

Techniki oceny klasyfikacji  
SK.UMA.7  
Dokumentacja końcowa

Projekt realizowany w ramach przedmiotu  
Uczenia maszynowego  
autorstwa Karoliny Romanowskiej i Marianny Gromadzkiej  
kierunek: Informatyka  
Wydział Elektroniki i Technik Informacyjnych  
Politechnika Warszawska

# Spis treści

1	Wstęp	2
2	Interpretacja problemu	2
3	Zbiory danych	2
3.1	Wybrany zbiór	2
3.2	Atrybuty	2
3.3	Analiza danych	2
3.4	Wnioski	4
3.5	Zbiór trenujący i testowy	4
4	Badane algorytmy	5
5	Wykorzystane metryki	5
5.1	Tablica pomyłek	5
5.2	Błąd pierwszego rodzaju — Miara fałszywie pozytywna	6
5.3	Błąd drugiego rodzaju — Miara fałszywie negatywna	6
5.4	Swoistość — Odsetek prawdziwie negatywnych	6
5.5	Czułość — Odsetek prawdziwie pozytywnych	6
5.6	Precyzja — Wartość predykcyjna dodatnia	6
5.7	Wartość predykcyjna ujemna	7
5.8	Wskaźnik fałszywych odkryć	7
5.9	Dokładność	7
5.10	Współczynnik korelacji Matthews’a	7
5.11	Wskaźnik F1	7
5.12	Wskaźnik F-Beta	7
5.13	Strata logarytmiczna — Strata entropii krzyżowej — Strata logistyczna	8
5.14	Współczynnik Kappa Cochena	8
5.15	Krzywa ROC	8
5.16	Wskaźnik ROC AUC	8
5.17	Wynik Briera	8
5.18	Krzywa PR	8
5.19	Wskaźnik PR AUC	8
5.20	Wykres skumulowanego zysku	9
5.21	Wykres Kolmogorov-Smirnov	9
5.22	Wskaźnik Kolmogorov Smirnov	9
6	Ocena jakości modeli przy pomocy metryk	9
6.1	Wybór metryk dla zbioru Breast Cancer Wisconsin	9
6.2	Przedstawienie wyników metryk badanych klasyfikatorów	10
6.3	Podsumowanie	12
7	Implementacja	14
7.1	Opis funkcjonalności	14
7.2	Struktura plików źródłowych	14
7.3	Opis wykorzystanych narzędzi i bibliotek	14
8	Bibliografia	14

# 1 Wstęp

W ramach projektu przedmiotu Uczenie maszynowe został nam przydzielony poniższy temat:

Zaimplementuj techniki oceny klasyfikacji dla zestawów danych dotyczących raka piersi, które dostępne są w <http://archive.ics.uci.edu/ml/datasets>

## 2 Interpretacja problemu

Skupiamy się na różnych metodach oceny klasyfikatorów, dlatego zaimplementujemy jedenaście różnych algorytmów za pomocą biblioteki scikit-learn. Jakość predykcji każdego z nich zostanie poddana analizie na podstawie metryk dla klasyfikatorów binarnych.

## 3 Zbiory danych

### 3.1 Wybrany zbiór

Wybrałyśmy zbiór Breast Cancer Wisconsin (Dataset) Data Set składający się z 569 rekordów o 30 atrybutach w postaci liczb rzeczywistych. 357 przypadków to przypadki łagodne, 212 złośliwe.

