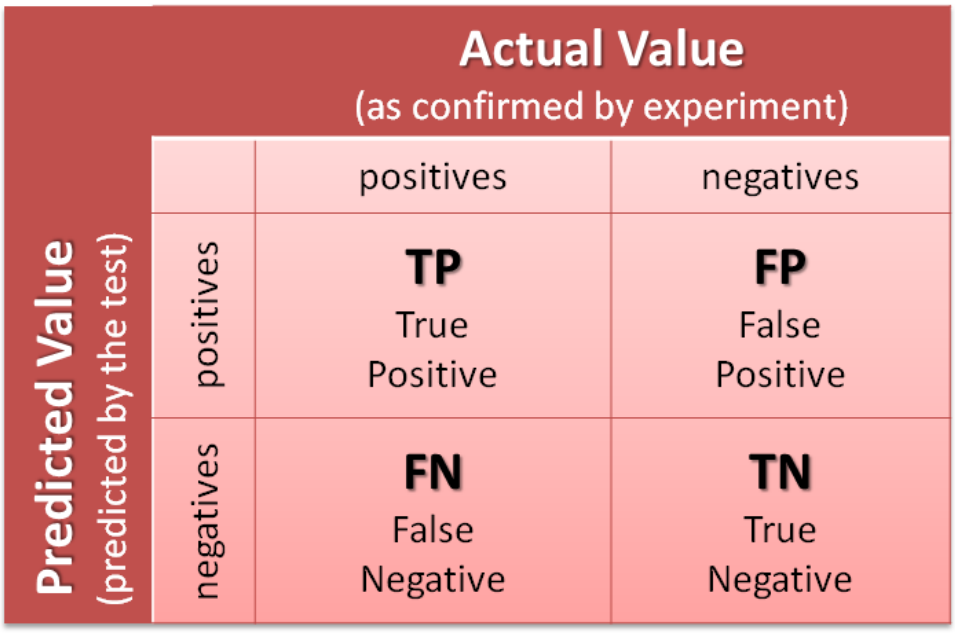
temu fakt, iż jedynie dwie zmienne spośród naszych miały charakter ilościowy (płeć oraz obecność kamienia nazębnego), które miały po dwie kategorie (M/K, obecność/brak), stąd nie było potrzeby ich przetwarzania.

Pierwszym krokiem, aby przygotować dane było podzielenie ich na zbiór uczący i testowy. Zbiór danych podzielono w proporcjach 80% - zbiór uczący, 20% - zbiór testowy.

Dla metody KNN wykonano 4 modele:

- Model I – model surowy
- Model II – ze skalowaniem (standaryzacja)
- Model III – z normalizacją
- Model IV – z większą liczbą k

Po przeprowadzeniu klasyfikacji na zbiorze uczącym, a następnie na zbiorze testowym, można wygenerować dla każdego modelu macierz błędu. Mówi ona o tym, jak prognoza na podstawie modelu jest skuteczna.



Rysunek 8 Macierz błędu

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

Precyzja - ile wśród przykładów zaprognozowanych pozytywnie jest rzeczywiście pozytywnych. Im większa wartość tym lepiej.

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

Czułość - to miara zasięgu/pokrycia, która bada, w jakiej części klasa pozytywna została pokryta przewidywaniem pozytywnym. Im większa wartosć tym lepiej.

$$\textbf{F - measure} = \frac{2*Recall*Precision}{Recall + Precision}$$

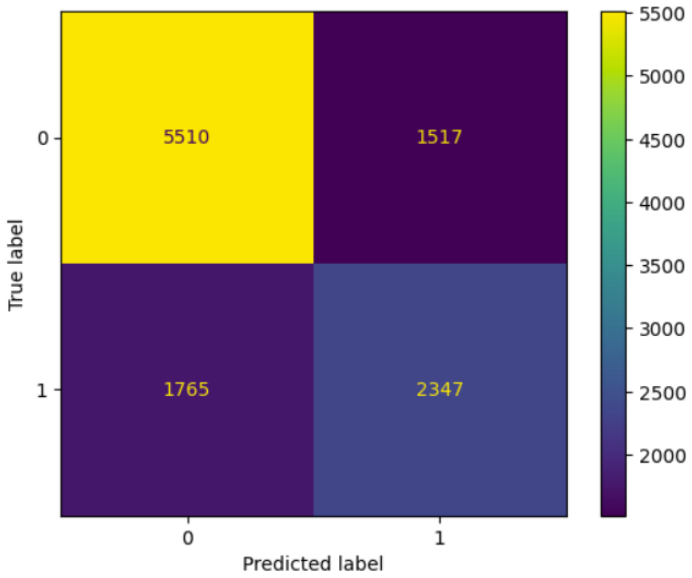
F1-score - F1-score to średnia harmoniczna pomiędzy precyzją (precision) i czułością (recall). Im bliższa jest jedynki, tym lepiej to świadczy o algorytmie klasyfikującym. W najlepszym przypadku przyjmuje wartość 1, kiedy mamy do czynienia z idealną czułością i precyzją.

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}$$

Dokładność - mówi nam o tym, jaka część zbioru testowego została prawidłowo przypisana.

AUC zapewnia zbiorczy pomiar skuteczności we wszystkich możliwych progach klasyfikacji.

Model I – model surowy

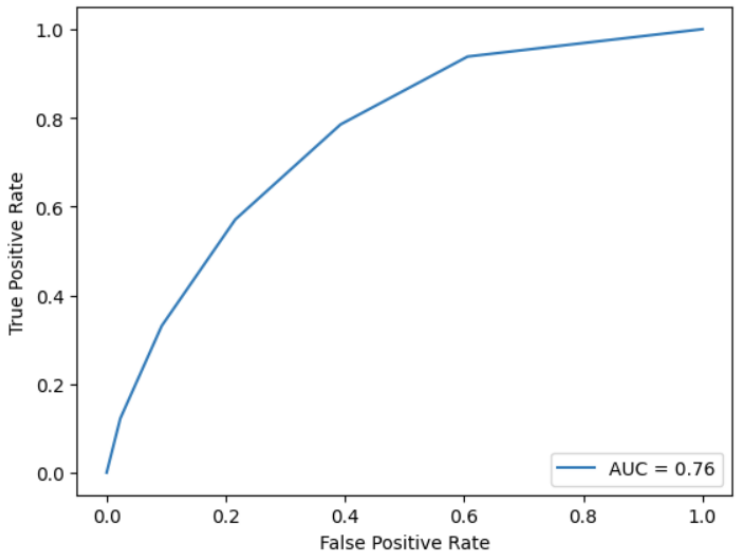


Rysunek 9 Macierz błędów modelu I (KNN)

Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0.607402	0.570768	0.588516	0.70536	0.75832	KNN_raw