### 3.2 Atrybuty

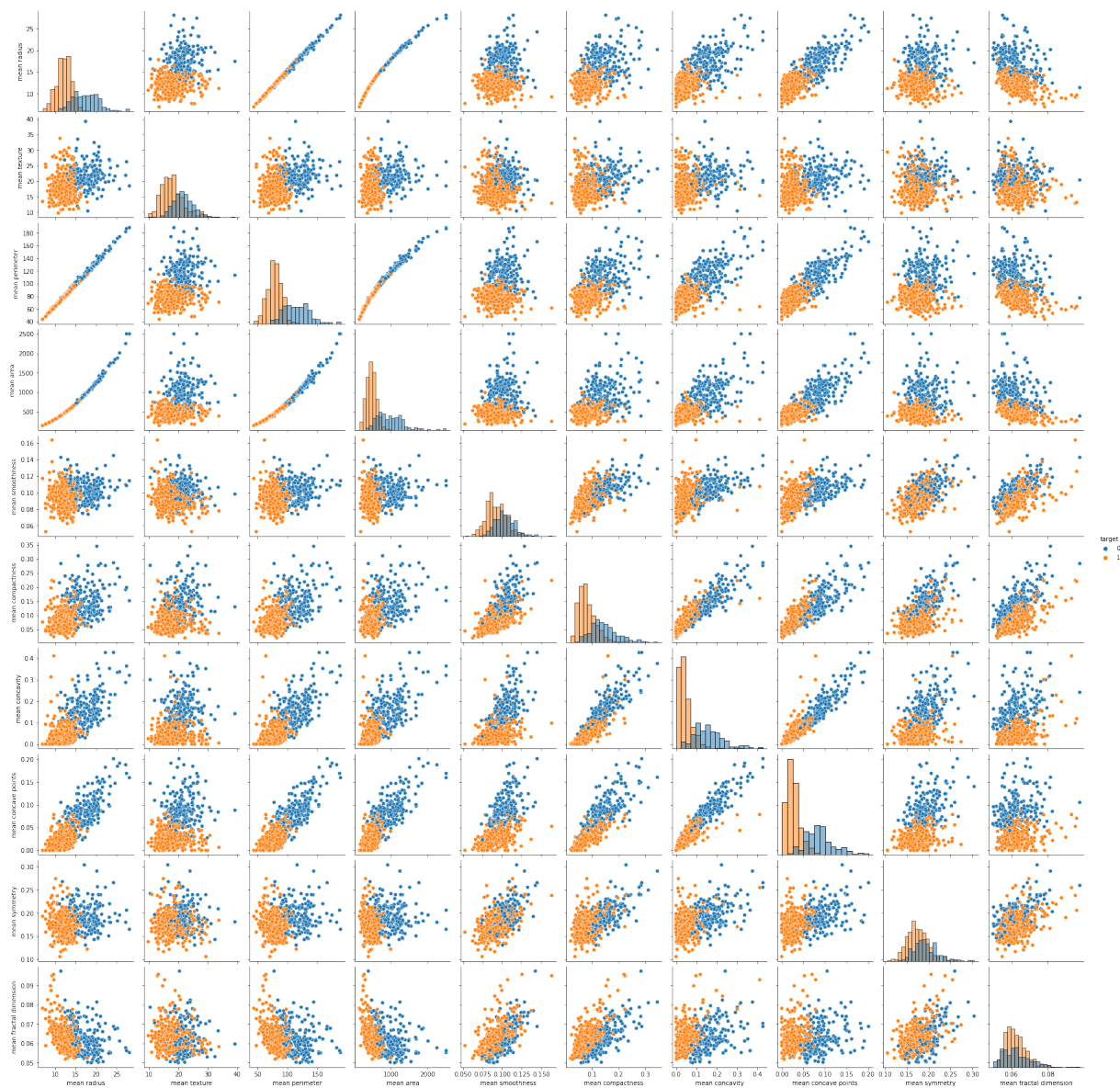
Atrybuty opisują właściwości jąder komórkowych widocznych na zdjęciach z biopsji:

- promień
- tekstura
- obwód
- powierzchnia
- gładkość (lokalna zmienność długości promieni)
- wklęsłość (dotkliwość wklęsłych fragmentów konturu)
- punkty wklęsłe (liczba wklęsłych części konturu)
- symetria
- wymiar podobieństwa
- ścisłość

Dla każdej właściwości podana jest średnia, błąd standardowy oraz "najgorsze przypadki" (średnia trzech największych wartości dla danej właściwości).

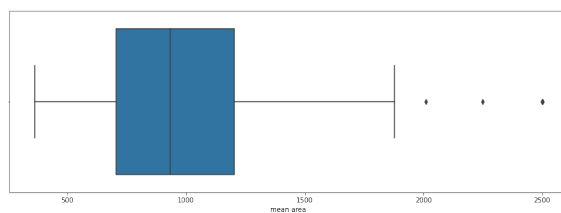
### 3.3 Analiza danych

Na wykresach przedstawiliśmy histogramy (na przekątnej) średnich wartości właściwości komórek oraz pairploty. Można zauważyć, że dla tekstury, symetrii oraz wymiaru podobieństwa histogramy pokrywają się i klasy są trudne do rozdzielenia.

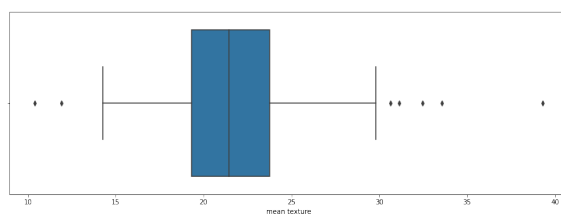


Rysunek 1: Legenda: pomarańczowy -przypadek łagodny, niebieski - przypadek złośliwy

W obu klasach w wielu atrybutach możemy zaobserwować górne wartości ekstremalne, np.:



Rysunek 2: Przypadek złośliwy



Rysunek 3: Przypadek łagodny

### 3.4 Wnioski

Z powodu za dużego podobieństwa w wartościach między klasami odrzucimy atrybuty: tekstura, symetria i wymiar podobieństwa. Odrzucone też zostaną przypadki w których wartości atrybutów są odstające.

### 3.5 Zbiór trenujący i testowy

Dataset zostanie podzielony na zbiór testowy (50%), walidacyjny (20%) i trenujący (30%).

## 4 Badane algorytmy

Algorytm	Opis
Maszyna wektorów nośnych	Ma na celu wyznaczenie hiperpłaszczyzny rozdzielającej przykłady z klas z maksymalnym marginesem
Perceptron wielowarstwowy	Sztuczna sieć neuronowa składająca się z warstw pojedynczych neuronów
K najbliższych sąsiadów	Wyznacza k badanych sąsiadów, do których badany element ma najbliżej dla wybranej metryki
Klasyfikacja procesem gaussowskim	Wykorzystuje leniwe uczenie się i miarę podobieństwa między punktami (funkcję jądra), aby przewidzieć wartość niewidocznego punktu na podstawie danych uczących
Drzewo decyzyjne	Etykietowane drzewo, w którym każdy węzeł wewnętrzny odpowiada przeprowadzeniu pewnego testu na wartościach atrybutów i wyznacza optymalny punkt podziału
ExtraTree	Etykietowane drzewo, w którym każdy węzeł wewnętrzny odpowiada przeprowadzeniu pewnego testu na wartościach atrybutów i wyznacza losowy punkt podziału
Las losowy	Polega na konstruowaniu wielu drzew decyzyjnych w czasie uczenia i generowaniu klasy, która jest dominantą klas (klasyfikacja) lub przewidywaną średnią (regresja) poszczególnych drzew
Klasyfikator AdaBoost	Silny klasyfikator, łącząc wiele klasyfikatorów o słabej wydajności. Podstawową koncepcją Adaboost jest ustawienie wag klasyfikatorów i próbek danych uczących w każdej iteracji, tak aby zapewnić dokładne przewidywania nietypowych obserwacji
Naiwny Klasyfikator Bayesowski	Prosty klasyfikator probabilistyczny oparty na założeniu o wzajemnej niezależności zmiennych objaśniających.
Kwadratowa analiza dyskryminacyjna	Jej zadaniem jest rozstrzyganie, które zmienne niezależne w najlepszy sposób dzielą dany zbiór przypadków na występujące w naturalny sposób grupy, opisane jako ściową zmienną zależną.
Regresja logistyczna	Jest szczególnym przypadkiem uogólnionego modelu liniowego, znajdującym zastosowanie, gdy zmienna zależna jest dychotomiczna.

Tabela 1: Opisy rozważanych algorytmów

## 5 Wykorzystane metryki

### 5.1 Tablica pomyłek

Tablica stosowana podczas oceny klasyfikacji binarnej. Wykorzystywane są tutaj etykiety przypisane przez model oraz klasy rzeczywiste. W tablicy można znaleźć ilość próbek prawdziwie pozytywnych - tp, fałszywie pozytywnych - fp, fałszywie negatywnych - fn i prawdziwie negatywnych - tn. Oś Y przedstawia klasy rzeczywiste, oś X przedstawia klasy przewidywane.

	Klasyfikacja pozytywna	Klasyfikacja negatywna
Stan pozytywny	Prawdziwie dodatnia - TP	Fałszywie ujemna - FN
Stan negatywny	Fałszywie dodatnia - FP	Prawdziwie ujemna - TN

Tabela 2: Tablica pomyłek

## 5.2 Błąd pierwszego rodzaju — Miara fałszywie pozytywna

Błąd pierwszego rodzaju przedstawia ilość przypadków, w których odrzucona została hipoteza zerowa, pomimo jej prawdziwości. Często stosuje się tę metrykę jako pomocniczą do innej metryki, zazwyczaj do pary ze wskaźnikiem fałszywych odkryć. Można ją interpretować jako częstość fałszywych alarmów

$$FPR = \frac{fp}{fp + tn} \quad (1)$$

## 5.3 Błąd drugiego rodzaju — Miara fałszywie negatywna

Błąd drugiego rodzaju przedstawia ilość przypadków, w których hipoteza zerowa nie została odrzucona, pomimo jej fałszywości. Często stosuje się tę metrykę jako pomocniczą do innej metryki.

$$FNR = \frac{fn}{tp + fn} \quad (2)$$

## 5.4 Swoistość — Odsetek prawdziwie negatywnych

Swoistość przedstawia stosunek ilości wyników prawdziwie negatywnych do sumy prawdziwie negatywnych oraz fałszywie pozytywnych. Przedstawia w jakim procencie klasa będąca w rzeczywistości negatywną, została przewidziana jako negatywna. Wraz z czułością metryka wskazuje poziom dokładności predykcji. Używana dla potwierdzenia poprawności predykcji. Warto z niej skorzystać gdy w naszych przewidywaniach chcemy uchwycić jak najwięcej przypadków negatywnych.

$$TNR = \frac{tn}{fp + tn} \quad (3)$$

## 5.5 Czułość — Odsetek prawdziwie pozytywnych

Czułość przedstawia stosunek ilości wyników prawdziwie pozytywnych do sumy prawdziwie pozytywnych oraz fałszywie negatywnych. Można ją interpretować jako prawdopodobieństwo poprawności klasyfikacji pod warunkiem rozpatrywania przypadku pozytywnego. Wraz ze swoistością metryka wskazuje poziom dokładności predykcji. Warto z niej skorzystać gdy w naszych przewidywaniach chcemy uchwycić jak najwięcej przypadków pozytywnych.