Rysunek 10 Statystyki modelu I (KNN)

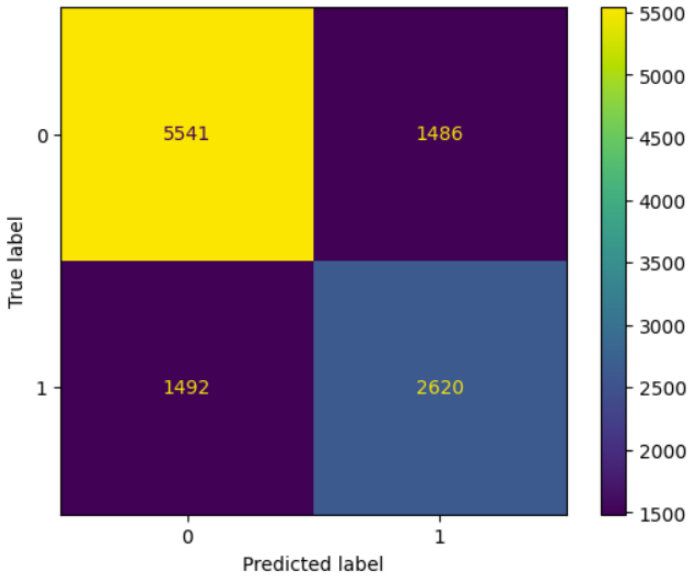
Powyżej zaprezentowane są metryki dla modelu “surowego”.



Rysunek 11 Krzywa ROC modelu I (KNN)

Wartość AUC dla krzywej ROC wynosi 0,76. Im większa wartość tym lepiej.

Model II – ze skalowaniem (standaryzacja)

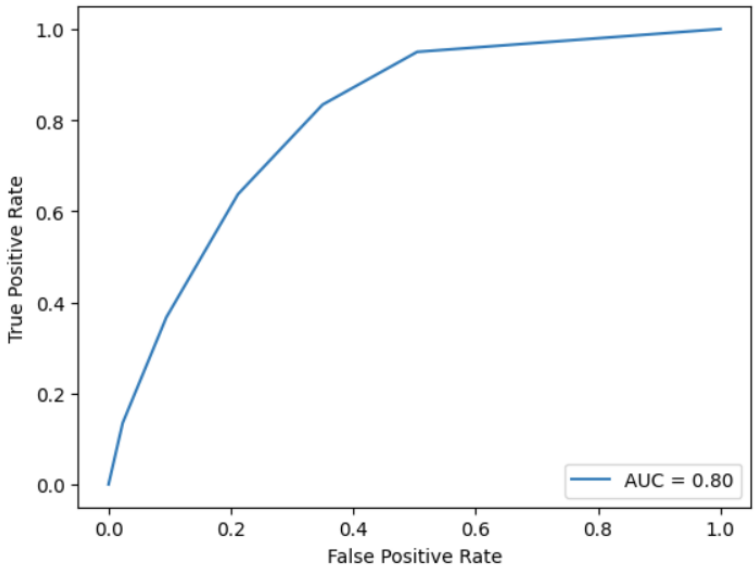


Rysunek 12 Macierz błędów modelu II (KNN)

Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0.638091	0.63716	0.637625	0.732651	0.801018	KNN+standarization

Rysunek 13 Statystyki modelu II (KNN)

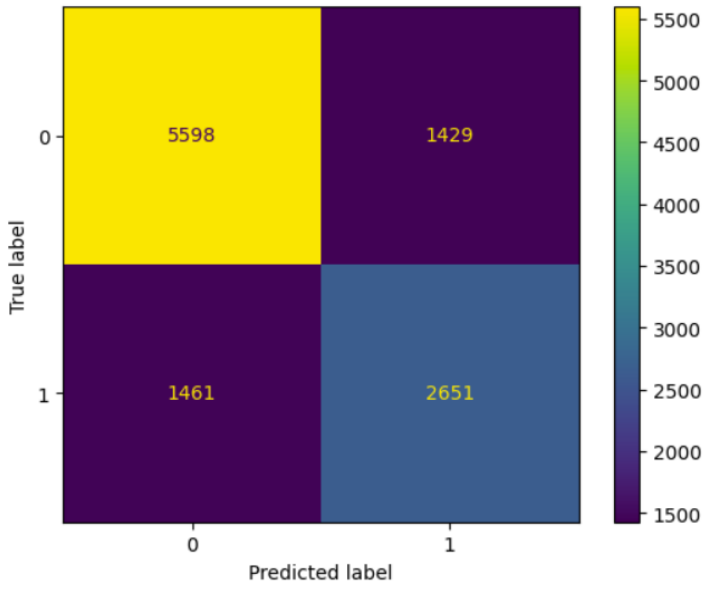
Powyżej zaprezentowane są metryki dla modelu ze skalowaniem.



Rysunek 14 Krzywa ROC modelu II (KNN)

Wartość AUC dla krzywej ROC wynosi 0,8.

Model III – z normalizacją

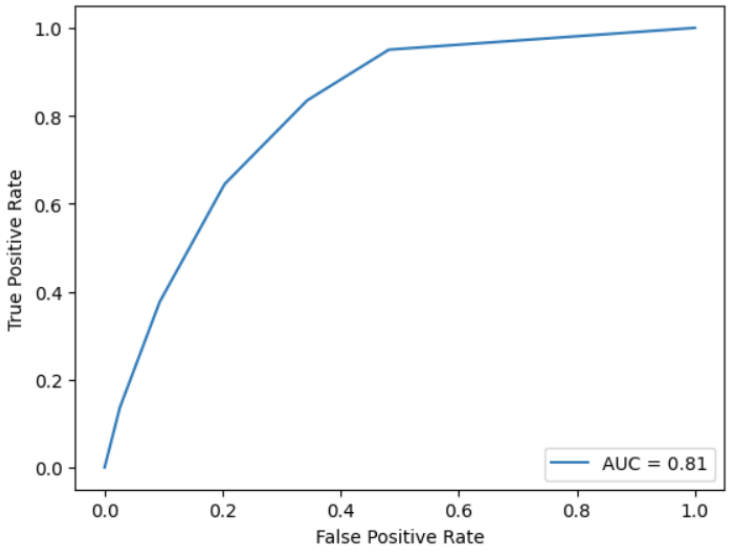


Rysunek 15 Macierz błędów modelu III (KNN)

Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0.649755	0.644698	0.647217	0.740551	0.807807	KNN+minmaxscaling

Rysunek 16 Statystyki modelu III (KNN)