$$TPR = \frac{tp}{tp + fn} \quad (4)$$

## 5.6 Precyzja — Wartość predykcyjna dodatnia

Precyzja przedstawia stosunek ilości wyników prawdziwie pozytywnych do sumy prawdziwie pozytywnych oraz fałszywie pozytywnych. Metryka ta mierzy ile predykcji pozytywnych jest w rzeczywistości pozytywnych. Stosowana gdy zależy nam na wysokiej pewności w stosunku do prawdziwości naszych predykcji

$$PPV = \frac{tp}{tp + fp} \quad (5)$$

## 5.7 Wartość predykcyjna ujemna

Wartość predykcyjna ujemna przedstawia stosunek ilości wyników prawdziwie negatywnych do sumy prawdziwie negatywnych i fałszywie negatywnych. Metryka ta mierzy ilość prawdziwie negatywnych przypadków która została zakwalifikowana poprawnie. Można ją interpretować jako precyzję wyznaczania klas negatywnych.

$$NPV = \frac{tn}{tn + fn} \quad (6)$$

## 5.8 Wskaźnik fałszywych odkryć

Wskaźnik fałszywych odkryć mierzy ile predykcji ze wszystkich pozytywnych było w rzeczywistości fałszywymi. Można ją interpretować jako częstość fałszywych odkryć. Często podaje się tę miarę wraz z Miarą fałszywie pozytywną.

$$FDR = \frac{fp}{fp + tp} \quad (7)$$

## 5.9 Dokładność

Dokładność mierzy ile predykcji, zarówno pozytywnych jak i negatywnych, była poprawnie zaklasyfikowana. Nie należy używać tej miary przy niezbalansowanych zbiorach danych dla problemu - wtedy łatwo uzyskać wysoką dokładność, poprzez klasyfikację wszystkich obserwacji do klasy większościowej. Metryki warto używać gdy każda z klas jest dla nas równie ważna.

$$ACC = \frac{tp + tn}{tp + tn + fp + fn} \quad (8)$$

## 5.10 Współczynnik korelacji Matthews'a

Współczynnik korelacji Matthews'a to korelacja między przewidywaną klasą a obserwowaną. Przyjmuje wartości w zakresie  $[-1, 1]$ , gdzie wartość 1 - oznacza klasyfikację bez pomyłek, 0 - jakość klasyfikacji równą klasyfikatorowi losowemu i 1 - klasyfikator zawsze zwracający błędne wyniki. Współczynnik uważany jest za zrównoważoną miarę - oznacza to, że znajduje swoje zastosowanie, również gdy nasze dane są niezbalansowane pod względem przynależności do klas.

$$MCC = \frac{tp * tn - fp * fn}{(tp + fp)(tp + fn)(tn + fp)(tn + fn)} \quad (9)$$

## 5.11 Wskaźnik F1

Wskaźnik F1 zdefiniowana jakoś średnia harmoniczna z precyzji i czułości. Średnia harmoniczna, w porównaniu do średniej arytmetycznej, jest nie czuła na wysokie wartości pojedynczych parametrów. Z tego powodu wykorzystywana jest gdy chcemy sprawdzić czy istnieje balans między precyzją a czułością w naszym modelu.

$$F_1 = \frac{2 * PPV * TPR}{PPV + TPR} \quad (10)$$

## 5.12 Wskaźnik F-Beta

Wskaźnik F-Beta pozwala na dostosowanie metryki do naszych potrzeb. Im bardziej zależy nam na czułości, tym wyższe beta powinniśmy zastosować. Parametr ten przyjmuje zazwyczaj wartości od 0.1 do 10. Szczególnie często przyjmuje wartości 0.5 - gdy zależy nam bardziej na precyzji i 2 - gdy zależy nam na czułości.

$$F_{beta} = (1 + \beta^2) \frac{PPV * TPR}{\beta^2 * PPV + TPR} \quad (11)$$



### 5.13 Strata logarytmiczna — Strata entropii krzyżowej — Strata logistyczna

Metryka ta jest używana gdy wyjściem klasyfikatora jest prawdopodobieństwo przewidywania klasy. Przyjmuje wartości w zakresie  $[0, \infty)$ . Wartości bliskie 0 wskazują na dokładność predykcji. Dążymy do minimalizacji tego wskaźnika.

$$\text{logloss}_{(N=1)} = -\frac{1}{N} \sum_{i=1}^N [y_i * \ln(p_i) + (1 - y_i) * \ln(1 - p_i)] \quad (12)$$

### 5.14 Współczynnik Kappa Cochena

Współczynnik Kappa Cochena przedstawia o ile oceniany model jest lepszy od losowego klasyfikatora. Parametr  $p_e$  oznacza sumę prawdopodobieństwa przypadkowej zgodności prognozy z rzeczywistymi wartościami klasy pierwszej oraz prawdopodobieństwo przypadkowej zgodności z rzeczywistymi wartościami klasy drugiej. Współczynnik ten przyjmuje wartości w zakresie  $[-1, 1]$ . Współczynnik ten wyniesie niższe wartości dla danych niezrównoważonych.