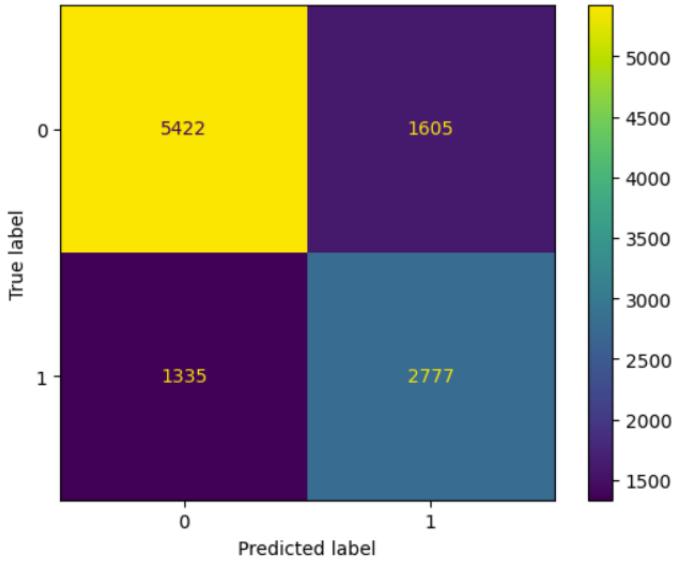
Powyżej zaprezentowane są metryki dla modelu ze skalowaniem.



Rysunek 17 Krzywa ROC modelu III (KNN)

Wartość AUC dla krzywej ROC wynosi 0,81.

Model IV – z większą liczbą k

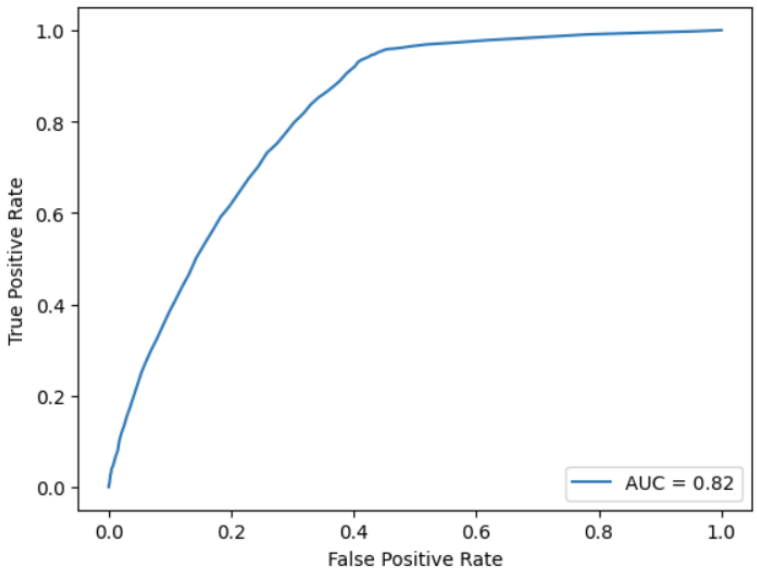


Rysunek 18 Macierz błędów modelu IV (KNN)

Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0.633729	0.67534	0.653873	0.736062	0.818528	KNN_many_neighbours

Rysunek 19 Statystyki modelu IV (KNN)

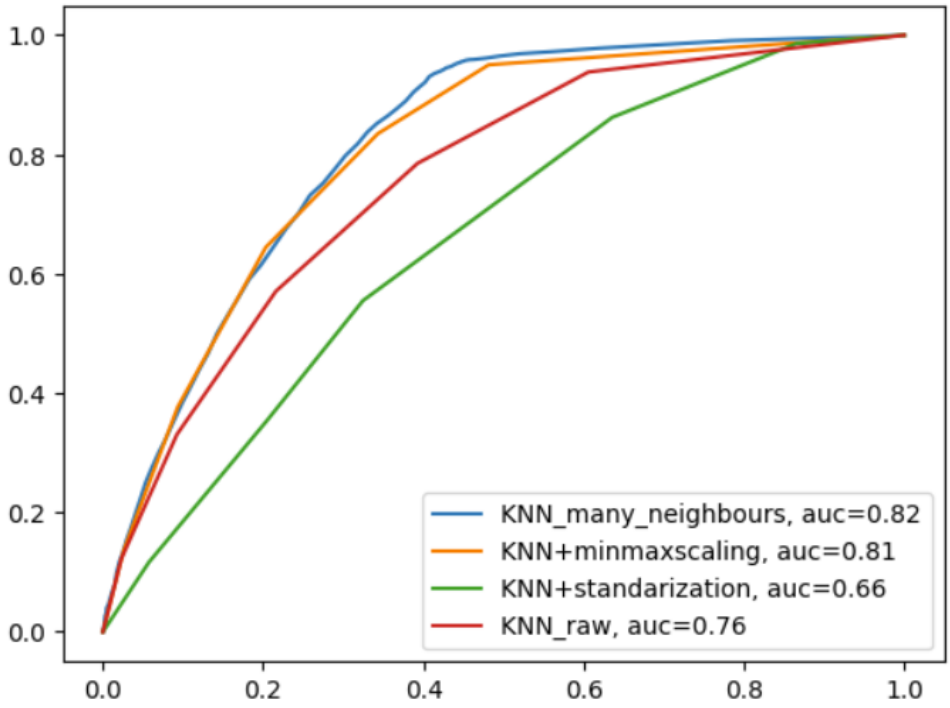
Powyżej zaprezentowane są metryki dla modelu ze skalowaniem.



Rysunek 20 Krzywa ROC modelu IV (KNN)

Wartość AUC dla krzywej ROC wynosi 0,82.

Porównanie modeli



Rysunek 21 Porównanie krzywych ROC (KNN)

Wykres przedstawia krzywe ROC dla 4 modeli. Różnice w wartości AUC dla modelu ze standaryzacją wynikają z innego sposobu obliczania.

Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0.607402	0.570768	0.588516	0.705360	0.758320	KNN_raw
0.638091	0.637160	0.637625	0.732651	0.801018	KNN+standarization
0.649755	0.644698	0.647217	0.740551	0.807807	KNN+minmaxscaling
0.633729	0.675340	0.653873	0.736062	0.818528	KNN_many_neighbours

Rysunek 22 Porównanie statystyk modeli (KNN)

Sugerując się wartością czułości dla danych modeli, jako najlepszy, wybrany został model IV z większą liczbą k.

Regresja logistyczna

Jak mówi sama nazwa, regresja logistyczna jest techniką regresyjną co oznacza, że jest ona zestawem narzędzi statystycznych służących do oszacowania zależności między zmiennymi. W regresji logistycznej, na podstawie zestawu cech ilościowych i jakościowych chcemy przewidzieć wartość zmiennej jakościowej.