$$\kappa = \frac{ACC - p_e}{1 - p_e} \quad (13)$$

### 5.15 Krzywa ROC

Jest to wykres który obrazuje zależność między metrykami TPR (czułością) i FPR (błędem pierwszego rodzaju). Dla każdego progu wyliczamy miary i wykreślamy je na wykresie. Krzywa ROC pozwala na szybkie porównanie ze sobą klasyfikatorów oraz wyboru dla nich progu odcięcia klas. Dążymy do uzyskania wysokiej wartości TPR i niskiej FPR.

### 5.16 Wskaźnik ROC AUC

Wskaźnik ROC AUC przedstawia wartość pola powierzchni pod krzywą ROC. Jest to prawdopodobieństwo, że oceniany model klasyfikacji binarnej przydzieli wyższe prawdopodobieństwo losowo wybranemu przypadkowi pozytywnemu, niż losowo wybranemu przypadkowi negatywnemu. Chcemy maksymalizować tę metrykę tak aby nasza krzywa ROC sięgała jak najbliżej lewemu górnemu rogu wykresu.

### 5.17 Wynik Brier

Wynik Brier to metryka służąca do sprawdzenia poprawności przewidywanego wyniku. Jest bardzo podobna do błędu średniokwadratowego, ale można ją zastosować jedynie dla prawdopodobieństwa wyniku z modelu będącego w zakresie  $[0, 1]$  Metryka używana do kalibracji prawdopodobieństwa modeli uczenia maszynowego. Stanowi dobre uzupełnienie metryki ROC AUC.

$$\text{brierloss} = (y_{pred} - y_{true})^2 \quad (14)$$

### 5.18 Krzywa PR

Krzywa PR obrazuje zależności między metrykami PPV i TPR. Pomaga dobrać próg odcięcia klas.

### 5.19 Wskaźnik PR AUC

Podobnie jak ROC AUC, metryka PR AUC przedstawia wartość pola powierzchni pod krzywą PR. Ta metryka może być również interpretowana jako średnia z precyzji obliczana dla każdej wartości czułości.

## 5.20 Wykres skumulowanego zysku

Wykres skumulowanego zysku pozwala ocenić, jak duży będzie zysk z wykorzystania modelu klasyfikacji w porównaniu z modelem losowym dla danej miary najwyżej ocenionych predykcji. Należy uporządkować predykcje od najwyższych do najniższych i dla każdego percentyla wyliczana jest miara prawdziwych pozytywów. Wykres pozwala dostrzec korzyści wynikające z zastosowania klasyfikatora.

## 5.21 Wykres Kolmogorov-Smirnov

Wykres Kolmogorova-Smirnova przedstawia dystans pomiędzy proporcją sukcesu:

$$\frac{\text{liczba pozytywow w danym percentylu}}{\text{liczba wszystkich pozytywow}} \quad (15)$$

a proporcją porażki:

$$\frac{\text{liczba negatywow w danym percentylu}}{\text{liczba wszystkich pozytywow}} \quad (16)$$

wyliczaną dla kolejnych percentyli uporządkowanych malejąco predykcji.

## 5.22 Wskaźnik Kolmogorov Smirnov

Jest to największy dystans pomiędzy wykresami Kolmorova-Smirnova. Przydatny gdy zależy nam równo na klasach pozytywnych co negatywnych.

# 6 Ocena jakości modeli przy pomocy metryk

## 6.1 Wybór metryk dla zbioru Breast Cancer Wisconsin

Po zapoznaniu się ze zbiorem danych oraz dostępnymi metrykami, wybrałyśmy najodpowiedniejsze naszym zdaniem metryki do klasyfikowanej dziedziny. Nasz klasyfikator, ma za zadanie przydzielić na podstawie parametrów, czy pacjent ma nowotwór łagodny czy złośliwy. Zbiór jest dość zbalansowany (w stosunku 3:2), z tego powodu przy wybieraniu metryk nie musieliśmy decydować na tej podstawie.

Na wstępie odrzuciliśmy tablicę pomyłek, ze względu na jej binarny charakter, z obawy przed niepoprawnym utartym zrozumieniem wyników przedstawionych przez metrykę. Dotyczył on podziału na przypadki pozytywne i negatywne, u nas jednak klasyfikacja binarna polegała na przydziale do klasy pacjenta z nowotworem złośliwym i łagodnym.

Aby porównać jakość modeli dla rozważanego problemu posłużyłyśmy się krzywą ROC, na podstawie której, wybrałyśmy z 11 początkowo rozważanych modeli, 6 do dalszej analizy.