Dla regresji logistycznej wykonano 6 modeli

- Model I – model ze wszystkimi zmiennymi
- Model II – model z usuniętymi zmiennymi nieistotnymi
- Model III – model ze zestandaryzowanymi zmiennymi
- Model IV – wybranie zmiennych z największą (najbardziej odległą od 0) korelacją
- Model V – regresja logistyczna z regularyzacją
- Model VI – model z CV

Model I

Logit Regression Results

Dep. Variable:	smoking	No. Observations:	44553
Model:	Logit	Df Residuals:	44530
Method:	MLE	Df Model:	22
Date:	Sun, 20 Nov 2022	Pseudo R-squ.:	0.2839
Time:	17:28:49	Log-Likelihood:	-20969.
converged:	True	LL-Null:	-29282.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-6.4133	0.480	-13.348	0.000	-7.355	-5.472
gender	2.9514	0.058	51.176	0.000	2.838	3.064
age	-0.0002	0.001	-0.129	0.897	-0.003	0.002
height(cm)	0.0206	0.002	8.302	0.000	0.016	0.025
weight(kg)	-0.0093	0.002	-3.876	0.000	-0.014	-0.005
waist(cm)	-0.0023	0.003	-0.817	0.414	-0.008	0.003
systolic	-0.0148	0.001	-10.491	0.000	-0.018	-0.012
relaxation	0.0096	0.002	4.976	0.000	0.006	0.013
fasting blood sugar	0.0032	0.001	5.251	0.000	0.002	0.004
Cholesterol	-0.0023	0.001	-3.927	0.000	-0.003	-0.001
...						
tartar	0.3334	0.024	13.647	0.000	0.285	0.381
eyesight	-0.0305	0.030	-1.012	0.312	-0.089	0.029
hearing	-0.2225	0.092	-2.416	0.016	-0.403	-0.042

Rysunek 23 Regresja logistyczna model I

Niektóre ze zmiennych w modelu są nieistotne. Należy je wykluczyć z dalszej analizy.

Model II

```
Optimization terminated successfully.
         Current function value: 0.470748
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                smoking   No. Observations:                44553
Model:                          Logit   Df Residuals:                    44537
Method:                           MLE   Df Model:                           15
Date:                Sun, 20 Nov 2022   Pseudo R-squ.:                  0.2838
Time:                        22:58:11   Log-Likelihood:                -20973.
converged:                       True   LL-Null:                       -29282.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -6.6140      0.407    -16.251      0.000      -7.412      -5.816
gender         2.9307      0.057     51.472      0.000       2.819       3.042
height(cm)     0.0220      0.002      9.751      0.000       0.018       0.026
weight(kg)    -0.0115      0.001     -8.245      0.000      -0.014      -0.009
systolic      -0.0147      0.001    -10.521      0.000      -0.017      -0.012
relaxation     0.0096      0.002      4.969      0.000       0.006       0.013
fasting blood sugar     0.0032      0.001      5.336      0.000       0.002       0.004
Cholesterol   -0.0023      0.000     -6.538      0.000      -0.003      -0.002
triglyceride   0.0046      0.000     24.239      0.000       0.004       0.005
hemoglobin     0.1445      0.012     12.025      0.000       0.121       0.168
serum creatinine      -0.9087      0.075    -12.102      0.000      -1.056      -0.761
ALT           -0.0062      0.001     -9.678      0.000      -0.007      -0.005
Gtp            0.0078      0.000     21.314      0.000       0.007       0.008
dental caries  0.3389      0.029     11.783      0.000       0.283       0.395
tartar         0.3329      0.024     13.651      0.000       0.285       0.381
hearing       -0.2355      0.090     -2.611      0.009      -0.412      -0.059
==============================================================================
```

Rysunek 24 Regresja logistyczna model II

Niska wartość pseudo R2 oznacza, że model wyjaśnia niewielką część wariancji.

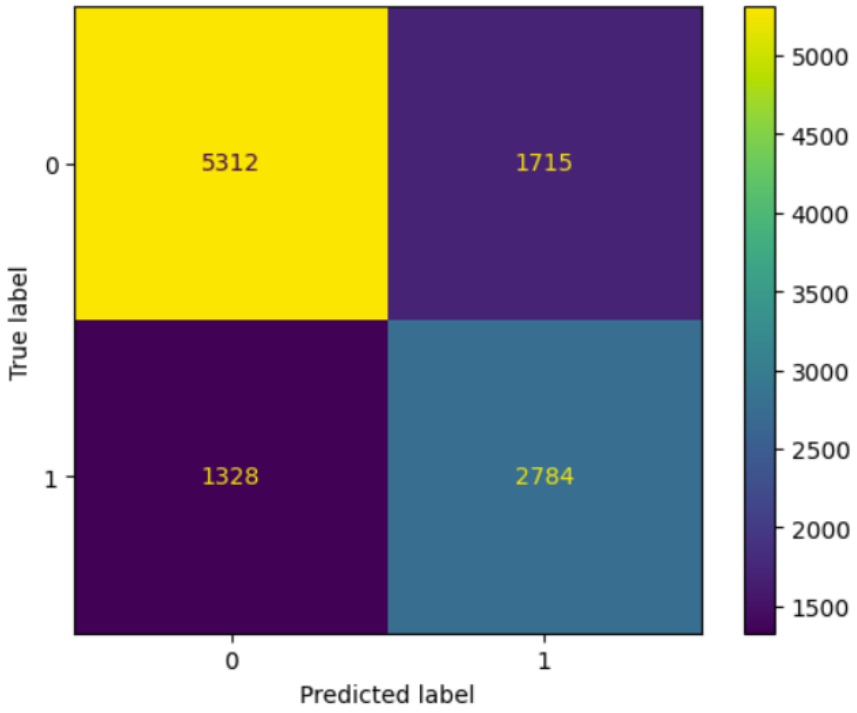
Model III

```
==============================================================================
Dep. Variable:                smoking   No. Observations:                44553
Model:                          Logit   Df Residuals:                    44534
Method:                           MLE   Df Model:                           18
Date:                Sun, 20 Nov 2022   Pseudo R-squ.:                  0.2838
Time:                        17:28:51   Log-Likelihood:                -20972.
converged:                       True   LL-Null:                       -29282.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.1265      0.145    -28.391      0.000      -4.411      -3.842
gender         2.9408      0.057     51.198      0.000       2.828       3.053
age           -0.0317      0.079     -0.403      0.687      -0.186       0.122
height(cm)     1.1493      0.136      8.428      0.000       0.882       1.417
weight(kg)    -1.0068      0.253     -3.984      0.000      -1.502      -0.511
waist(cm)     -0.2163      0.223     -0.971      0.332      -0.653       0.221
systolic      -2.4771      0.238    -10.399      0.000      -2.944      -2.010
relaxation     1.0197      0.205      4.978      0.000       0.618       1.421
fasting blood sugar     1.2398      0.232      5.349      0.000       0.786       1.694
Cholesterol   -0.8970      0.138     -6.510      0.000      -1.167      -0.627
...
tartar         0.3326      0.024     13.621      0.000       0.285       0.380
eyesight      -0.2925      0.295     -0.992      0.321      -0.870       0.286
hearing       -0.2233      0.092     -2.425      0.015      -0.404      -0.043
==============================================================================
```

Rysunek 25 Regresja logistyczna model III

Z przedstawionych danych wynika, że standaryzacja nie ma wpływu na jakość modelu (równy wynik pseudo R2).