Sześć kolejnych modeli zostało ze sobą porównanych na podstawie kilku metryk, które naszym zdaniem, dobrze oddają istotę problemu klasyfikacji. Jako pierwszą, zbadaliśmy dokładność predykcji naszych modeli. Do naszego zbioru danych wybrałyśmy metrykę dokładności (eng. Accuracy) ze względu na prostą interpretację oraz zbalansowane dane zbioru Breast Cancer Wisconsin. Podobną metryką przedstawiającą dokładność jest Wskaźnik F1, który szczególnie zalecany jest dla oceny modeli uczonych na danych niezbalansowanych. Pomimo, że nasz zbiór taki aktualnie nie jest, to uznaliśmy, że warto również dodać metrykę, aby w razie dalszego douczania modelu, łatwo zauważyć moment, w którym wskaźnik Accuracy, przestanie prawidłowo, przedstawiać wyniki klasyfikacji naszego modelu.

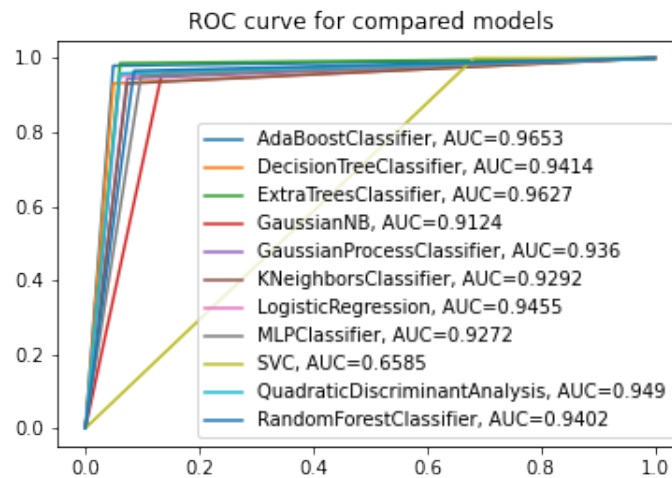
Aby poprawnie ocenić jakość klasyfikacji chcemy zbadać jakość korelacji między przewidywaną klasą a obserwowaną. Taką zależność można przedstawić przy pomocy współczynnika Kappa Cohena, jednak jest on wrażliwy na ocenę modeli uczących się na danych niezbalansowanych. Współczynnik korelacji Matthews'a takiej niedoskonałości nie posiada, z tego też powodu zdecydowałyśmy się go wybrać jako trzecią metrykę.

W ramach oceny jakości klasyfikacji, uznaliśmy, że powinniśmy sprawdzić pewność predykcji klasyfikatora, z tego powodu dla każdej predykcji, sprawdzimy również wysokość prawdopodobieństwa. Do oceny prawdopodobieństwa użyjemy metryki straty logarytmicznej.

Po porównaniu wyniku tych metryk, zdecydowaliśmy się na wybór modelu dla tego problemu. W ramach przedstawienia korzyści wynikające z wdrożenia klasyfikatora, wykreśliłyśmy wykres skumulowanego zysku.

## 6.2 Przedstawienie wyników metryk badanych klasyfikatorów

Dla wszystkich jedenastu klasyfikatorów została wykreślona krzywa ROC. Jest to częsta metoda eliminacji modeli oraz ich rankingowania. Podczas porównywania należy patrzeć, która z krzywych znajduje się najbliżej lewego górnego rogu.

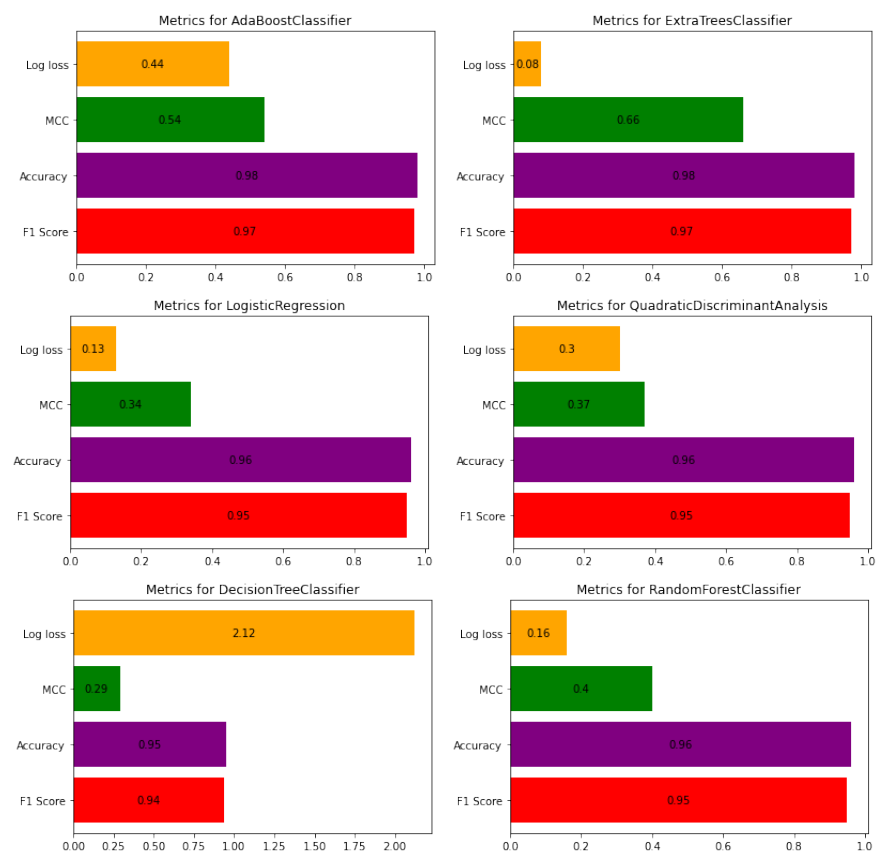


Rysunek 4: Krzywa ROC dla wszystkich modeli

Na podstawie krzywej wytypowałyśmy 6 modeli do dalszej analizy. Są to:

- Klasyfikator AdaBoost
- ExtraTree
- Regresja logistyczna
- Kwadratowa analiza dyskryminacyjna
- Drzewo decyzyjne
- Las losowy

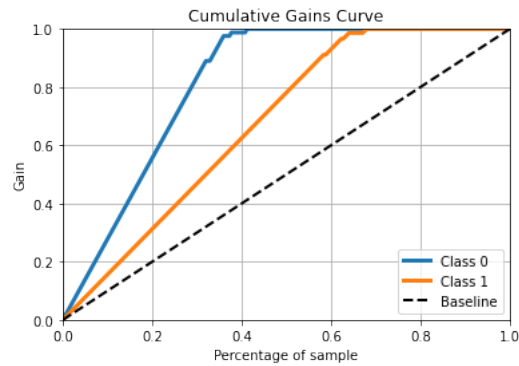
Następnie dla każdego z 6 modeli, sprawdziliśmy wyniki metryk Dokładności, wskaźnika F1, współczynnika korelacji Matthews'a oraz wysokość straty logarytmicznej. Wyniki zwizualizowane są poniżej.



Rysunek 5: Porównanie modeli przy pomocy wybranych metryk

Odpowiadające sobie wielkości metryki Accuracy i wskaźnika F1 potwierdzają, że mamy do czynienia ze zbalansowanymi danymi. Metryka Accuracy dla wybranych pięciu modeli jest w granicach 95%-98%. Kolejną metryką, na którą warto zwrócić uwagę jest współczynnik korelacji Matthews'a - chcemy aby był on bliski wartości jeden. Wartości bliskie zero, oznaczałyby, że nasz model nie jest lepszy niż losowy klasyfikator. Rozważane modele mają wartości zaczynający się od 0.29. Dwa najlepsze modele osiągają wynik 0.54 i 0.66. Ostatnią metryką, którą bierzemy pod uwagę jest strata logarytmiczna. Model jest tym lepszy im wartość tej metryki jest bliższa zero.

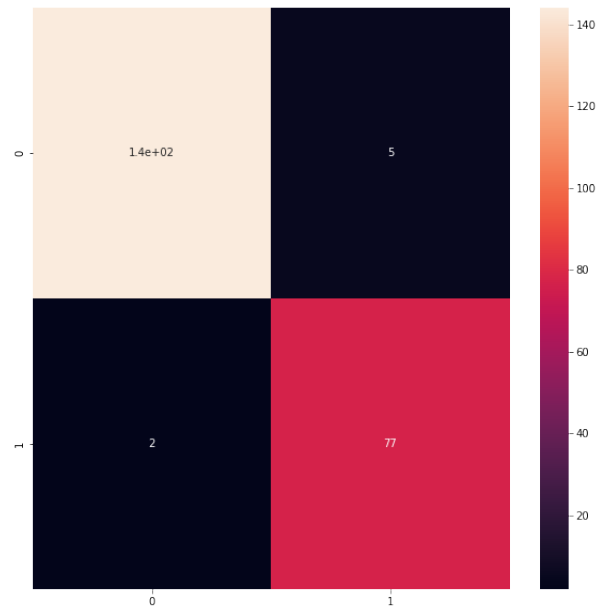
Na podstawie wyznaczonych metryk, oczywistym staje się wybór modelu ExtraTreeClassifier. Osiąga on najlepsze wyniki dla powierzonego mu zadania.