Model IV

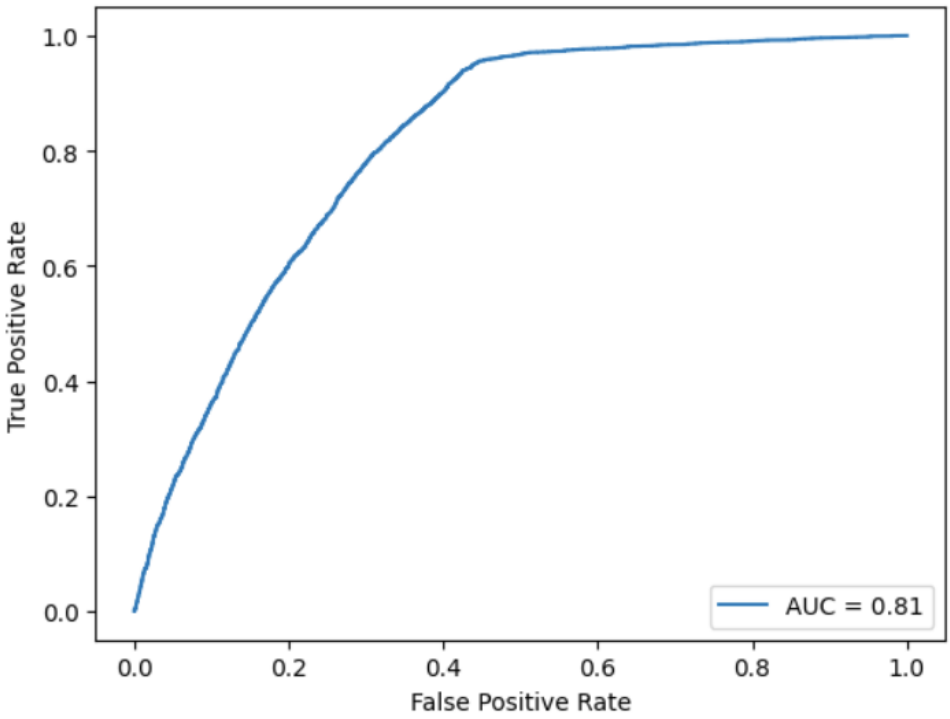


Rysunek 26 Macierz błędów model IV (regresja logistyczna)

Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0.618804	0.677043	0.646615	0.726816	0.810852	Logit 4 variables

Rysunek 27 Statystyki modelu IV (regresja logistyczna)

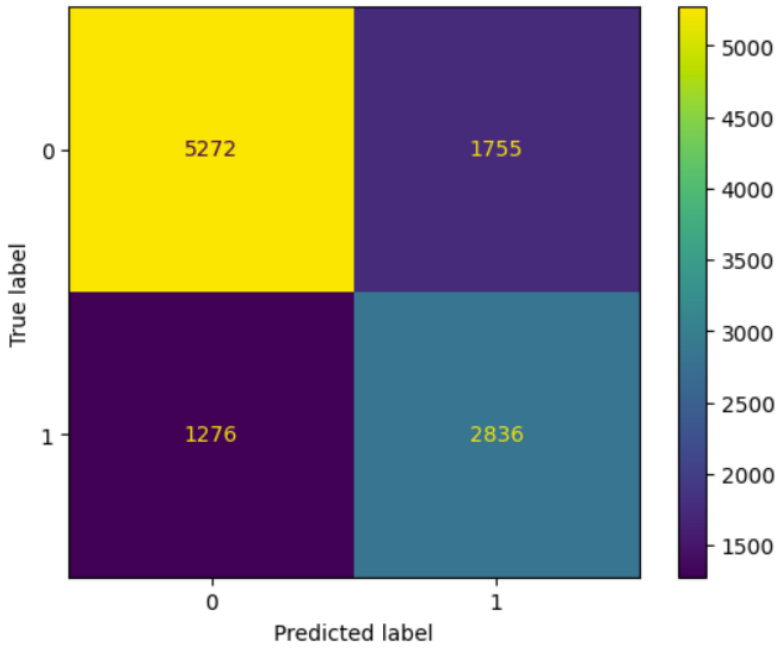
Dokładność dla modelu wynosi około 73%. Czułość oraz precyzja osiągają wartość powyżej 60%.



Rysunek 28 Krzywa ROC dla modelu IV (regresja logistyczna)

AUC dla modelu wynosi 0,81. Jest to zadowalający wynik.

Model V

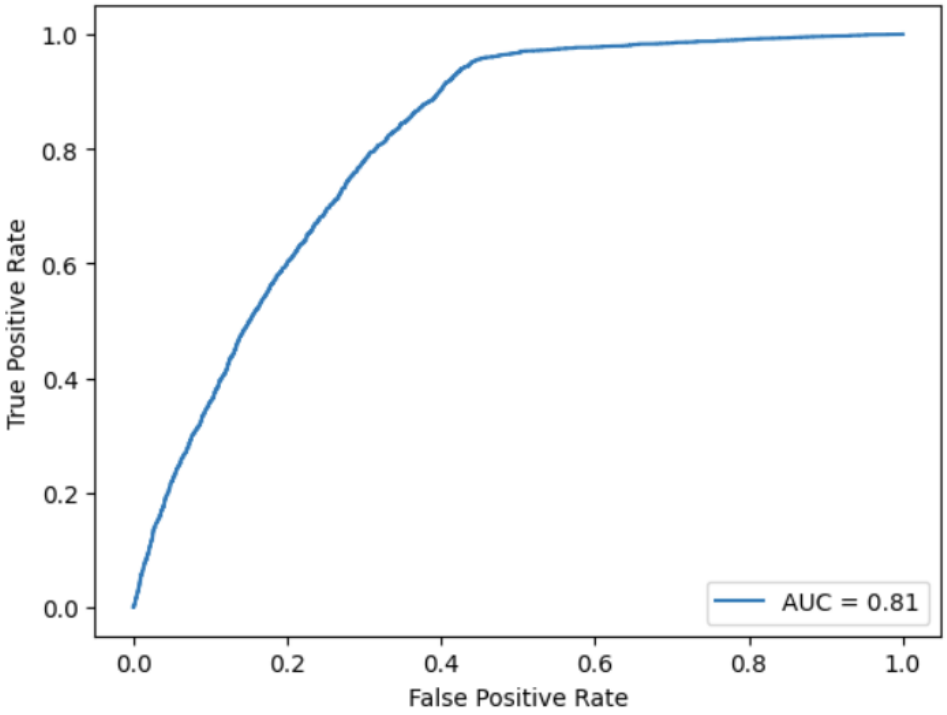


Rysunek 29 Macierz błędów model V (regresja logistyczna)

Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0.61773	0.689689	0.651729	0.727893	0.81087	Logit_4_vars+reg

Rysunek 30 Statystyki modelu V (regresja logistyczna)

Wyniki nieznacznie lepsze niż w modelu poprzednim.



Rysunek 31 Krzywa ROC modelu V (regresja logistyczna)

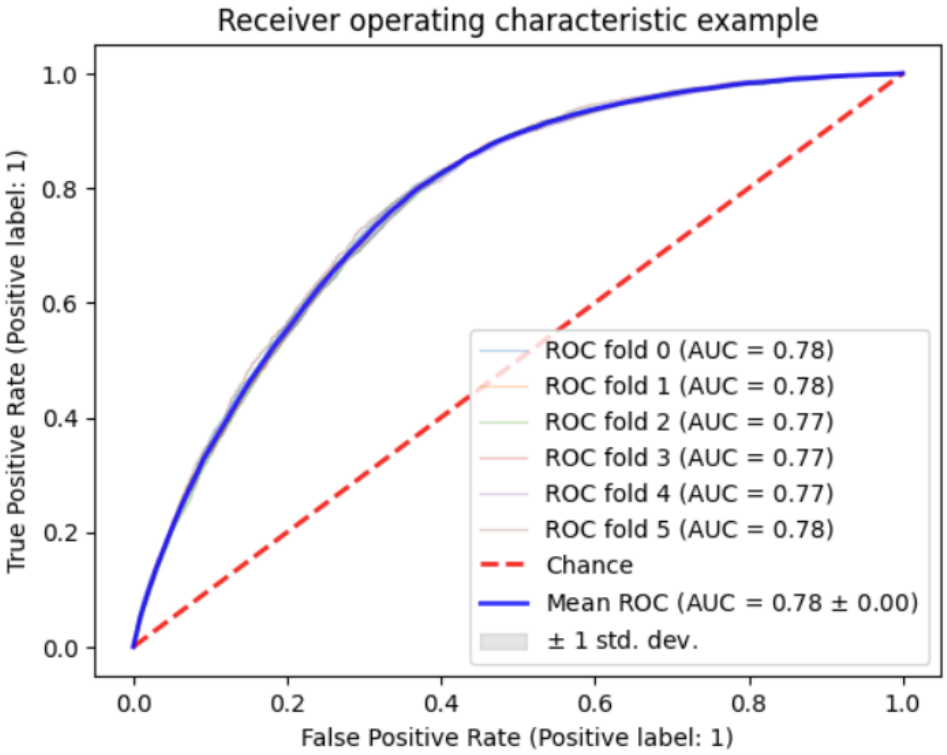
AUC dla modelu wyniosło 0,81.

Model VI

Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0.61773	0.689689	0.651729	0.727893	0.81087	Model V

Rysunek 32 Statystyki modelu VI (regresja logistyczna)

Model zbudowany na bazie V.

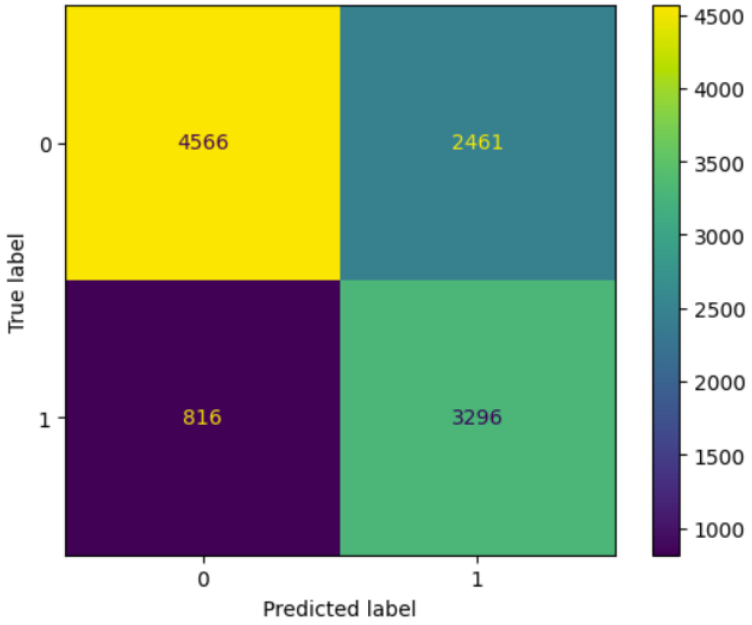


Rysunek 33 Krzywa ROC dla modelu z cross-validacją

Metoda naiwna Bayesa

Metoda naiwna Bayesa jest schematem, który klasyfikuje przypadki opierając się na tw. Bayesa. Jest to typ klasyfikacji statystycznej, których chce przewidzieć prawdopodobieństwo przynależności obiektu do klasy. Wyznaczając P(Y|X), chcę się dowiedzieć, jakie jest prawdopodobieństwo bycia obiektu w stanie X, jeśli charakteryzuje się cechami opisanymi zmienną X. Porównując prawdopodobieństwa dla różnych stanów Y, dokonujemy klasyfikacji obiektu. Metoda jest „naiwna”, gdyż zakłada, że zmienne A i B opisujące obiekt X są niezależne. Jest to konieczne, aby obliczenia nie były zbyt skomplikowane.

Dla metody Bayesa wykonano jeden model.