Rysunek 6: Wykres skumulowanego zysku dla ExtraTreeClassifier

### 6.3 Podsumowanie

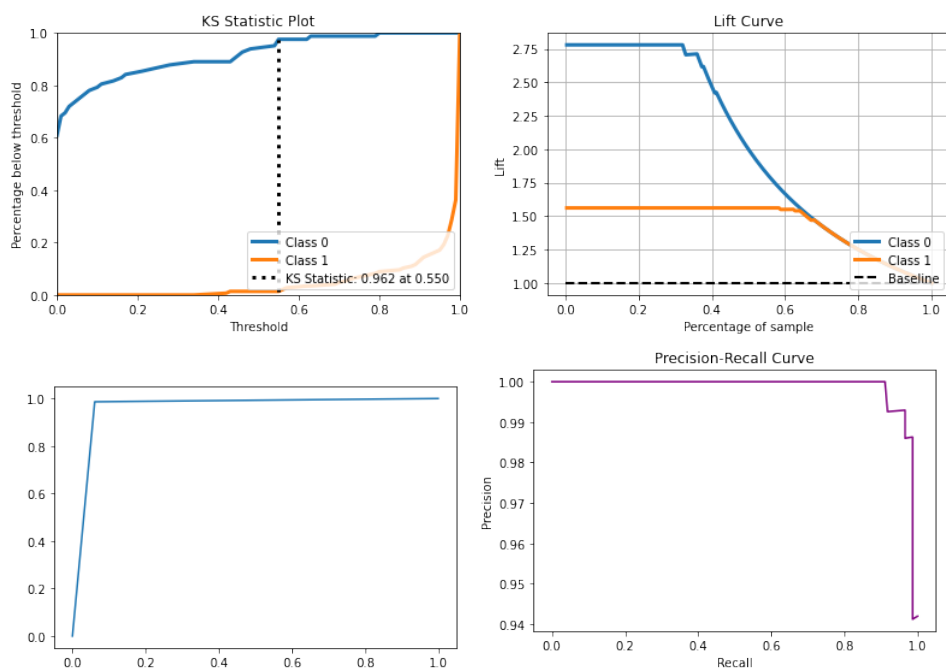
Po ocenie różnych modeli, okazało się, że najlepszym klasyfikatorem dla naszego problemu jest ExtraTreeClassifier. W ramach głębszej analizy modelu, przedstawiamy pozostałe metryki, które mogą być przydatne podczas dalszego utrzymania modelu.



Rysunek 7: Macierz pomyłek

Nazwa	Wartość
Błąd pierwszego rodzaju - Miara fałszywie pozytywna	0.061
Błąd drugiego rodzaju - Miara fałszywie negatywna	0.014
Swoistość - Odsetek prawdziwie negatywnych	0.014
Czułość - Odsetek prawdziwie pozytywnych	0.986
Precyzja - Wartość predykcyjna dodatnia	0.966
Wartość predykcyjna ujemna	0.975
Wskaźnik fałszywych odkryć	0.034
Dokładność	0.969
Współczynnik korelacji Matthews'a	0.664
Wskaźnik F1	0.976
Wskaźnik F-Beta dla Beta=0.5	0.97
Wskaźnik F2	0.982
Strata logarytmiczna	0.076
Współczynnik Kappa Cochena	0.933
Wskaźnik ROC AUC	0.963
Wynik Briera	0.031
Wskaźnik PR AUC	0.962

Tabela 3: Wartość metryk dla ExtraTreeClassifier



Rysunek 8: Wykresy metryk dla ExtraTreeClasification

## 7 Implementacja

### 7.1 Opis funkcjonalności

W ramach projektu powstały trzy notatniki Jupyter Notebook. W notatniku Analysis znajduje się analiza zbioru danych, dla którego będziemy poszukiwać odpowiednich metryk do oceny modeli trenowanych na tym zbiorze. Notatnik Models, generuje i trenuje modele klasyfikacji binarnej, które następnie w notatniku Metrics zostaną poddane analizie. Na podstawie wartości metryk można wybrać model odpowiadający postawionemu problemowi.

### 7.2 Struktura plików źródłowych

Poniżej przedstawiona jest struktura plików źródłowych wraz z krótkim opisem zawartości każdego z nich

```
UMA-Classification-Evaluation/
├── plots/                <- Wygenerowane do sprawozdania wykresy
├── reports/              <- Dokumentacje opisujące projekt
├── saved_models/         <- Zapisane wytrenowane modele
├── Analysis.ipynb        <- Analiza zbioru Breast Cancer Wisconsin
├── Metrics.ipynb         <- Notatnik wyliczający metryki dla klasyfikacji binarnej
├── Models.ipynb          <- Notatnik generujący modele klasyfikacji binarnej dla zadanego zbioru
├── README.md             <- Opis projektu
└── requirements.txt      <- Opis wymagań
```

### 7.3 Opis wykorzystanych narzędzi i bibliotek

Projekt został zaimplementowany przy pomocy Pythona 3.8.8, w Jupyter Notebook, w środowisku Google Colab. W ramach projektu skorzystałyśmy z bibliotek: scikit-plot, matplotlib w wersji 3.4.0, seaborn oraz sklearn.

## 8 Bibliografia

Klasyfikacja procesem gaussowskim

Regresja logistyczna

Maszyna wektorów nośnych

Drzewa decyzyjne

Las losowy i extra tree

Klasyfikator AdaBoost

Naiwny klasyfikator bayesowski

Analiza dyskryminacyjna

Regresja logistyczna

Dokument porównujący modele drzew na podstawie metryk