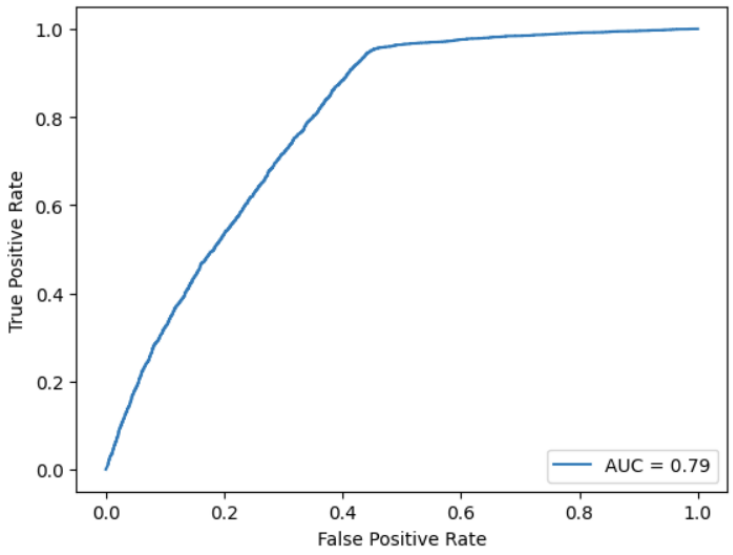


Rysunek 34 Macierz błędów (metoda naiwna Bayesa)

Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0.57252	0.801556	0.66795	0.705808	0.790565	Model_I

Rysunek 35 Statystyki modelu (metoda naiwna Bayesa)

Dokładność dla modelu wyniosła około 71%. Warto zauważyć wysoką wartość czułości dla modelu.



Rysunek 36 Krzywa ROC (metoda naiwna Bayesa)

AUC dla modelu wyniosło 0,79.

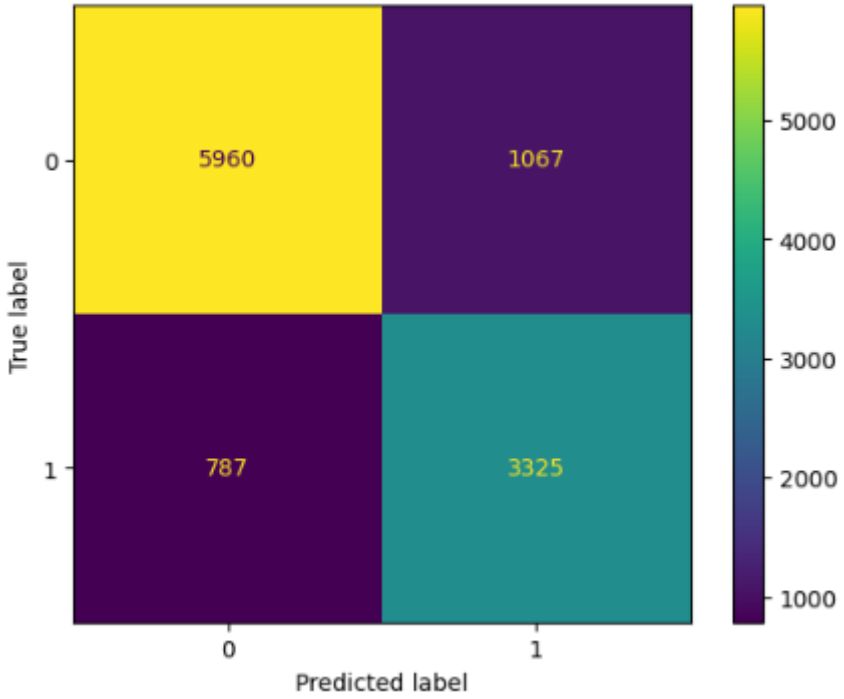
Sekcja najlepszych modeli

Na końcu zdecydowaliśmy się na zastosowanie bardziej zaawansowanego modelu uczenia maszynowego, jakim jest model lasu losowego. Idea, która stoi za ww. modelem polega na konstruowaniu wielu drzew decyzyjnych w czasie uczenia i generowaniu klasy, która jest dominantą klas (klasyfikacja) lub przewidywaną średnią (regresja) poszczególnych drzew.

Już bez większej ingerencji w hiperparametry modelu udało nam się osiągnąć zdecydowaną poprawę zbieranych metryk.

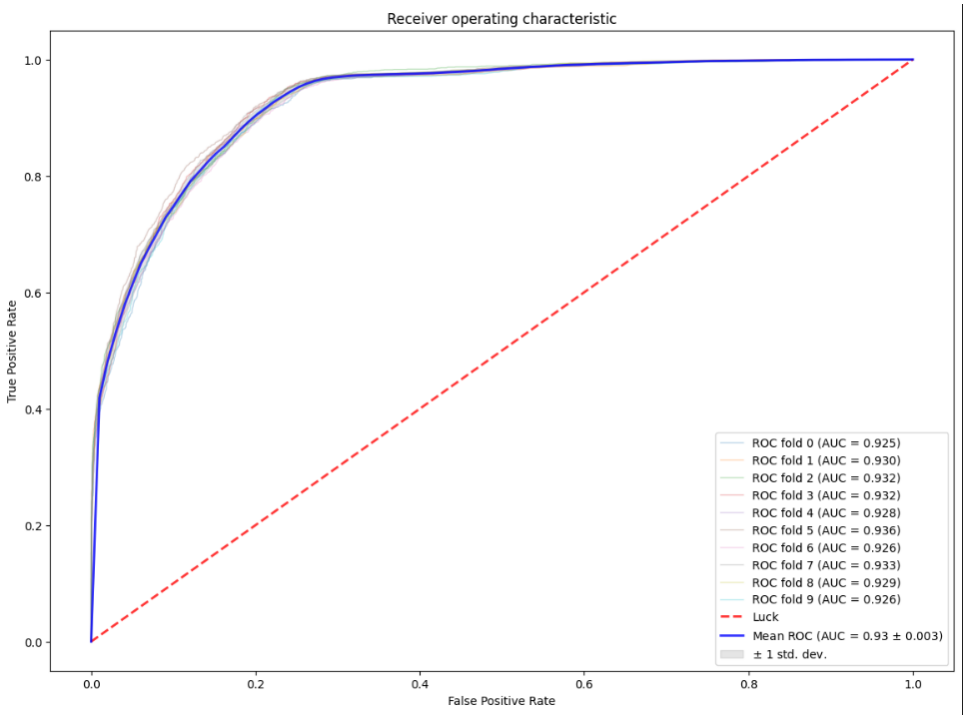
Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0.757058	0.808609	0.781985	0.833558	0.918964	Model_I

Rysunek 37 Zebrane metryki dla modelu lasu losowego



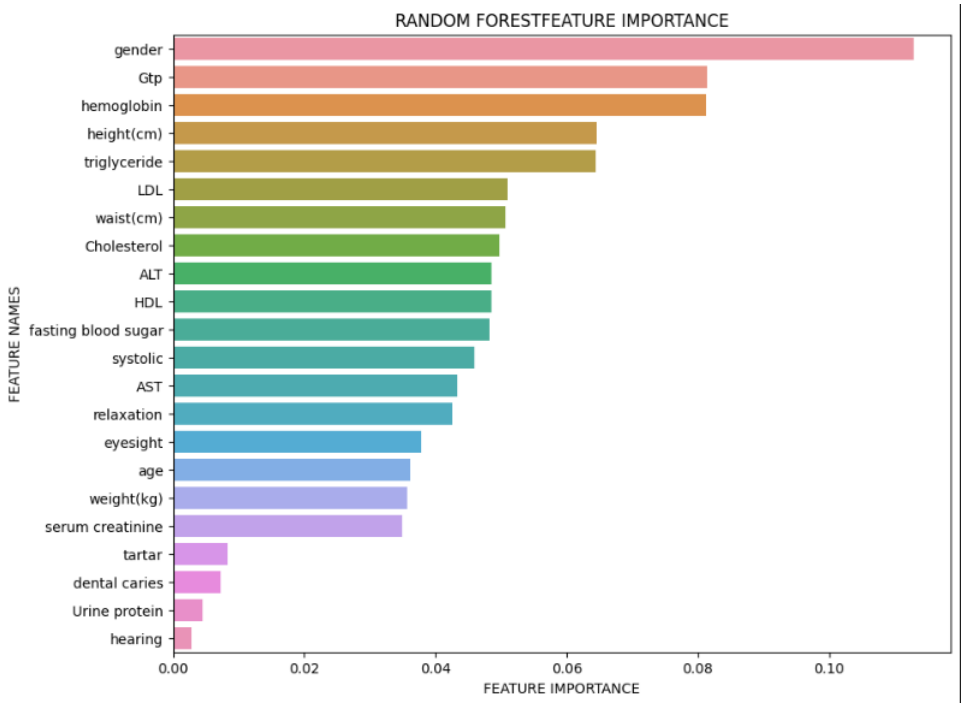
Rysunek 38 Macierz pomyłek dla modelu lasu losowego

Dla powyższego modelu wykonaliśmy również walidację krzyżową dla k=10 i otrzymaliśmy średnią wartość pola pod krzywą ROC na poziomie 93%.



Rysunek 39 Krzywa ROC dla modelu lasu losowego

Algorytmy uczenia maszynowego oparte na drzewach, takie jak las losowy, posiadają atrybut ważności cech(ang. Feature importance), który wyprowadza tablicę zawierającą wartości od 0 do 100 dla każdej cechy, reprezentującą jak bardzo przydatna jest każda cecha w modelu w próbie predykcji. Daje nam to możliwość przeanalizowania, co przyczyniło się do dokładności modelu, a jakie cechy były tylko szumem. Z tymi informacjami możemy sprawdzić, czy model działa tak, jak byśmy tego oczekiwali, odrzucić cechy, jeśli uważamy, że nie dodają żadnej wartości. Poniżej prezentuje się grafika ilustrująca atrybuty ważności cech.



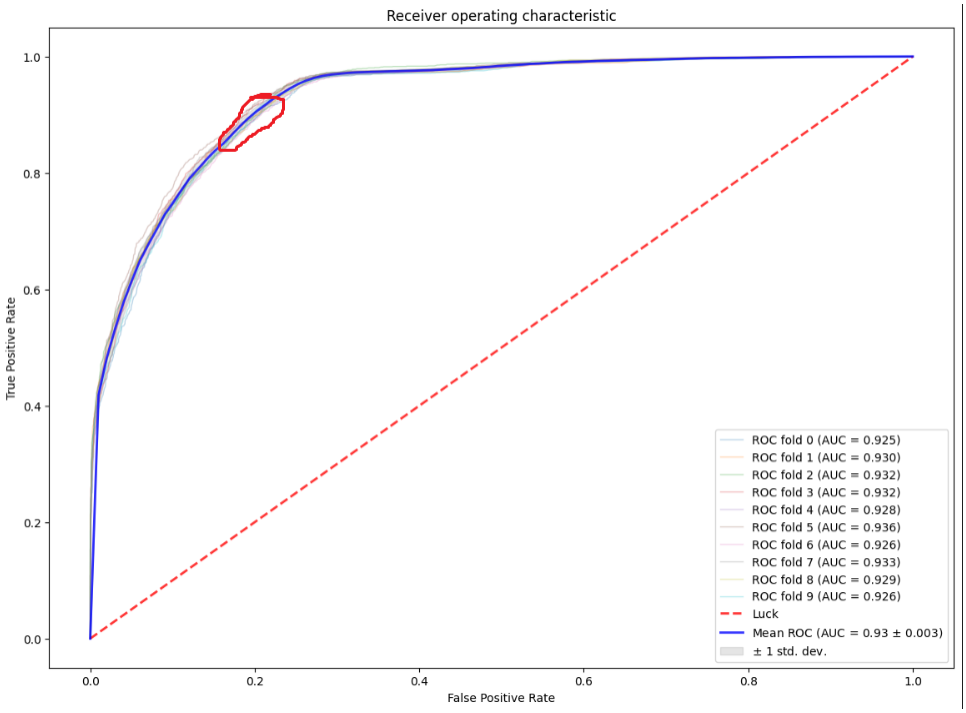
Rysunek 40 Wykres atrybutów ważności cech.

Z wykresu wynika, że zmienne płeć, Gtp i hemoglobin okazują się być najlepsze do predykcji palenia (natomiast słuch, białko w moczu, próchnica i kamień nazębny praktycznie nie wnoszą nic do modelu.

Porównanie i wnioski

Podsumowując, ze wszystkich wybranych modeli najlepszym dla naszego problemu okazał się model lasu losowego. W każdej z badanych metryk(pomijając dokładność ze względu na dysbalans danych) las losowy okazał się “lepszy” od pozostałych.

	Precision	Recall	F1 Score	Accuracy	Roc_auc_score	Name
0	0.641497	0.687986	0.663929	0.742885	0.823637	Logit_raw
0	0.618804	0.677043	0.646615	0.726816	0.810852	Logit 4 variables
0	0.617730	0.689689	0.651729	0.727893	0.810870	Logit_4_vars+reg
0	0.607402	0.570768	0.588516	0.705360	0.758320	KNN_raw
0	0.638091	0.637160	0.637625	0.732651	0.801018	KNN+standarization
0	0.649755	0.644698	0.647217	0.740551	0.807807	KNN+minmaxscaling
0	0.633729	0.675340	0.653873	0.736062	0.818528	KNN_many_neighbours
0	0.57252	0.801556	0.66795	0.705808	0.790565	Naive Bayes
0	0.757645	0.807393	0.781728	0.833558	0.919243	Random Forest



Rysunek 41 Krzywa ROC modelu lasu losowego

Na rysunku przedstawiliśmy optymalny próg odcięcia(wg naszej opinii) dla naszego najlepszego modelu. W zaznaczonej na czerwono przestrzeni mamy zachowaną wysoką pożądaną czułość modelu i wciąż jeszcze niską wartość 1-swoistość.